Chapter 18
Algorithmic Censorship of Online Visual Content – Ethical, Political, and Economic Rationales

Eberhard Ortland

I. Introduction

Images can be powerful, exciting or incisive, appealing or appalling and therefore, sometimes controversial. Some images move us to act, some call for comments, and some leave us speechless. But some images can also be harmful. An encounter with certain images may disturb some or most of the people who see them. What has been seen cannot be unseen again. The circulation and accessibility of certain images may also be offensive to people who do not even feel offended themselves, but who, for certain reasons, do not want others to see those images, at least not unless they meet certain conditions. The accessibility of some images may violate legitimate interests – of the persons depicted in the image or of people who are not even depicted in the image but whose relatives, friends, idols or belongings are shown in a way they would not like to share with others, or in a way they find objectionable for some reason. There may also be economic or political reasons why some people do not want certain images to be freely accessible to anybody, or why they want to subject the availability of certain images to certain conditions.

Arguably, some of these reasons can be regarded as legitimate, while some may be contested. But how and according to which sets of criteria can we decide which of the reasons cited for limiting the production, distribution or accessibility of certain images ought to be accepted as legitimate and which should be rejected – either in general, or at least under certain circumstances – to preserve some generally recognized, superior good? And how should we deal with conflicts about the circulation and accessibility of certain images that cannot be resolved on common grounds?

Conflicts about the circulation and accessibility of certain images increase considerably with the spreading of digital media and communication technologies, for obvious reasons: There are now exponentially more, and more diverse, images than ever before in human history. The number of people involved in making or taking images as well as the number

of automatically operating cameras have likewise increased in totally unprecedented ways. Digital computer networks allow images to circulate around the world in light speed on countless channels. Thus, they are now accessible to unprecedented audiences from most heterogeneous cultural backgrounds. The global circulation of images brings about countless occasions of cross-cultural short circuit experiences, when certain audiences take offence at images that were considered completely innocent, rather funny, or maybe daring, but definitively legal by those who produced them or by those who liked them and distributed them online because they wanted to share them with others.

The central question is: Who shall decide, according to which criteria, under which circumstances which images may or may not be shown? While it is certainly true that these issues are at the centre of any debate on the ethics of images, it is, however, rather surprising that it is hardly ever made explicit that all the activities deemed necessary to avoid or reduce conflicts about images – such as content moderation or some sort of regulation – inevitably involve some form of censorship.

II. Censorship

Censorship is a highly contentious issue. It has been contentious for centuries. It was officially abolished in most European states and in the so-called free world since the American and the French Revolution. However, the process was lengthy and included severe set-backs and was not accomplished in most countries before the middle or even the end of the twentieth century. Yet, some sort of censorship was and continues to be executed in many countries most of the time.

I have previously argued that censorship "is not only the suppression of speech, public communication, or other kinds of expression or information … by the government or church authorities". Rather, "censorship occurs whenever people, institutions or organizations succeed in imposing their political or moral values or particular interests on others by suppressing the circulation of certain words, images, or ideas that they find offensive or otherwise objectionable".[1] Such a broad definition of censorship is useful and necessary given today's fragmented and multi-layered political, sectarian and corporate powers.

---

1 Ortland (2018) 129.

Of course, it is important that state authorities do not directly and officially control the mass media in countries where democracy and the freedom of the press flourish. However, direct intervention by state authorities is not the only kind of censorship to be concerned about. It does not help us very much if we may trust that the government will refrain from censoring the mass media, as long as other actors or agencies are legally entitled or even obliged to control and restrict the public circulation of certain types of information or expression, words, images or sound recordings. I think it is time that we face the reality of censorship and start a realistic discussion about the various types of censorship as well as the reasons why certain kinds of censorship are regarded as desirable and legitimate, at least under certain circumstances, while others are not.[2]

Most discussions about censorship focus primarily or exclusively on the interruption and suppression of communication, and on the prohibition of certain words, signs or images. These are the cases when conflicts about the right to communicate or to express oneself become most strikingly apparent. However, as we have learned from Foucault,[3] it is important to understand that power is not necessarily and not even primarily repressive, but that it is always directed at certain positive goals: at influencing the way people behave, how they see things and what they regard as desirable or detestable. To understand the impact of censorship, we must be aware that in most cases the intervention of the censor does not lead to the suppression of objectionable content, but rather to the approval of all kinds of expression deemed acceptable.

Censorship is a filter. The purpose of any filtering is to separate the filtrate that can pass from the residue that is retained or extracted from the filtered mixture. Sometimes we filter because we are interested in the residue, usually we want the filtrate to be clear of certain stuff. Censorship is driven by a strong negative interest in certain types of expression that the censor – or those who have installed the censor – do not want to circulate in public. Yet the aim of censorship is not to retain large chunks of ongoing communication, but rather to direct people to use only expressions which can be expected to be acceptable under the censors' critical eyes.

Thus, censorship can be regarded, in a way, as a special type of co-authorship imposed on the primary author by a co-author who is typically

---

2 For a discussion of different types of censorship see Jones (2001); Green/Karolides (2005); Mathiesen (2008); Ortland (2018) 129–136; Heldt (2019) §§ 9–12 and 19–24.

3 Foucault (1975).

absent and is perhaps not even a fellow human being but rather the anonymous and elusive "powers that be". The censor rarely discusses the wanted results with the primary author, but silently demands that the expression submitted by the author meets certain expectations to be approved.[4] Under a censorship regime, authorship becomes shady and the expression, corrupt. People must learn to read, whatever they get to read, with double vision. They distrust the surface, search for hidden meanings, try to guess what the author might have wanted to say but had to avoid or camouflage to pass the censorship filter.

III. Reasons for Censorship

Censorship is deeply ambivalent, neither always bad nor simply good. Obviously, censorship stifles free speech. It distorts the way we express ourselves (or whatever). It creates or reinforces power imbalances. And it channels, in often problematic ways, the stream of communication that is vital for each of us. We may be appalled by such distortions and power imbalances, and often rightly so. But then we must ask if we could really imagine a society that allows all kinds of expression – words, images, sounds and information about anything and anybody that anybody wants to share with others, or to throw at others – to circulate without any restrictions.

There are reasons why some kind of censorship is deemed necessary under certain circumstances. Some of these reasons are political and will always be controversial. As the saying goes, one person's or state's terrorist is another person's or state's freedom fighter.[5] Other reasons are mainly economic, for example concerning the enforcement of copyright claims or trademarks.[6] But there are, I think, also serious ethical or moral concerns, rooted in normative ideas about rights we think we have or ought to have, and in ideas about what is relevant for a good life, for men, women and children, living together more or less closely but not always peacefully. We must recognise and analyse these reasons,[7] and figure out how they relate to the political and economic rationales that are also relevant for both the legitimacy, and the limits of legitimacy, of certain kinds of censorship.

---

4 Steinhauer (forthcoming).
5 Ganor (2002).
6 Aufderheide (2020).
7 Ortland (2018).

Even fervent advocates of free speech[8] had to concede that some kinds of images, texts or other kinds of content that can be easily uploaded and thus made accessible to the public by virtually anybody at any time, can, indeed cause serious harm.[9] Typical examples include depictions of or incitement to child abuse, graphic violence or cruelty, terrorist propaganda videos, hate speech and harassment.[10] More controversial is whether state authorities or internet service providers should be entitled or even obliged to protect users from unwanted exposition to pornography, explicit adult content or other kinds of unwanted content regarded as spam.[11] Furthermore, online content moderation is already applied to large extents,[12] and will undoubtedly be applied in the foreseeable future,[13] not only to block or delete obviously illegal content of the types just mentioned, but likewise for primarily economic reasons, in order to enforce copyright and other types of intellectual property claims.[14]

All major social media platforms have developed so called "community standards" or "guidelines" that rule out making available any material that falls under these categories, and many have some other more specific rules. Platform operators may be quite optimistically assuming that a considerable overlap exists between the policies of the large platforms and the applicable laws in the countries where these platforms offer their services, and also with our widely held ethical views. Besides broad areas of overlap, however, there are still relevant points where our ethical views might diverge. There are still differences between the applicable laws of different countries. Furthermore, there are differences between the laws of several countries and particular ethical claims either against or in favour of the public accessibility of certain kinds of images, and also between those particular laws and particular ethical convictions and the more or less uniform, "global" approach of the platform policies.

Indeed, we can no longer ignore that there are profound cross-cultural differences that make it very difficult to find a balance that will be acceptable in most societies. The question, then, is, of course, of how we are going to deal with such differences. Must we learn to avoid certain words

---

8 Garton Ash (2016).
9 Waldron (2012).
10 Kabasfrooshan (2019).
11 Cf. e.g., Cambridge Consultants (2019) 31; see also Ortland (2018) 136–161 (with further references).
12 Gillespie (2018).
13 Sartor/Loreggia 2020.
14 Valais (2015); Bar-Ziv/Elkin-Koren (2017); Husovec (2018).

and images that, for example, Muslims or Christian fundamentalists may find offensive? Or should they instead learn to respect, if not to embrace, liberal ideas about free speech?

IV. Algorithmic Judgment and the Pragmatics of Pictorial Speech Acts

Even if we might someday, hopefully, arrive at widely shared criteria for legitimate and necessary distinctions between acceptable and unacceptable images, there will still be the problem of how these criteria should be applied to the vast amount of information and misinformation circulating on the internet. Three or four years ago, there were broad discussions in the news media about overburdened content moderators at Facebook and other social media platforms and the law enforcement officers who must watch hours of sickening material, mostly images or videos, every day. "As the amount of user-generated content that platform users upload continues to accelerate, it has become impossible to identify and remove harmful content using traditional human-led moderation approaches at the speed and scale necessary".[15]

Will Artificial Intelligence sooner or later be able to solve this problem? Perhaps it might cause even more severe problems than those it is supposed to solve. Obviously, there are serious problems concerning the accountability of anonymous and non-transparent algorithms making decisions that may affect people's freedom of expression or other fundamental rights.[16] In order to sustain the rule of law, there will always be a need for human review and due process.[17]

But there are also considerable technical limits that need to be addressed. With automated hash-matching, further advances in deep neural network computing, machine learning and artificial intelligence technology, algorithms can now perform certain pre-moderation tasks in sifting large amounts of data. In other words, they can significantly improve the training data and assist human moderation.[18] It is important to note, however, that the classification of images according to the applicable criteria for blocking as well as for unblocking certain kinds of content poses con-

---

15 Cambridge Consultants (2019) 4.
16 Perel/Elkin-Koren (2017); Sinnreich (2018); Senftleben (2020).
17 Reinhardt (2020) 260–62.
18 For a brief summary of the technological development, see Cambridge Consultants (2019) 16–22; Jani/Chaudhuri/Patel/Shah (2020); Llansó/van Hoboken/Leerssen/ Harambam (2020).

siderable challenges that no automated algorithm so far can overcome, and there are reasons why this problem cannot be solved by simply improving the approaches to AI that have so far been developed.[19]

Whether we like it or not – and whether we like to call it 'censorship'[20] or not – I am afraid, algorithmic content moderation is here to stay. It has only evolved in recent years, but we may – and must – expect it to further increase and literally become pervasive.

Still, the automatic, algorithmic application of the criteria we (or whoever gets to decide) want to see applied on the vast stream of images and other forms of content circulating online is a huge problem and will probably remain a huge problem for any currently foreseeable future. Basically, this is the problem of judgment.[21] Of course, there are many kinds of decisions, i.e., determining judgments in Kant's sense,[22] which do *not* pose any particularly difficult problems of judgment, and which can be delegated to more or less automatic algorithms. By now, automatic pattern recognition software can determine incredibly complex and subtle patterns in images, surfaces, 3D physical bodies, environments or continuous temporal forms better than human observers could. We rely on such algorithms every day in many ways.

Facial recognition is still a huge challenge. But then we are told that some algorithms are now much better at it than even the most attentive humans.[23] Indeed, it is often difficult for most of us to recognize strangers merely from a photograph. Sometimes we won't recognise even our close acquaintances if we encounter them in places where we did not expect them, or we mistake strangers for friends.

What is really difficult for pattern recognition algorithms to figure out, however, is the distinction between relevant and irrelevant context. Such distinctions require a different type of judgment than the simple 'match' or 'no match' distinction. Regarding images, the question of context goes beyond the problem of framing — what is or ought to be regarded as part of the picture, and what ought to be distinguished from the picture as belonging to its background or surroundings. We have probabilistic

---

19 See Gorwa/Binns/Katzenbach (2020); Gillespie (2020); Castets-Renard (2020); Elkin-Koren (2020), from a constitutional law perspective and with an interesting suggestion for an "adversarial approach" by implementing "contesting algorithms"; Käde (2022).

20 Sirichit (2015); Cobbe (2019); Armijo (2020).

21 Cf. Kant (1790) A vii = AA V, 169.

22 Kant (1790) AA V, 179; Hanna (2017) suppl. 4.

23 Babcock (2015); Klosowski (2020); MacCarthy (2020).

algorithms that can deal with such questions in many cases quite success-fully, and we are, indeed, so used to expecting the algorithms to work that we find it startling to see them fail. But artificial intelligence algorithms so far have not been particularly successful in cases that require reflective judgment in Kant's terms,[24] or sound inductive generalisations from very limited data.

With regard to images, the main difficulty for algorithms that are ex-pected to decide which images ought to be blocked or deleted concerns the pragmatics of "pictorial speech acts",[25] rather than the syntactics of visual configurations or typical forms of depiction. People can do very different kinds of things with the same or quite similar pictures. And pictures that can be used to do the same or similar things may look totally different from one another.

From the Google *Transparency Report*, we can see that algorithms for "automated flagging" have been relevant for 95 % of the removal or block-ing decisions. Apparently, "automated flagging" does not yet mean auto-mated blocking or deletion, at least not at Google. Rather, a workforce of some 10,000 employees occupied with censorship tasks at Google account for these decisions.[26] Now, what kind of content can these automated flag-ging algorithms recognise as probably illegal? Does this recognition work only with the help of a huge database of material marked for blocking for various reasons, from criminal law or hate speech regulations to the en-forcement of copyright claims? Or can the algorithm recognize, for exam-ple, on a probabilistic account previously unknown images that should be blocked because they fall under one of the categories like child pornogra-phy, incitement to violence or terrorism? Other questions that are going to remain relevant, as far as we can see, include: How much time would an average content moderator usually have to decide about a video? How of-ten does the judgement of human moderators differ from the preliminary assessment provided by the algorithm? And to what degree is the judgment process accelerated with the help of the automated flagging algorithms? These are only some of the questions that will obviously need to be dis-cussed when formulating an ethics of images in the light of censorship that invariably comes with it.

---

24 Kant (1790) AA V, 179–181.
25 Kjørup (1978), see also Scholz (1991) 123–130; Bredekamp (2015) 58–59.
26 Google (2020).

References

Armijo, Enrique (2020): 'Reasonableness as Censorship – Algorithmic Content Moderation, The First Amendment, and Section 230 Reform', CSAS Working Paper 20–10 Fairfax, VA: George Mason University, Antonin Scalia Law School, Gray Center for the Study of the Administrative State, Spring 2020) (online available at https://administrativestate.gmu.edu/wp-content/uploads/sites/29/202 0/03/Armijo-Reasonableness-As-Censorship.pdf)

Aufderheide, Patricia (2020): 'How much copyright censorship is OK?' American University, Center for Media and Social Impact, Fair Use Blog, May 30, 2020, https://cmsimpact.org/fair-use-blog/much-copyright-censorship-ok/

Babcock, Ernest J. (2015): *Privacy Impact Assessment for the Facial Analysis, Comparison, and Evaluation (FACE) Services Unit* (Washington D.C.: FBI Services, Information Management, May 1, 2015, https://www.fbi.gov/services/informatio n-management/foipa/privacy-impact-assessments/facial-analysis-comparison-and -evaluation-face-services-unit)

Bar-Ziv, Sharon/Elkin-Koren, Niva (2018): 'Behind the Scenes of Online Copyright Enforcement – Empirical Evidence on Notice & Takedown', 50 Connecticut Law Review, No. 2 (2018) 339–385 (online available at https://opencommons.uc onn.edu/cgi/viewcontent.cgi?article=1395&context=law_review)

Bredekamp, Horst (2015): *Der Bildakt* (Berlin: Wagenbach, 2015)

Cambridge Consultants (2019): *Use of AI in Online Content Moderation – Report on behalf of Ofcom* (Cambridge: Cambridge Consultants, 2019) (online available at https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consul tants-ai-content-moderation.pdf)

Castets-Renard, Céline (2020): 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement', 20 Illinois Journal of Law, Technology & Policy, No. 2 (2020) 283–323 (online available at http://illinoisjltp.co m/journal/wp-content/uploads/2020/12/Castets-Renard_Final (1).pdf)

Cobbe, Jennifer (2020): 'Algorithmic Censorship by Social Platforms – Power and Resistance', 34 Philosophy & Technology (2020) 739–766 (online available at https://doi.org/10.1007/s13347-020-00429-0)

Elkin-Koren, Niva (2020): 'Contesting algorithms – Restoring the public interest in content filtering by artificial intelligence', Big Data & Society (July-December 2020) 1–13 (online available at https://doi.org/10.1177/2053951720932296

Foucault, Michel (1975): *Surveiller et punir – Naissance de la prison* (Paris: Gallimard, 1975)

Ganor, Boaz (2002): 'Defining Terrorism', 29 Media Asia, No. 3 (2002) 123–133 (online available at https://doi.org/10.1080/01296612.2002.11726675)

Garton Ash, Timothy (2016): *Free Speech – Ten Principles for a Connected World* (New Haven, CT/London: Yale University Press, 2016)

Gillespie, Tarleton (2018): *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (New Haven and London: Yale University Press, 2018)

Gillespie, Tarleton (2020): 'Content moderation, AI, and the question of scale', Big Data & Society (2020), https://doi.org/10.1177/2053951720943234

Google (2020): *Transparency Report*, https://transparencyreport.google.com/ and https://transparencyreport.google.com/copyright/overview (quoted data from 18 Sept. 2020)

Gorwa, Robert/Binns, Reuben/Katzenbach, Christian (2020): 'Algorithmic content moderation – Technical and political challenges in the automation of platform governance', Big Data & Society (February 28, 2020) (online available at https://doi.org/10.1177/2053951719897945)

Green, Jonathan/Karolides, Nicholas J. (2005): *Encyclopedia of Censorship* (New York: Facts on File, 2005)

Hanna, Robert (2004): 'Kant's Theory of Judgment', in: Zalta, E.N.: *Stanford Encyclopedia of Philosophy* (first published Wed Jul 28, 2004; substantive revision Mon Oct 23, 2017, https://plato.stanford.edu/archives/win2018/entries/kant-judgment/)

Heldt, Amélie Pia (2019): 'Upload-Filters: Bypassing Classical Concepts of Censorship?', 10 JIPITEC, No. 2 (2019) 57–65, https://www.jipitec.eu/issues/jipitec-10-1-2019/4877/JIPITEC_10_1_2019_57_Heldt-JIPITEC_10_1_2019_57_Heldt-Upload-Filters

Husovec, Martin (2018): 'The Promises of Algorithmic Copyright Enforcement – Takedown or Staydown? Which is Superior? And Why?', 42 Columbia Journal of Law & the Arts (2018) 53–84 (online available at https://journals.library.columbia.edu/index.php/lawandarts/article/view/2009/971)

Jani, Kunal/Chaudhuri, Mainak/Patel, Hrishitva/Shah, Manan (2020): 'Machine learning in films – An approach towards automation in film censoring', 2 Journal of Data, Information and Management, No. 2 (2020) 55–64 (https://doi.org/10.1007/s42488-019-00016-9)

Jones, Derek (2001): *Censorship – A World Encyclopedia* (London/New York: Routledge 2001)

Kabasfrooshan, Ashkan (2019): 'Counter Strike – Exposing Worst ContentID Abusers', Context TV, Business Battleground (2019), https://www.youtube.com/watch?v=--e59ApQino

Käde, Lisa (2022): 'The Issue of the Image of Algorithms', in: Dreier, Thomas/Andina, Tiziana (eds), *Digital Ethics – The Issue of Images* (Baden-Baden: Nomos 2022) 335–353

Kant, Immanuel (1790): *Kritik der Urtheilskraft* (1790), in: Windelband, Wilhelm (ed), *Kants gesammelte Schriften*, Akademie-Ausgabe, Vol. V (Berlin: de Gruyter 1908) 165–485

Kjørup, Søren (1978): 'Pictorial Speech Acts', 12 Erkenntnis, No. 1 (1978) 55–71

Klosowski, Thorin (2020): 'Facial Recognition Is Everywhere. Here's What We Can Do About It', The New York Times, July 15, 2020 (online available at https://www.nytimes.com/wirecutter/blog/how-facial-recognition-works/)

Llansó, Emma/van Hoboken, Joris/Leerssen, Paddy/Harambam, Jaron (2020): *Artificial Intelligence, Content Moderation, and Freedom of Expression – A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression* (2020) (online available at https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf)

Matthiesen, Kay (2008): 'Censorship and Access to Information', in: Himma, Kenneth E./Tavani, Herman T. (eds), *Handbook of Information and Computer Ethics* (Hoboken, NJ: J. Wiley & Sons 2008) 573–587

MacCarthy, Mark (2020): 'Who thought it was a good idea to have facial recognition software?', Brookings, August 20, 2020 (online available at https://www.brookings.edu/research/who-thought-it-was-a-good-idea-to-have-facial-recognition-software/)

Ortland, Eberhard (2018): 'A Moral Need for Censorship, Access Restrictions, and Surveillance? The Ethics of Copying and the Freedom of Communication', 26 Jahrbuch für Recht und Ethik/Annual Review of Law and Ethics (2018) 125–180

Perel, Maayan/Elkin-Koren, Niva (2017): 'Black Box Tinkering – Beyond Disclosure in Algorithmic Enforcement', 69 Florida Law Review (2017) 181–221

Reinhardt, Jörn (2020): 'Algorithmizität und Sichtbarkeit – Konflikte um Bilder in den sozialen Medien', in: E. Schürmann, Eva/von Plato, Levno (eds), *Rechtsästhetik in rechtsphilosophischer Absicht – Untersuchungen zu Formen und Wahrnehmungen des Rechts* (Baden-Baden: Nomos 2020) 249–262

Sartor, Giovanni/Loreggia, Andrea (2020): *The impact of algorithms for online content filtering or moderation – Study, requested by the JURI committee, PE 657.101* (Brussels: European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs 2020) (online available at https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf)

Scholz, Oliver R. (1991): *Bild, Darstellung, Zeichen – Philosophische Theorien bildhafter Darstellung*, (Freiburg/München: Alber 1991)

Senftleben, Martin (2020): 'Institutionalized Algorithmic Enforcement – The Pros and Cons of the EU Approach to UGC Platform Liability', Florida International University Law Review (2020) 299–328 (online available at https://ecollections.law.fiu.edu/lawreview/vol14/iss2/11/)

Sinnreich, Aram (2018): 'Four Crises in Algorithmic Governance', 26 Jahrbuch für Recht und Ethik/Annual Review of Law and Ethics (2018) 181–190

Sirichit, Methaya (2015): 'Censorship by Intermediary and Moral Rights – Strengthening Authors' Control Over Online Expressions Through the Right of Respect and Integrity', 3 Journal of Law, Technology and Public Policy (2015) 54–159

Steinhauer, Eric (forthcoming): 'Zensur', in: Ortland, Eberhard/Dreier, Thomas/Schmücker, Reinold (eds), *Original und Kopie – Ein Sachlexikon* (Hamburg: F. Meiner forthcoming)

Valais, Sophie (2015): 'Copyright enforcement online: policies and mechanisms', IRIS-plus (Strasbourg: European Audiovisual Observatory, 2015–3) (online available at https://rm.coe.int/1680783480)

Waldron, Jeremy (2012): *The Harm in Hate Speech* (Cambridge, MA: Harvard University Press, 2012)

Zuckerberg, Mark (2018): Facebook CEO Mark Zuckerberg Hearing on Data Privacy and Protection, *C-SPAN* (10 April, 2018, https://www.c-span.org/video/?443 543-1/facebook-ceo-mark-zuckerberg-testifies-dataprotection)