“Enlisting Big Data in the Fight Against Coronavirus”

Senate Committee on Commerce, Science, and Transportation

April 9, 2020

Written Testimony of Professor Ryan Calo[*]

Mr. Chairman, Ranking Member Cantwell, and Members of the Committee:

I am deeply grateful for the opportunity to testify before this Committee regarding the promise and peril of using data analytics to combat the novel coronavirus. There can be little doubt that better access to and analysis of information will play a prominent role in addressing the ongoing pandemic. Yet even as we bring to bear the considerable ingenuity of our academic, public, and private institutions, my research into privacy and technology counsels a measure of humility and caution regarding the use of data analytics to address this crisis.

In this testimony, I will address some of the ways people and institutions propose to use data analytics and other technology to respond to coronavirus. The first set of examples involves gaining a better understanding of the virus and its effects on American life. By and large I support these efforts; the value proposition is clear and the privacy harms less pronounced. The second set of examples involves the attempt to track the spread of COVID-19 at an individual level using mobile software applications (“apps”). I am more skeptical of this approach as I fear that it threatens privacy and civil liberties while doing little to address the pandemic. Finally, I conclude with the recommendation that, however we leverage data to fight this pandemic, policymakers limit use cases to the emergency itself, and not permit mission creep or downstream secondary uses that surprise the consumer.

A recent United Nations and World Health Organization study concluded that proposals to address COVID-19 using big data and artificial intelligence operate at three scales.[1] At the *medical scale*, researchers propose using data analytics to diagnosis patients and recommend individual treatment. At the *molecular scale*, researchers propose data-driven methods to better understand the structure of the virus, improve testing, and discover new treatments or precautions. My remarks

[*] Lane Powell and D. Wayne Gittinger Associate Professor, School of Law; Associate Professor (by courtesy), Paul G. Allen School of Computer Science and Engineering; Associate Professor (by courtesy), Information School; Co-Director, Tech Policy Lab and Center for an Informed Public; University of Washington. I am indebted to Ashkan Soltani and Eric Sears for early feedback and PhD candidate María Angel for her excellent and timely research assistance. All opinions expressed are my own and institutional affiliations are for identification purposes only.

[1] Joseph Bullock et al., Mapping the Landscape of Artificial Intelligence Applications Against COVID-19, United Nations Global Pulse (March 2020), online at https://arxiv.org/pdf/2003.11336.pdf.

largely concern the use of data and digital technology to address COVID-19 at the *societal scale*—efforts such as encouraging social distancing, predicting where the next outbreak will occur, or determining who has been infected or exposed.

Apart from widespread testing, among the most powerful tools at our disposal in combating COVID-19 is social distancing. Social distancing slows the spread of the virus so that hospitals are not overwhelmed. Of course, this distancing is painful; people are by nature social animals and avoiding one another has profound social, cultural, and economic consequences. Setting aside the many Americans who perform essential services such as providing food or healthcare, not every individual or community is heeding the advice of epidemiologists or their government to shelter in place.

Data held by companies could help promote social distancing. Google's COVID-19 Community Mobility Report sheds light on social distancing compliance across the country and the world by displaying month-by-month reports on how much given communities are traveling to work or using public transportation relative to a pre-coronavirus baseline.[2] Google is using consumer location information, which is a highly sensitive form of data.[3] But because the data is aggregated and displayed only as a relative percentage, the risks to individuals are mitigated. Meanwhile, the data is useful to policymakers in determining where additional social distancing measures might be needed and to health officials in assessing the correlation between social distancing and rates of viral transmission.

Tragically, severe and fatal cases of the novel coronavirus have already overwhelmed healthcare systems in several nations and in our own nation's largest city. Other areas are at risk of being overrun. Institutions are trying to harness the predictive power of data analytics to determine the location and severity of the next outbreak so that communities can better prepare and local resources can be bolstered. Some approaches compare hospital capacity and resources (such as ventilators) against population density and infection rates.[4] Others use a single metric—such as rates of fever or oxygenation levels—as a heuristic for the incidence of disease in a particular community. According to reporting by the New York Times, the connected thermometer company Kinsa Health was able to see evidence that social distancing was slowing the spread of coronavirus within a day of new social distancing measures being put in place.[5]

[2] The reports from Google can be found here: https://www.google.com/covid19/mobility/.

[3] In *United States v. Carpenter*, the Supreme Court acknowledged that location records "hold for many Americans the 'privacies of life'" and that a government with access to historic location data "achieves near perfect surveillance." 585 U. S. ____, *12-13 (2018) (internal citations omitted).

[4] The COVID-19 Capacity Predictor from Definitive Healthcare and Esri, for example, is available here: https://www.definitivehc.com/resources/covid-19-capacity-predictor.

[5] Donald G. McNeil, Jr., Restrictions Are Slowing Coronavirus Infections, New Data Suggest, New York Times (March 30, 2020).

Obviously it would be of tremendous value to know where the novel coronavirus threatens to overrun a community's healthcare resources. Yet there is reason for caution and humility before making decisions of material consequence on the basis of artificial intelligence or other methods of data analytics. Google Flu Trends applied complex mathematical models to consumer search terms in order to predict flu outbreaks in 2009 around the time of the H1N1 pandemic, garnering public acclaim and a nod from the Center for Disease Control. Yet just a few years later, the model ceased to predict the incidence of flu with any accuracy, as chronicled in both *Nature* and *Science* in 2014.[6] Google quietly shuttered the project.

Even when artificial intelligence works, it does not always work for everyone. As research by Ruha Benjamin, Safiya Umoja Noble, Virginia Eubanks, Kate Crawford, and other leading academics has shown, the vulnerable and marginalized seldom realize the full benefits of AI systems.[7] Imagine, for example, that public health officials were to allocate coronavirus resources on the basis of data trends from connected thermometers like Kinsa Health (retail cost: $35.99 – $69.99) or connected pulse oximeters like iHealth Air (retail cost: $69.99). Only communities where sufficient numbers of consumers were aware of such devices and could afford them would receive an early warning or stockpiled support.

I have described several efforts with the potential to assist policymakers, public health officials, and others in making wiser decisions around coronavirus. There has been considerable public attention focused instead on technologies that perform contact tracing—the use of data to try to determine who may have come into contact with COVID-19. Technology-enabled or digital contact tracing has played a conspicuously visible part of the pandemic responses of South Korea, Singapore, Israel, and other nations. Several American and European institutions now propose mobile software apps that crowd-source data in order to track who has been infected by or exposed to COVID-19.

I understand the intuition behind digital contact tracing. But I see the gains in the fight against the virus as unproven and the potential for unintended consequences, misuse, and encroachment on privacy and civil liberties to be significant.

Contact tracing apps generally involve combining self-reported data about health status and location with other sources of data to help users avoid exposure to the novel coronavirus. The idea behind the technology is to inform the public where the risk of contracting COVID-19 is highest and to alert individuals if they may have come into direct or indirect contact with someone who is infected. For example, a

[6] For a discussion, see David Lazer et al., The Parable of Google Flu: Traps in Big Data Analysis, 343 Science 1203 (2014).

[7] Ruha Benjamin, Race After Technology (2019); Safiya Umoja Noble, Algorithms of Oppression (2018); Virginia Eubanks, Automating Inequality (2017); Kate Crawford and Ryan Calo, There is a blind spot in AI research, 538 Nature 311 (October 13, 2016).

person who has download the app would receive a notice if they purchased groceries at a store where an infected person had recently shopped.

The appeal of contact tracing apps is intuitive. Many Americans today face a Hobson's choice: remain at home in isolation, leaving social relations (and the economy) in tatters, or venture out into the world and potentially contract and spread COVID-19. The developers of contact tracing apps hope to offer a third way: safe mobility even in the absence of herd or vaccine immunity by crowd-sourcing the detection and avoidance. Laudable as this goal may be, the technique is unproven and the drawbacks potentially significant.

Today, most household name technology companies—whether Google, Facebook, Twitter, or Uber—are under a consent decree with the Federal Trade Commission for privacy and security lapses, notwithstanding enormous resources. Contact tracing apps collect and combine two highly sensitive categories of information: location and health status. It seems fair to wonder whether these apps, developed by small teams, will be able to keep such sensitive information private and secure. To the extent digital contact tracing—or any private, technology-driven response to the pandemic—involves the sharing of health care data with private parties, there is also the specter of inadequate transparency or consent.[8]

Several digital contact tracing efforts are attending to privacy and security concerns. Of particular note, a large team of European academics recently developed a decentralized platform that safeguards individual privacy and helps guard against government abuse.[9] Even when a system is well-architected from a privacy perspective, however, many pitfalls remain. If participation is voluntary, for example, then communities with few downloaders will look relatively safe just because no one is using the app to report their condition. If health and location status are self-reported, then asymptomatic carriers—who apparently comprise a significant percentage of contagious individuals[10]—will not show up in the results.

It is not hard to imagine nefarious use cases as well. A foreign operative who wished to sow chaos, an unscrupulous political operative who wished to dampen political participation, or a desperate business owner who sought to shut down the competition, all could use self-reported instances of COVID-19 in an anonymous fashion to achieve their goals. The process of threat modeling apps that purport to trace the prevalence of coronavirus is limited or nonexistent.

[8] E.g., Julia Powles & Hal Hodson, Google DeepMind and healthcare in an age of algorithms, 7 Health Technologies 351 (2017).

[9] Carmela Troncoso et al., Decentralized Privacy-Preserving Proximity Tracing (April 3, 2020), draft on file with author.

[10] The director of the Center for Disease Control recently stated that up to twenty five percent of individuals infected with COVID-19 show no discernible symptoms. Apoorva Mandavilli, Infected but Feeling Fine: The Unwitting Coronavirus Spreaders, New York Times (March 31, 2020).

South Korea, Taiwan, Israel, Singapore, and other jurisdictions have apparently used widespread digital contact tracing alongside aggressive investigation and quarantine in order to contain the spread of COVID-19. There is reason to question how important digital technology has been to these efforts; some see widespread availability of testing and early physical distancing measures as the primary way these governments contained the novel coronavirus.[11] But to the extent that technology-based contact tracing has been effective in these jurisdictions, they have not been voluntary, self-reported, or involved self-help. Rather, public officials have forced compliance and dispatched investigators to interview and, if necessary, forcibly quarantine exposed individuals. I see it as an open question whether Americans would be comfortable with this level of state expenditure and intervention. At any rate, the experiences of these nations are not a ready analogy.

There are myriad potential applications of technology to the fight against the novel coronavirus—too many to detail here. Each carries with it a measure of promise and of peril. Perhaps artificial intelligence will be faster than people in identifying pandemic-related misinformation on the internet, but sometimes censor important commentary or information.[12] Perhaps drones can help local authorities safely disperse crowds of people who are not respecting social distancing, but create the impression of a police state in an already frightened population. Perhaps school districts will turn to a technology platform like Zoom to maintain a connection between pupils and teachers, but in the process gather granular commercial intelligence about students on an unparalleled scale. At some level, the question is always the same: does this intervention do enough in the fight against the novel coronavirus to offset its impact on privacy, civil liberties, or other important values? I submit that not all proposed interventions will meet this simple test.

I am not opposed to leveraging every tool in our technical arsenal to address the current pandemic. We are facing a near unprecedented global crisis. I note in conclusion that there will be measures that are appropriate in this context, but not beyond it. Americans and their representatives should be vigilant that whatever techniques we use today to combat coronavirus do not wind up being used tomorrow to address other behaviors or achieve other goals. To paraphrase the late Justice Robert Jackson, a problem with emergency powers is that they tend to kindle emergencies.[13]

[11] E.g., Mark Zastrow, South Korea is reporting intimate details of COVID-19 cases: has it helped?, Nature News (March 18, 2020). One issue is that infected persons may be reticent to self-report if the consequences are greater government surveillance and control. See also Countries are using apps and data networks to keep tabs on the pandemic, The Economist (March 26, 2020).

[12] Elizabeth Dwoskin and Nitasha Tiku, Facebook sent home thousands of human moderators due to the coronavirus. Now the algorithms are in charge, Washington Post (March 24, 2020).

[13] Youngstown Sheet & Tube Company v. Sawyer, 343 U.S. 579, 650 (1952) ("We may also suspect that [the Founders] suspected that emergency powers would tend to kindle emergencies.") (Jackson, J., concurring).

In national security, critics speak in terms of *mission creep*, as when vast surveillance powers conferred to fight terrorism end up being used to enforce against narcotics trafficking or unlawful immigration. In consumer privacy, much thought is given to the prospect of *secondary use,* i.e., the possibility that data collected for one purpose will be used by a company to effectuate a second, more questionable purpose without asking the data subject for additional permissions. No consumer would or should expect that the absence of certain antibodies in their blood, gathered for the purpose of tracing a lethal disease, could lead to higher health insurance premiums down the line. There is also a simpler danger that Americans will become acclimated to more invasive surveillance partnerships between industry and government.[14] My hope is that policymakers will expressly ensure that any accommodations privacy must concede to the pandemic will not outlive the crisis.

Thank you again for the opportunity to testify on this important and pressing issue. I am honored to be able to share these remarks and eager to answer any questions the Committee may have.

[14] In their edited volume Security Games: Surveillance and Control at Mega-Events (2011), Collin Bennett and Kevin Haggerty collect work tending to show that security precautions taken in connection with large events like the Olympics have a tendency to stick around even after the conclusion of the event.