TRANSNATIONAL INTERNET

Preservation of Freedoms and Privacy

VINTON G. CERF

Before the Internet there was the ARPANET (Advanced Research Projects Agency Network), a US Defence department project to implement packet-switching technology in support of computer communication. Launched in late 1969, one of the ARPANET's first popular applications to emerge in 1971 was networked electronic mail ('e-mail') and associated mailing lists. The general public did not have access to the ARPANET, as it was reserved for researchers supported by the US Defense Advanced Research Projects Agency (DARPA). There emerged, however, experimentation with Bulletin Board Systems, and with Usenet[1] in the late 1970s and early 1980s, for example. These were public information-sharing systems containing content generated by the users, and fostered dialogue, discourse, debate and the production of a very wide range of content.

The arrival of the Internet in 1983 and the World Wide Web (WWW) in 1991 signalled a wave of new media development. Web pages led the way to weblogs ('blogs'). All older media—print, imagery, sound and video—could be transported on the Internet/Web.

One of the surprising aspects of the arrival of the Web was the phenomenal sharing of information on the platform. People generated vast quantities of content on web pages, blogs, commentaries and e-mail. So much was produced that search engines were developed to help find content of interest. An advertising model evolved in which users were provided with free access to tools and content in exchange for viewing advertisements posted on web pages or associated with responses to search queries.

Winter 2018–Spring 2019, Volume 45, Numbers 3 & 4

For a time, ads were even shown to users who were reading their e-mail, but that particular practice seems to have dissipated over time. For the most part, users were motivated to share their information, not for pecuniary gain, but simply because it was gratifying to know that what you shared was useful to others.

The arrival of social media, such as Facebook, Twitter, YouTube, Instagram, among others, opened up a new avalanche of user-generated content and commentary. These media were supported by advertising business models. Users got to use the platforms for free, while advertisers paid the platform operators for the privilege of showing ads to the users. Eventually, the users and generators of content were also allowed to share in some of the revenue generated by the ads shown in association with user-generated content.

These so-called new media sparked considerable commentary from viewers, not all of it constructive. In addition, metrics were offered to the generators of content: 'likes' on Facebook, followers on Twitter and views on YouTube. Users were *rewarded* by these metrics and this feedback loop led to behaviours that may not have been anticipated by the new media creators. To achieve higher metric scores, users were given incentives to produce more and more extreme content. Extremism gets attention. In the news world, a common recognition of this phenomenon is found in the aphorism: 'If it bleeds, it leads.' Looking back, it is not so surprising that extreme content or comments can be found in these media. Indeed there is even a pecuniary motivation if the higher level of attention leads to larger advertising revenue that is shared by the user generating the content. The more controversial a web page, blog, video or tweet, the more attention it gets, and the more ads it may expose to viewers and the more revenue it generates. Even in the absence of revenue generation, high levels of attention, measured in the metrics of social media, can drive content towards extremism.

Consequently, we are seeing in real time the evolution of an online social phenomenon with real-world consequences. As of this writing, the International Telecommunications Union estimates that 50 per cent of the world's population is now online, many of them by means of smart phones. An apparently increasing number of these users get *news* by way of social media rather than traditional print, radio, television or even online news sources. The social

media are not typically subject to historical journalistic and publishing guidelines. They are frequently hives of misinformation, disinformation, extremism and closed bubble effects. There exists strong evidence that nation states have used these media to achieve political ends, disrupt normative processes, foster political tension and division, incite violence, interfere with elections, and degrade trust in critical societal institutions. In some ways, this phenomenon is an extension of the effects of cable television with its hundreds (thousands?) of channels catering to special interests and splinter groups.

Another phenomenon associated with online media must also be taken into consideration. Computers have become tools that empower and amplify human capabilities. No human could possibly search through the body of the Web to find useful information in any reasonable period of time. It is the power of computing that allows search engines to sift through billions of Web pages to index them and to help users discover content of interest. There are, however, more nefarious ways of employing computers to exploit their amplifying capacity. Programmable devices of all kinds (servers, desktops, laptops, mobiles, pads, Internet-enabled appliances, etc.) can be taken over by hackers who then use the computing power to generate spam e-mail, malware distribution, denial of service attacks and a host of other harmful practices. Thanks to this amplification effect, a single user or small group can exercise the same power as a nation state in the online world.

Weak operating systems and applications that cannot withstand determined hacking attacks are the primary reason such robot armies can be created. These so-called 'botnets' can be used to create and animate social media identities, emulating human users sufficiently well that it is hard to tell them from real users. Botnets pose extremely serious problems, not only in the social media space, but more generally in the context of computer and network security. Critical infrastructure that depends on computing and networking is at risk from serious and massive or extremely subtle attack. While that is not the primary topic of this essay, it is worth noting that it has proven to be extremely difficult to inhibit the creation and use of bots in all kinds of contexts. To make matters worse, botnet attacks cross international borders with impunity, partly because the Internet is largely insensitive to such boundaries.

Victims can be in one country and perpetrators in another, making mitigation complicated.

The matter has become so visible that in the United States, hearings have been held to expose the problems and to try to hold social media operators to account for abusive content. Of course, the definition of 'abusive content' may vary dramatically depending on the party defining the term. Demands that private sector companies censor content according to an often unspecified standard collides with freedom of expression so valued in the Universal Declaration of Human Rights that recently celebrated its 70th anniversary. From time to time, it is proposed that Artificial Intelligence and Machine Learning tools will be able to winnow the wheat from the chaff, but this is far from assured.

First, these algorithms usually need to be 'trained' against a body of content so as to correctly detect unwanted material. But the training content may have built-in biases, producing results that are not considered acceptable by all the parties calling for filtering. These algorithms are also subject to a kind of brittleness in which content that was intended to have been filtered is not, and vice versa. Second, there are proposals to use crowd-sourcing to apply the so-called 'wisdom of crowds' to distinguish acceptable from unacceptable content. But these proposals fail in the face of bots that are indistinguishable from human users.

Readers may recall the famous 'Turing Test' in which a person is exchanging text messages with a human and a computer. The person conducting the test tries to distinguish the human from the computer solely by evaluating the responses received. If the tester cannot distinguish between the computer and the human, the computer is said to have passed the Turing Test. Now let us imagine a different situation in which a computer programme is exchanging messages with another computer and with a human. The task of the computing programme is to distinguish between the computer and the human. If it fails to make the distinction, we can say that the programme conducting the test has failed what we will call Turing Test Two.

This is not a purely speculative scenario. Social media platforms, in an effort to filter unwanted content, may perceive botnet-produced content as coming from real users and reach wrong conclusions about crowd-sourced opinion. The bots confirm that

which is not true, in effect. This leads to the natural question, 'What is to be done?'

While I do not offer solutions in this brief essay, I would proffer a few thoughts for consideration. First, it does not seem appropriate to ask the private sector to filter user-generated content without guidance. Some platform operators have Terms of Service that allow them to exclude some content as a condition of use by users. Nudity, incitement of violence, racist expression and denial of the Holocaust are examples of such voluntary prohibitions. Second, operators might benefit from a broader harmonisation of guidance, allowing them to distinguish acceptable from unacceptable content. If not global then perhaps at least national guidance might be helpful. There is risk that such guidance may conflict in serious ways with the concepts of freedom of expression and freedom of access to information. Indeed, attempts to dictate filtering may reinforce or provide justification for authoritarian regimes claiming that their censorship is simply a reasonable tool for maintaining public safety.

We are presented with a conundrum. Human rights ought to be preserved as much in cyberspace as in the physical world. Among those rights is the right to privacy, and we see efforts to enforce this with such legislation as the European General Data Protection Regulation (GDPR) that has effect not only in Europe, but extraterritorially wherever personal information about a citizen of the European Union is held. At the same time, nation states also express the need to combat malicious behaviour, often amplified through botnets, even when the behaviour crosses national borders. Discovering the perpetrators of harm often falls to law enforcement and a tension can then arise between protection of privacy and protection of citizens from online harm.

It seems clear that the struggle to protect users from online harm, while protecting other rights, will prove to be a thorny matter, especially in a global context. A number of commissions and panels have been created in a wide range of contexts to seek solutions to these problems. Among them is the High Level Panel on Digital Cooperation set up by the Secretary General of the United Nations. There is also a Global Commission on the Stability of Cyberspace. These are but two of many efforts to wrestle with these challenges and to produce actionable recommendations. It seems to me essential that these problems be viewed through a multi-stakeholder lens.

So many parties are at interest in government, the private sector, civil society and the technical community, that only a concerted effort to explore objectives, means and transnational cooperation has a chance of success.

If solutions emerge at all, I am persuaded that they will include more attack-resistant software systems, wider agreement on trans-border law enforcement, better tools to allow users to protect themselves (e.g., two-factor identity authentication), broader agreement on the definition of harmful content, and application of individual critical thinking to the diverse content discoverable on the Internet.

◆

NOTE

1. https://en.wikipedia.org/wiki/Usenet.

◆◆