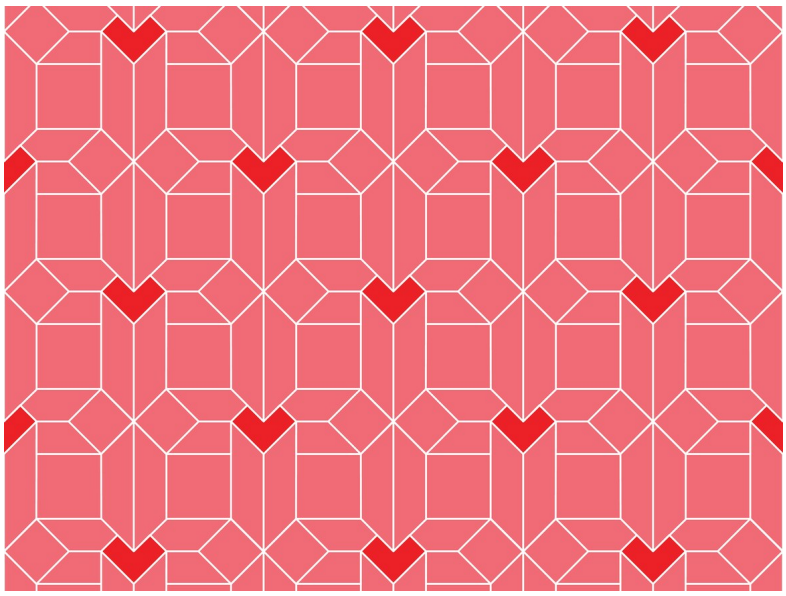
MICHAEL ZIMMER OPINION 05.14.16 07:00 AM

OKCUPID STUDY REVEALS THE PERILS OF BIG-DATA SCIENCE



GETTY IMAGES

ON MAY 8, a group of Danish researchers publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, what kind of relationship (or sex) they're interested in, personality traits, and answers to thousands of profiling questions used by the site.

When asked whether the researchers attempted to anonymize the dataset, Aarhus University graduate student Emil O. W. Kirkegaard, who was lead on the work, replied bluntly: "No. Data is already public." This sentiment is repeated in the accompanying draft paper, "The OKCupid dataset: A very large public dataset of dating site users," posted to the online peer-review forums of *Open Differential Psychology*, an open-access online journal also run by Kirkegaard:

> Some may object to the ethics of gathering and releasing this data. However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form.

For those concerned about privacy, research ethics, and the growing practice of publicly releasing large data sets, this logic of *"but the data is already public"* is an all-too-familiar refrain used to gloss over thorny ethical concerns. The most important, and often least understood, concern is that even if someone knowingly shares a single piece of information, big data analysis can publicize and amplify it in a way the person never intended or agreed.

WIRED OPINION

ABOUT
Michael Zimmer, PhD, is a privacy and Internet ethics scholar. He is an Associate Professor in the School of Information Studies at the University of Wisconsin-Milwaukee, and Director of the Center for Information Policy Research.

The “already public” excuse was used in 2008, when Harvard researchers released the first wave of their “Tastes, Ties and Time” dataset comprising four years’ worth of complete Facebook profile data harvested from the accounts of cohort of 1,700 college students. And it appeared again in 2010, when Pete Warden, a former Apple engineer, exploited a flaw in Facebook’s architecture to amass a database of names, fan pages, and lists of friends for 215 million public Facebook accounts, and announced plans to make his database of over 100 GB of user data publicly available for further academic research. The “publicness” of social media activity is also used to explain why we should not be overly concerned that the Library of Congress intends to archive and make available all public Twitter activity.

Public Does Not Equal Consent

In each of these cases, researchers hoped to advance our understanding of a phenomenon by making publicly available large datasets of user information they considered already in the public domain. As Kirkegaard stated: “Data is already public.” No harm, no ethical foul right?

Wrong.

Many of the basic requirements of research ethics—protecting the privacy of subjects, obtaining informed consent, maintaining the confidentiality of any data collected, minimizing harm—are not sufficiently addressed in this scenario.

Moreover, it remains unclear whether the OkCupid profiles scraped by Kirkegaard’s team really were publicly accessible. Their paper reveals that initially they designed a bot to scrape profile data, but that this first method was dropped because it was “a decidedly non-random approach to find users to scrape because it selected users that were suggested to the profile the bot was using.” This implies that the researchers created an OkCupid profile from which to access the data and run the scraping bot. Since OkCupid users have the option to restrict the visibility of their profiles to logged-in users only, it is likely the researchers collected—and subsequently released—profiles that were intended to *not* be publicly viewable. The final methodology used to access the data is not fully explained in the article, and the question of whether the researchers respected the privacy intentions of 70,000 people who used OkCupid remains unanswered.

There Must Be Guidelines

I contacted Kirkegaard with a set of questions to clarify the methods used to gather this dataset, since internet research ethics is my area of study. While he replied, so far he has refused to answer my questions or engage in a meaningful discussion (he is currently at a conference in London). Numerous posts interrogating the ethical dimensions of the research methodology have been removed from the OpenPsych.net open peer-review forum for the draft article, since they constitute, in Kirkegaard’s eyes, “non-scientific discussion.” (It should be noted that Kirkegaard is one of the authors of the article *and* the moderator of the forum intended to provide open peer-review of the research.) When contacted by Motherboard for comment, Kirkegaard was dismissive, stating he “would like to wait until the heat has declined a bit before doing any interviews. Not to fan the flames on the social justice warriors.”

scientists. Rather, we should highlight this episode as one among the growing list of big data research projects that rely on some notion of "public" social media data, yet ultimately fail to stand up to ethical scrutiny. The Harvard "Tastes, Ties, and Time" dataset is no longer publicly accessible. Peter Warden ultimately destroyed his data. And it appears Kirkegaard, at least for the time being, has removed the OkCupid data from his open repository. There are serious ethical issues that big data scientists must be willing to address head on—and head on early enough in the research to avoid unintentionally hurting people caught up in the data dragnet.

In my critique of the Harvard Facebook study from 2010, I warned:

> The...research project might very well be ushering in "a new way of doing social science," but it is our responsibility as scholars to ensure our research methods and processes remain rooted in long-standing ethical practices. Concerns over consent, privacy and anonymity do not disappear simply because subjects participate in online social networks; rather, they become even more important.

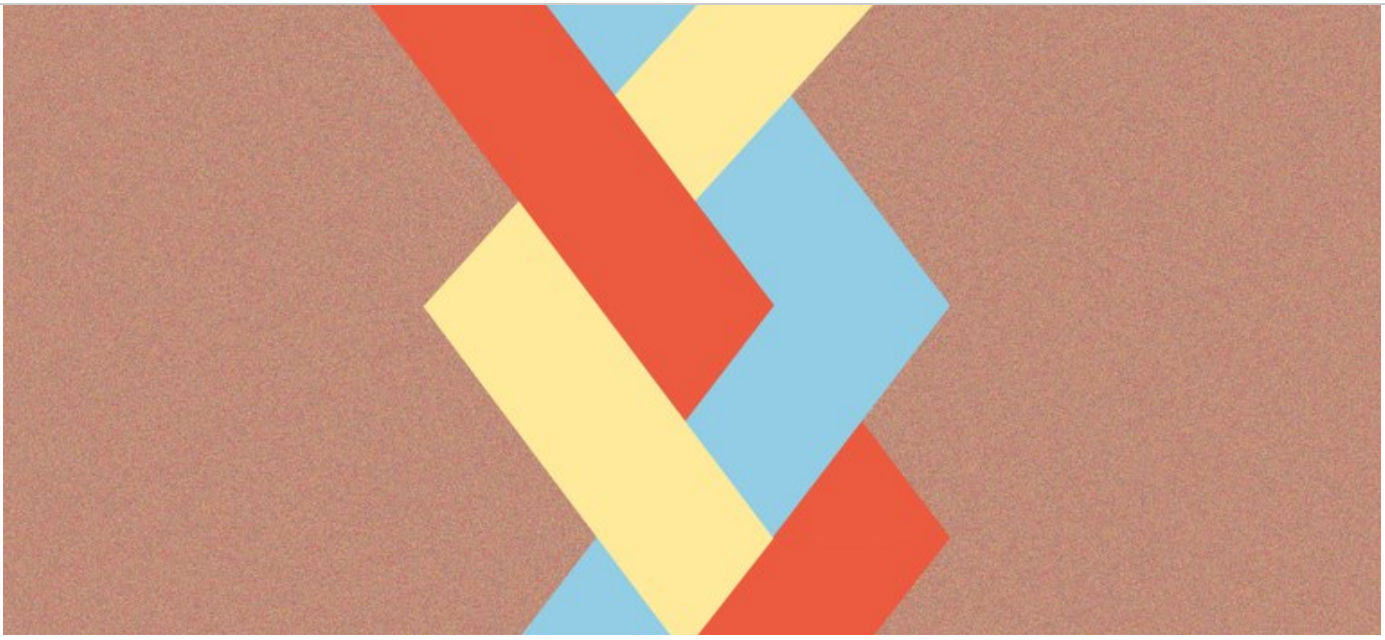
Six years later, this warning remains true. The OkCupid data release reminds us that the ethical, research, and regulatory communities must work together to find consensus and minimize harm. We must address the conceptual muddles present in big data research. We must reframe the inherent ethical dilemmas in these projects. We must expand educational and outreach efforts. And we must continue to develop policy guidance focused on the unique challenges of big data studies. That is the only way can ensure innovative research—like the kind Kirkegaard hopes to pursue—can take place while protecting the rights of people an the ethical integrity of research broadly.

#DATING #WIRED OPINION

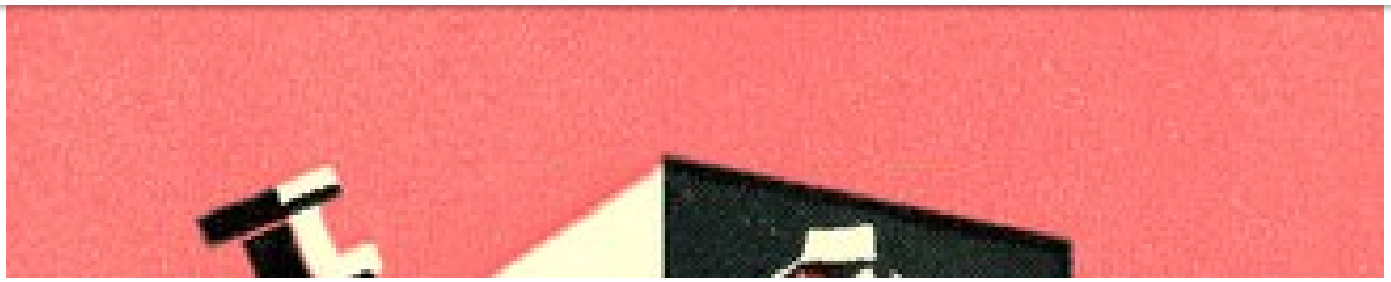
MICHAEL DEMPSEY
How the US Can Counter Threats from DIY Weapons and Automation

CHELSEA BARABAS
Decentralized Social Networks Sound Great. Too Bad They'll Never Work

MORE OPINION



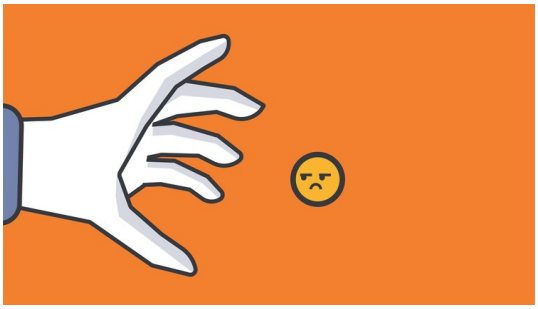
WIRED OPINION
The World is Ruled by Net States. Ignore Them at Your Peril.
ALEXIS WICHOWSKI



WIRED OPINION

Workers Displaced by Machines Could Become Caregivers for Humans

OREN ETZIONI



OPINION

The Solution to Too Much Facebook Isn't More Facebook

ANTONIO GARCÍA MARTÍNEZ

WIRED OPINION

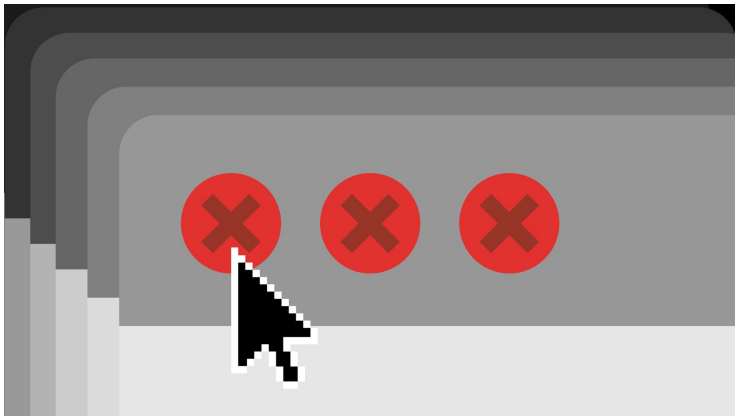
Digital Solutions Can Help Even the Poorest Nations Prosper

BJORN LOMBORG

WIRED OPINION

Supreme Court's Cell Phone Tracking Case Could Hurt Privacy

NICK SIBILLA



WIRED OPINION

How Federal Law Protects Online Sex Traffickers

ROB PORTMAN