Douwe Korff
Emeritus Professor of International Law, London Metropolitan University
Associate, Oxford Martin School, University of Oxford; Fellow, Centre for Internet & Human Rights, Eur. Univ. Viadrina, Berlin
http://douwe.korff.co.uk

General monitoring of communications in order to block "undesirable" Internet content*

Increasingly, demands are made that "something be done" about "undesirable" and "harmful" material on the Internet: online child abuse images and other criminal pornography; "extremism"; "incitement to violence"; "hate speech"; - and more recently, "fake news". Organisations representing holders of intellectual property rights similarly demand that measures be taken to prevent the sharing of IP-protected materials online. There is a widespread assumption that the "Internet Giants" (Google, Apple, Facebook, Twitter) have the means and resources to meet these demands, and should be forced to do so.[1]

Yet it is problematic to leave such action to the discretion of the companies that provide the platforms for the sharing of such content, both in terms of standards and in terms of processes.[2] (Although leaving it to states to determine what is "true" or "fake" news is also problematic: over the years, states themselves, and state agencies, and political parties, have all been engaged in lies and black propaganda.)[3]

Special problems arise in relation to what the Special Rapporteur on Freedom of Expression, David Kaye, calls "automation or algorithmic filtering".[4] The typical example given is the use of Microsoft's "PhotoDNA" to detect online child abuse pictures.[5] However, there are two main problems with such tools. First of all, as Peter Sommer explains in a recent blog, while artificial intelligence (AI) tools may be able to identify materials that have already been judged to constitute online child abuse pictures, by matching hashes (albeit it not even then if the material is modified), it is much more difficult to use them to try and identify text or images that *may* constitute objectionable or even illegal content, i.e., where some judgment is required. In such cases, the context will often be critical:[6]

> The same photo may appear on a site promoted by a terrorist group and by a news organisation. Some sexually explicit photos may be justified in the context of medical and educational research – or law enforcement. …
>
> How do you reliably distinguish a 16-year-old from an 18-year-old, and for all ethnicities? How does an AI system distinguish the artistic from the exploitative or when in a sexual situation there is an absence of consent?

This applies also, especially, to automated systems that use so-called natural language processing (NLP) tools for analysing the text of social media posts. As the Center for Democracy puts it:[7]

> Today's tools for automating social media content analysis have limited ability to parse the nuanced meaning of human communication, or to detect the intent or motivation of the speaker. Policymakers must understand these limitations before endorsing or adopting automate content analysis tools. Without proper safeguards, these tools can facilitate overbroad censorship and biased enforcement of laws and of platforms' terms of service.

* This paper is an excerpt from a wider paper, First do no harm: The potential of harm being caused to fundamental rights and freedoms by state cybersecurity interventions, February 2018, to be included in Ben Wagner, Matthias C. Kettemann and Kilian Vieth, Research Handbook on Human Rights & Digital Technology: Global Politics, Law & International Relations, CIHR, Berlin, due for publication later in 2018.

Unfortunately, companies keep trying to sell snake oil tools to governments, such as a tool which, it is claimed, "can detect 94% of Isis propaganda with a 99.99% success rate in tests" – and politicians keep buying such impossible claims (and such useless snake oil).[8]

Even when the standard is clear, the judgment may be difficult. Depictions of people having sex with animals are clearly defined as illegal in UK law. But does this mean all historical or even recent depictions of Leda and the Swan (and other depictions of mythological god/human-beast encounters) must be blocked? (Leda and her swan are currently not blocked from Google searches.)

L'HISTOIRE
D'UN PECHE
DZIEJE GRZECHU
P

(Poster by Jerzy Flisak for Walerian Borowczyk's 1975 film *The Story of a Sin*)

In relation to IP-protected material, there is the special, major problem of limits to and exceptions from such protection, e.g., in relation to fair comment/reporting, criticism, parody, caricature and pastiche, or to facilitate access to people with disabilities. The scope and application of those exceptions is difficult to determine in individual cases by lawyers and courts – and well beyond the capabilities of so-called "artificial intelligence" and NLP.

*Put simply: because of their unavoidable limitations and incapabilities, "algorithmic filter" tools are **inherently inappropriate** for the purpose of determining whether speech or text (or pictures or videos) amounts to "hate speech", "incitement to violence", "support for extremism", etc. – or whether, if it seems to comprise copyright-protected materials, their use is covered by one of the exceptions in the applicable national law (or indeed to determine which law that is).*

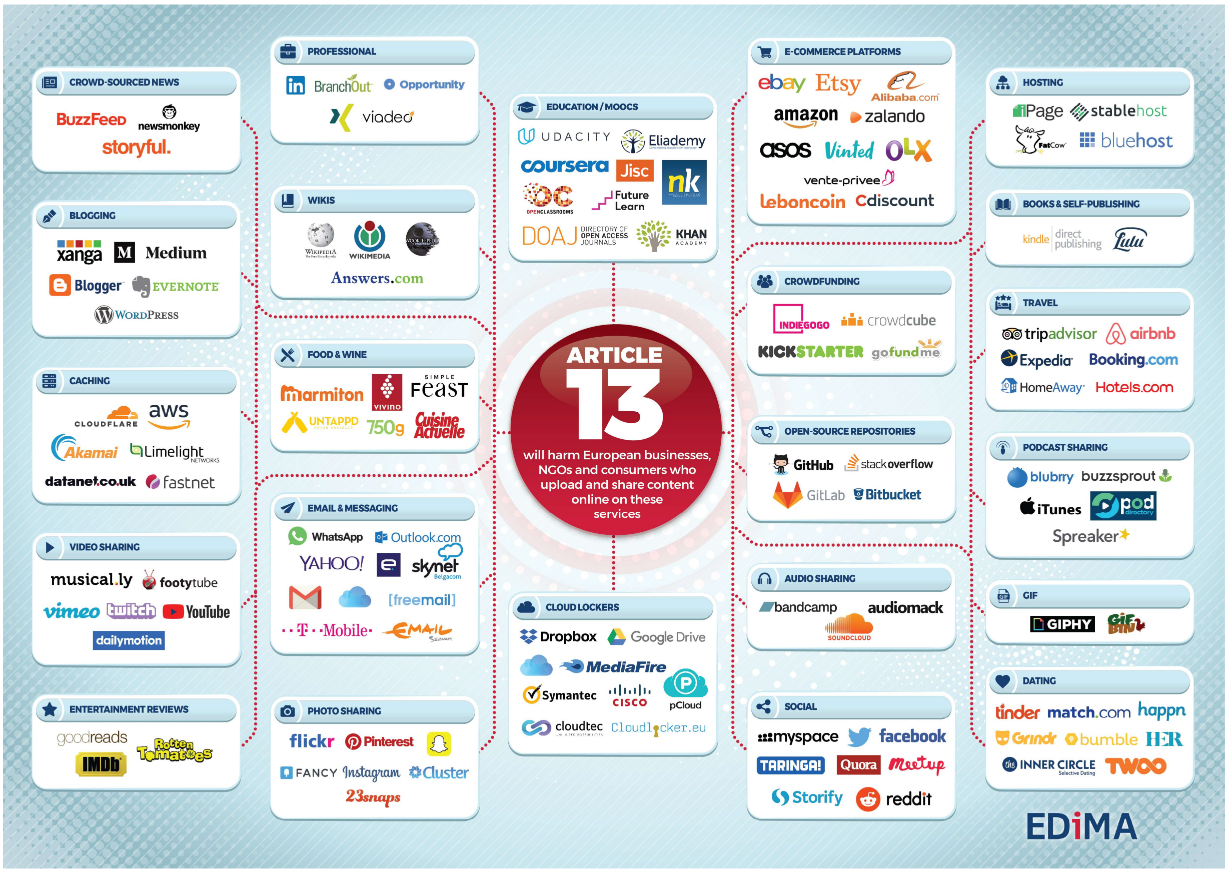
Yet astonishingly, the European Commission is proposing that precisely those tools are to be used by information society service providers to detect copyright-protected materials and "prevent" them from being made available on their sites.

Specifically, Article 13 of the proposed Copyright Directive,[9] if adopted as proposed by the Commission, would impose on all such providers a legal duty to conclude "agreements" with rightholders (in practice, royalty collecting agencies) under which those service providers must "prevent the availability on their services of works or other subject-matter identified by rightholders" (read: as protected by copyright). In other words, they will be required by law to implement a system to block copyright-protected materials (as identified by rightsholders).

The proposal is seemingly less prescriptive when it comes to the **means** to be used to achieve such blocking: it appears to say, in quite general terms, that "[t]hose measures" – i.e., the measures used to block the relevant materials – "shall be appropriate and proportionate". But this is quite simply disingenuous, as is clear from the only example of such measures mentioned in the text of the draft directive (Art. 13(2)): "effective **content recognition technologies**".[10]

In fact, the *only* way to "prevent the availability" of – i.e., to preemptively block – copyright-protected content on the relevant platforms is to use such algorithmic filters. This would affect a vast multitude of services, as illustrated by the infographic, below.

INFOGRAPHIC



Source: http://edima-eu.org/wp-content/uploads/2018/01/Services-affected-by-Article-13-Infographic.jpg

Moreover, such tools can of course be used to preventively detect, and then block, any pre-determined content. They are a gift to any government wanting to suppress the free flow and sharing of information on the Internet.

Not surprisingly, European Digital Rights calls such tools "**Censorship Machines**".[11]

The Global Network Initiative (whose members include Facebook, Google, LinkedIn and Microsoft as well as human rights organisations)[12] has adopted the view that:[13]

> Governments should not mandate the use of filters or other automated content evaluation tools in laws regulating speech.

In its judgments in the *Digital Rights Ireland* and *Safe Harbor* cases,[14] the Court of Justice of the EU (CJEU) made clear that:[15]

> Legislation is not limited to what is strictly necessary [and thus incompatible with the EU Charter of Fundamental Rights] where it authorises, on a generalised basis, storage of all the personal data of [large numbers of people] without any differentiation, limitation or exception being made in the light of the objective pursued and without an objective criterion being laid down by which to determine the limits of the access of the public authorities to the data, and of its subsequent use, for purposes which are specific, strictly restricted and capable of justifying the interference which both access to that data and its use entail.
>
> In particular, legislation permitting the public authorities to have access on a generalised basis to the content of electronic communications must be regarded as compromising the essence of the fundamental right to respect for private life, as guaranteed by Article 7 of the Charter.

The algorithmic filters promoted by the Commission – and effectively required by Article 13 – would institutionalise precisely the kind of "generalised" monitoring of Internet content that the Court mentions. But surely, if states are not allowed, under the Charter, to grant their own agencies powers of "generalised" monitoring, because that would violate the very "essence" of the right to private life, then those same states should also not be allowed to allow – indeed, insidiously but effectively *require* – private companies to carry out such monitoring.

In sum: The use of "algorithmic filters" (or "content recognition-" and/or "content evaluation technologies") in order to detect and block objectionable or copyright-protected content in private sharing platforms must, because of the very function they are to fulfil, involve the very ***"generalised" monitoring*** *of the content of communications of whole populations that the Court of Justice of the EU denounced as incompatible with the EU Charter of Fundamental Rights in relation to the mass surveillance by states.*

Such tools are therefore both inappropriate for their stated aim and constitute major and disproportionate – and thus unlawful – interferences with the fundamental rights of the people in the populations against which they are used.

- o – O – o -

Douwe Korff (Em. Prof.)
Cambridge, February 2018

NOTES:

[1] Jeremy Darroch*, It's time the internet giants were made accountable*, Times, 28 November 2017, available at:
https://www.thetimes.co.uk/article/it-s-time-the-internet-giants-were-made-accountable-2w9r3brvd

[2] See the UN Special Rapporteur on Freedom of Expression, David Kaye's, "Concept Note" on Content Regulation in the Digital Age, prepared for his June 2018 Human Rights Council Report, available at:
https://freedex.org/wp-content/blogs.dir/2015/files/2017/09/Concept-Note-Social-Media-Search-and-FOE.pdf

European Digital Rights has noted problems relating to the deletion of perfectly legal content by providers relying on their Terms of Use; to overreliance on "trusted flaggers" of contentious content; to "review processes, record-keeping, assessing counterproductive effects, anti-competitive effects, over-deletion of content, complaints mechanisms for over-deletion"; and to the lack of "investigation or prosecutions of the serious crimes behind, for example, child abuse." See:
https://edri.org/leaked-document-does-the-eu-commission-actually-aim-to-tackle-illegal-content-online/.
See also this subsequent article:
https://edri.org/commissions-position-tackling-illegal-content-online-contradictory-dangerous-free-speech/.
On the problems in relation to "fake news", see: Tarlach McGonagle, ''Fake news'': False fears or real concerns?, Netherlands Quarterly of Human Rights 2017, Vol. 35(4) 203–209, available at:
http://journals.sagepub.com/doi/pdf/10.1177/0924051917738685

[3] See Kenan Malik, *Fake news has a long history. Beware the state being keeper of 'the truth'*, *Observer*, 11 February 2018, available at:
https://www.theguardian.com/commentisfree/2018/feb/11/fake-news-long-history-beware-state-involvement?

[4] See the UN Special Rapporteur on Freedom of Expression, David Kaye's, "Concept Note" on Content Regulation in the Digital Age, prepared for his June 2018 Human Rights Council Report, available at:
https://freedex.org/wp-content/blogs.dir/2015/files/2017/09/Concept-Note-Social-Media-Search-and-FOE.pdf

[5] See:
https://www.microsoft.com/en-us/photodna

[6] Peter Sommer, *blogspot*, 3 February 2018, available at:
https://pmsommer.blogspot.co.uk/.

[7] Mixed Messages? The Limits of Automated Social Media Content Analysis, Center for Democracy & Technology, November 2017, available at:
https://cdt.org/files/2017/11/Mixed-Messages-Paper.pdf.

[8] *Home Office unveils AI program to tackle Isis online propaganda*, *Guardian*, 13 February 2018, available at:
https://www.theguardian.com/uk-news/2018/feb/13/home-office-unveils-ai-program-to-tackle-isis-online-propaganda?CMP=Share_iOSApp_Other
Some further information on this magic tool is provided in the *Independent* of the same day: *Artificial intelligence 'can detect Isis videos even before they are uploaded'*, available at:
https://edition.independent.co.uk/editions/uk.co.independent.issue.130218/data/8207326/index.html
Typically, the impossible claim of a "99.99% success rate" is illiterate in scientific terms: in pattern recognition science, the outcomes should be measures in terms of *precision* and *recall*, rather than "accuracy", see:
https://en.wikipedia.org/wiki/Precision_and_recall

[9] Proposal for a Directive of the European Parliament and of The Council on copyright in the Digital Single Market, COM/2016/0593 final - 2016/0280 (COD), Brussels, 14 September 2016, available at:
http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52016PC0593

[10] Interestingly, a new "possible" text of Article 13, contained in an EU Presidency Discussion Paper on Art 11 and Article 13 [of the proposed Copyright Directive], distributed to Member State delegations on 6 February 2018, omits the "example" of "effective content recognition technologies":
https://www.parlament.gv.at/PAKT/EU/XXVI/EU/01/03/EU_10322/imfname_10784225.pdf

(see the alternative text on p. 13). This is likely a further disingenuous attempt to effectively hide the issue: the alternative "possible" text still retains the requirement that "online content sharing services" must "take effective measures to **prevent the availability** on its services of these unauthorised works or other subject-matter identified by the rightholders" (emphasis added), even though, as noted next in the main text of the present paper, in practice the *only* way of achieving this is to use algorithmic filters.

[11] See:
https://edri.org/civil-society-calls-for-the-deletion-of-the-censorshipmachine/
The European Commission belatedly responded to the open letter sent by EDRi and 56 other civil society organisations on 1 February 2018:
https://edri.org/files/copyright/20180201-EC_ReplyOpenLetter-Art13.pdf
For criticism of this response and of the Commission's continuing attempt to effectively require content-screening filters, see EDRi*,* 9 February 2018, at:
https://edri.org/smashing-the-law-without-breaking-it-a-commission-guide/

[12] "GNI is a unique multi-stakeholder forum bringing together Information and Communications Technology (ICT) companies, civil society organizations, investors, and academics to forge a common approach to protecting and advancing free expression and privacy online." See:
https://www.globalnetworkinitiative.org/
https://globalnetworkinitiative.org/participants/index.php

[13] GNI Submission to the UN Special Rapporteur on the right to freedom of opinion and expression, David Kaye, on Content Regulation in the Digital Age, 20 December 2017, p. 8, available at:
https://globalnetworkinitiative.org/sites/default/files/GNI-Submission-SR-Report-Content-Regulation.pdf
See also:
https://globalnetworkinitiative.org/news/gni-provides-input-un-report-content-regulation-digital-age.

[14] Respectively: CJEU Judgment in *Digital Rights Ireland*, C-293/12, 8 April 2014; CJEU Judgment in *Schrems*, C-362/14, 6 October 2015. The main considerations of the Court in these cases are summarised in Douwe Korff, Ben Wagner, Julia Powles, Renata Avila and Ulf Buermeyer, Boundaries of Law: Exploring Transparency, Accountability, and Oversight of Government Surveillance Regimes, global comparative report covering Colombia, DR Congo, Egypt, France, Germany, India, Kenya, Myanmar, Pakistan, Russia, South Africa, Turkey, UK, USA, prepared for the World Wide Web Foundation, January 2017, pp. 23 – 25, under the heading "*Case Law of the CJEU*". available at:
https://ssrn.com/abstract=2894490.

[15] The quote is from para. 93 of the *Schrems/Safe Harbor* judgment, echoing paras. 54ff. of the *Digital Rights Ireland* judgment. In the *Digital Rights Ireland* case, the population in question was all citizens of the EU; in the *Schrems/Safe Harbor* case it was all the persons whose data has been transferred from the European Union to the United States under the *Safe Harbor* agreement between the EU and the USA.

- o – O – o -