SIMON ROTHÖHLER

INFORMATIONEN, DIE BILDER HABEN

Zur Moderierbarkeit von visuellem Content

Wer zu rekonstruieren versucht, seit wann usergenerierter Content, der über sozialmediale Plattformen zirkuliert und deren ‹Leitwährung› ist,[1] in der allgemeinen Öffentlichkeit als etwas wahrgenommen wird, das einer Moderation – also einer Lenkung und Steuerung, Mäßigung und Dämpfung – unterliegt und bedarf, stößt ziemlich schnell auf einen initialen Artikel von Adrian Chen, der am 16. Februar 2012 auf Gawker publiziert wurde. Ausgangspunkt war ein geleaktes Richtliniendokument, das Chen über digitale Mikroarbeit verrichtende Vertragspartner eines kalifornischen Subunternehmens namens oDesk zugespielt wurde: «Facebook's Operation Manual for Content Moderators».[2] Auf den zu Ausbildungs- und Trainingszwecken erstellten Präsentationsfolien, deren Existenz bis dato weitgehend unbekannt gewesen war und die in der Folge durch zahlreiche ähnliche Leaks und Investigativrecherchen bestätigt wurden,[3] artikulieren sich Kriterien, Klassifikationen und Handlungsanweisungen, die menschliche Akteur_innen in die Lage versetzen sollen, unerwünschte Inhalte als solche zu identifizieren und einem bis zu Inhaltslöschung und Suspension von Nutzer_innen eskalierbaren Prozess der ‹Moderation› zuzuleiten.

Sortiert in allgemeinere Kategorien – «Sex and Nudity», «Illegal Drug Use», «Theft Vandalism and Fraud», «Bullying and Harassment», «Hate Content», «Graphic Content», «Self-harm», «Credible Threats» –, fallen die derart subsumierten Inhaltsbestimmungen schon im nächsten Schritt deutlich deskriptiver aus und lassen erahnen, was aus Sicht eines Plattformbetreibers – der seine Benutzeroberflächen einerseits werbeökonomisch rationalisiert, andererseits potenzielle Haftungsrisiken zu minimieren sucht – alles umzulenken und abzuschwächen ist: «Depicting the mutilation of people or animals, or decapitated, dismembered, charred, or burning humans», «People ‹using the bathroom›», «Images of drunk and unconscious people, or sleeping people with things drawn on their faces», «Mothers breastfeeding without clothes on» etc.

Neben vielleicht erwartbaren Empörungswellen, die sich an bestimmten Differenzierungsbemühungen der Facebook-Richtlinien entzündeten («male nipples

1 Vgl. Sarah T. Roberts: Digital detritus: ‹Error› and the logic of opacity in social media Content-Moderation, in: *First Monday*, Vol. 23, Nr. 3–5, März 2018.

2 Vgl. Adrian Chen: Inside Facebook's Outsourced Anti-Porn and Gore Brigade. Where ‹Camel Toes› are More Offensive Than ‹Crushed Heads›, in: *gawker.com*, 16.2.2012.

3 Vgl. Adrian Chen: The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed, in: *Wired*, dort datiert 23.10.2014, *wired.com/2014/10/content-moderation/*; Nick Hopkins: Revealed: Facebook's internal rule on sex, terrorism and violence, dort datiert 21.5.2017, in: *The Guardian*, *theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence*; zum Berliner «Löschteam» von Arvato vgl. Till Krause, Hannes Grassegger: Inside Facebook, in: *SZ-Magazin*, online unter *sz.de/1.3297138*; alles gesehen am 28.7.2018.

are ok»), ist den ersten Reaktionen aus heutiger Sicht vor allem zu entnehmen, dass Content-Moderation zum damaligen Zeitpunkt auch für Plattformanalysten ein relativ unerforschtes Gelände darstellte – insbesondere hinsichtlich der konkreten operativen Implementierung und Pragmatik der anbieterseitig institutionalisierten Moderationsroutinen. Chens Beitrag verstand sich denn auch ausdrücklich als erstes Kartierungsangebot, hinterließ aber vor allem den Eindruck, dass sich hier eine Reihe sehr grundsätzlicher Fragen stellen: Zu welchen medientechnischen Bedingungen regulieren Plattformen Inhaltserzeugungen ihrer User_innen? Welche normativen Vorannahmen werden wie in Richtlinien und Skripte übersetzt? Welche Legitimitätsansprüche und ökonomischen Kalküle sind diesen Handlungsvorschriften eingeschrieben? Zu welchen Operationsketten werden menschliche und nichtmenschliche Akteur_innen im Zuge der Moderationsleistung verbunden? Wie verhalten sich nach menschlichen Einzelfallbeurteilungen getroffene zu automatisch generierten Entscheidungen? Weshalb wird die dabei anfallende Arbeit vorwiegend im Unsichtbaren verrichtet – und: von wem, wo und unter welchen ökonomischen Rahmenbedingungen?

Nach einer gewissen Latenzphase ist die Frage nach Zuschnitt und Reichweite plattformspezifischer Content-Moderation mittlerweile auch im medienwissenschaftlichen Diskurs angekommen. Neben den Beiträgen von Sarah T. Roberts[4] liegt mit Tarleton Gillespies *Custodians of the Internet. Platforms, Content-Moderation, and the Hidden Decisions That Shape Social Media* nun eine erste Monografie vor, die nicht nur exemplarische Einzelfälle wie Facebooks temporäre Löschung des ikonischen Vietnamkriegsfotos von Nick Ut (*The Terror of War*, 1972),[5] «adult nudity» im Kontext fotografischer Holocaust-Dokumente[6] oder die ebenfalls breit diskutierten «breastfeeding photos» behandelt, sondern sich an einer systematischeren Klärung versucht.[7]

Gillespie versteht Moderation als konstitutiven Plattformprozess, der industrieweit ein flexibel konfigurierbares soziotechnisches Ensemble («moderation apparatus») hervorgebracht hat, in dem medientechnische Arrangements in Kooperation mit menschlichen Programmier-, Regelungs- und Bewertungsleistungen zu Plattformpolitiken führen, die als «Gemeinschaftsstandards» veröffentlicht werden, in wesentlichen Hinsichten aber intransparent bleiben. Die verfügbaren Handlungsmodelle der Inhaltserkennung, -prüfung und -löschung werden hier unterschiedlich gewichtet und prozedural kombiniert: Von der kompletten Inhaltssichtung durch den Betreiber (*editorial review*) über die quasiselbstregulative Inanspruchnahme von Nutzer_innengemeinschafen (*community flagging*), die zumindest Vorarbeiten hinsichtlich der Identifizierung und Meldung erbringen sollen, bis zu Versuchen, die Moderation weitgehend an technische Akteure zu delegieren (*automatic detection*) und editoriale wie nutzerseitige Entscheidungsspielräume entsprechend zu minimieren. Unabhängig von konkret umgesetzten Regimen sei die Moderationsfrage für Plattformen keine nachrangige, sondern betreffe – insbesondere in medienökonomischer Hinsicht[8] – deren Kern:

4 Vgl. Roberts: Digital deritus; dies.: Content-Moderation, in: Laurie A. Schintler, Connie L. McNeely, Geoffrey J. Golson (Hg.): *Encyclopedia of Big Data* (im Erscheinen); dies.: Commercial Content-Moderation: Digital laborers' dirty work, in: Safiya Umoja Noble, Brendesha M. Tynes (Hg.): *The Intersectional Internet: Race, Sex, Class and Culture Online*, New York 2016, 147–159.

5 Vgl. Roberts: Digital detritus.

6 Vgl. Alexis C. Madrigal: Inside Facebook's Fast-Growing Content-Moderation Effort, in: *The Atlantic*, dort datiert 7.2.2018, *theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/*, gesehen am 28.7.2018.

7 Vgl. Tarleton Gillespie: *Custodians of the Internet. Platforms, Content-Moderation, and the Hidden Decisions That Shape Social Media*, New Haven, London 2018, 141–175.

8 Das Primat werbeökonomischer Logiken der Content-Moderation war zuletzt im Fall von YouTube zu beobachten, vgl. Davey Alba: YouTubes Ad Problems Finally Blow Up in Google's Face, in: *wired.com*, 25.3.2017.

> [M]oderation is, in many ways, the commodity that platforms offer. Though part of the web, social media platforms promise to rise above it, by offering a better experience of all this information and sociality: curated, organized, archived, and moderated. [...] Moderation is not an ancillary aspect of what platforms do. It is essential, constitutional, definitional. Not only can platforms not survive without moderation, they are not platforms without it. Moderation is there from the beginning, and always; yet it must be largely disavowed, hidden, in part to maintain the illusion of an open platform and in part to avoid legal and cultural responsibility. Platforms face what may be an irreconcilable contradiction: they are represented as mere conduits and they are premised on making choices for what users see and say.[9]

Die auf einer komplexen soziotechnischen Logistik[10] basierende Pragmatik von Moderation fällt demnach mit der strategischen Kommunikation von Neutralität zusammen, die Gillespie als zentralen «Mythos»[11] der Plattformpolitik begreift, der Mitte der 2000er Jahre auf dem Feld des Copyrights (man denke beispielsweise an YouTubes 2007 eingeführte Content-ID oder an die aktuelle Debatte in der EU um leistungsschutzrechtliche Uploadfilter) ins Wanken geriet, bevor schließlich – Gillespie nennt den 2006 in Großbritannien erlassenen *Terrorism Act* als Wendepunkt (Plattformbetreibern bleibt seitdem zwei Tage Zeit, bevor ein nichtgelöschter Content als anbieterseitig *endorsed* gilt) – vor allem die sozialmediale Distribution von Inhalten in den Mittelpunkt rückte, die extremistische Gruppierungen zu Propaganda- und Rekrutierungszwecken lancieren.

Prinzipiell fragwürdig erscheint der Neutralitätsmythos zugleich mit Blick auf Moderationspraktiken, die effektiv als Zensur wirksam werden und dem kommunizierten Selbstverständnis als grundsätzlich offene, barrierelose «speech machines» zuwiderlaufen.[12] Hier geht es meist weniger um Piraterie, Pornografie oder Terrorismus, sondern um die mit ökonomischen Rationalitäten verbundene Anpassungsbereitschaft von Plattformbetreibern, was in einer Reihe von Fällen de facto zur Übernahme und Durchsetzung von Regulationsdoktrinen autokratischer Regime geführt hat.[13] Sarah Roberts argumentiert diesbezüglich, dass der Mythos einer gleichsam plattformtechnisch garantierten ‹Neutralität› einer bewussten Depolitisierung im Sinne der «brand protection» geschuldet ist: «[T]his operating logic of opacity serves to render platforms as objective in the public imagination, driven by machine/machine-like rote behavior.»[14] Die investierten normativen Gehalte des Moderationsdesigns werden demzufolge geblackboxt und verschwinden regelmäßig hinter der popularisierten Vorstellung einer ‹neutralen› technischen Delegierbarkeit, die vorgeblich ohne *human bias* ist.

Auch Gillespie betont, dass die konkrete Ausgestaltung gegenwärtig installierter Moderationsdispositive in erster Linie als Reaktion auf das gewaltige Volumen und die «echtzeitliche»[15] Zirkulation von usergenerierten Inhalten verstanden werden muss:

> Content is policed at scale, and most complaints are fielded at scale. More important, the ways moderators understand the problems have been formed and shaped by working at this scale. […] What to do with a questionable photo […] when you're facing not one violation but hundreds exactly like it, and thousands much like it,

9 Ebd., 13, 18.
10 Vgl. Ebd., 136 ff.
11 Ebd., 24 ff.
12 Ebd., 48.
13 Umgesetzt vor allem in Form von IP-Blocking, vgl. ebd., 190 ff. und Mike Isaac: Facebook Said to Create Censorship Tool to Get Back into China, in: *The New York Times*, 22.11.2016.
14 Roberts: Digital detritus.
15 Zur Unterscheidung plattformspezifisch feinmodulierter Formen von «realtimeness» vgl. Esther Weltevrede, Anne Helmond, Carolin Gerlitz: The Politics of Real-time: A Device Perspective on Social Media Platforms and Search Engines, in: *Theory, Culture & Society*, Vol. 31, Nr. 6, 2014, 125–150.

> but slightly different in a thousand ways. This is not just a difference of size, it is fundamentally a different problem. For large-scale platforms, moderation is industrial, not artisanal.[16]

Dass die flächendeckend beobachtbare Industrialisierung der Content-Moderation gerade kein rein algorithmisch-informationstechnischer Vorgang ist, sondern de facto auf kleinteilige Prozesse menschlicher Dateninterpretation angewiesen bleibt – deren *workload* weitgehend in den globalen Süden outgesourct wird –, hat Adrian Chen in einem Beitrag für das auf *visual journalism* spezialisierte Rechercheprojekt *Field of Vision* (das zum Medienkonzern First Look Media des eBay-Gründers Pierre Omidyar gehört) auch filmdokumentarisch nachvollzogen. *The Moderators*[17] beobachtet in Form einer medienethnografischen Miniatur, wie der globale Content-Strom auf die prekäre sozioökonomische Verteilungsrealität digitaler Mikroarbeit trifft. Die Dateninhalte, die hier zu moderieren sind, bestehen, wie sich im Verlauf der filmisch dokumentierten Ausbildungswoche in einem indischen Subunternehmen zeigt, in erster Linie aus visuellem, genauer: bildhaftem Content – Digitalfotografien, Streamvideos, GIFs –, der massenhaft über die Benutzeroberflächen sozialmedialer Plattformen distribuiert ist.

Auf den skriptgemäßen Reinigungsauftrag, auf die Inhaltsidentifizierung, -markierung und -kassation werden die angehenden Moderator_innen mittels kasuistischer Übungen vorbereitet. Die Interpretation des Materials mag im Einzelfall kompliziert, kontextabhängig und generell kulturell voraussetzungsreich sein – der Entscheidungsspielraum besteht aber in einem schlichten Binarismus: *ignore/delete*. Die Praxisform der hier zu erbringenden «Human Intelligence Tasks» (HIT) erweist sich, wie in vielen Feldern des *crowd working*, als Filterfabrikarbeit:[18] An einer Stelle ist von 2000 Bildern pro Stunde die Rede, die gleichsam am Bildschirmfließband gering entlohnter indischer Moderator_innen vorüberziehen. Wie auch in *The Cleaners*, einem weiteren Dokumentarfilm, der sich mit einem ähnlichen Subunternehmen auf den Philippinen beschäftigt, wird unmittelbar anschaulich, dass Content-Moderation realiter vor allem eines ist: serielle Bildbetrachtung und serielle Bildbeurteilung durch menschliche Akteur_innen.[19]

Während die Social-Media-Plattformbetreiber, die sich aus Haftungsgründen ausdrücklich nicht als «media companies»[20] verstehen wollen, noch bis vor kurzem[21] nicht nur in der Unternehmenskommunikation den Eindruck zu erwecken versuchten, dass jedwede Inhaltsprozessierung weitgehend medientechnisch automatisiert sei, zeigt sich in der alltäglichen Arbeitsrealität der Content-Moderation, dass Daten, die als digitale Bilder formatiert sind und auf Benutzeroberflächen entsprechend ikonisch materialisiert werden können, weiterhin eine spezifische Grenzfigur informationstechnischer Verarbeitung darstellen. Claus Pias' Hinweis, dass digitale Bilder aus medientheoretischer Sicht Bilder sind, «die Informationen haben»,[22] stellt sich mit Blick auf die nicht an technische Aktanten übertragbaren, nichtautomatisierbaren

16 Gillespie: *Custodians of the Internet*, 77.

17 Vgl. *The Moderators*, Regie: Ciaran Cassidy, Adrian Chen, USA 2017, online unter fieldofvision.org/the-moderators, gesehen am 10.7.2018.

18 Ayhan Aytes bemerkt zur Verteilungsrealität der arbeitsrechtlich weitgehend unregulierten *crowd work*: «[T]his sociotechnical system represents a crucial formation on a global scale as it facilitates the supply of cognitive labor needs to mainly Western information and communication technologies industries from a global workforce.», ders.: Return of the Crowds. Mechanical Turk And Neoliberal States of Exception, in: Trebor Scholz (Hg.): *Digital Labor. The Internet as Playground and Factory*, New York 2013, 79–97, hier 80.

19 Vgl. *The Cleaners*, Regie: Hans Block, Moritz Riesewick, D 2018. *The Cleaners* geht insofern über *The Moderators* hinaus, als Block/Riesewick zumindest punktuell versuchen, die Moderator_innen in ihren (oftmals prekären) lebensweltlichen Milieus zu verorten und dabei kulturelle Konfliktfelder und psychische Belastungsmomente der Moderationstätigkeit anzudeuten.

20 Gillespie: *Custodians of the Internet*, 7 f.

21 Facebook hat seit Mai 2017 die Strategie gewechselt und kommuniziert die Investitionen in überwiegend outgesourcte menschliche Moderationsteams nun als integralen Bestandteil der *brand protection*. Vgl. Samuel Gibbs: Facebook Live: Zuckerberg adds 3.000 moderators in wake of murders, in: theguardian.com, 3.5.2017.

22 Claus Pias: Das digitale Bild gibt es nicht. Über das (Nicht-) Wissen der Bilder und die informatorische Illusion, in: *zeitenblicke*, Nr. 2., H. 1, 2003.

Anteile gewöhnlicher Content-Moderation in gewisser Weise umgekehrt dar: Der *moderation apparatus* ist auf die hermeneutischen Ressourcen menschlicher Akteur_innen insbesondere dann angewiesen, wenn es in den durch Computernetzwerke zirkulierenden Datenströmen um Informationen geht, die Bilder haben.[23]

Über die in Frage stehenden sozialmedialen Digitalbilder, deren Informationseinheiten diskret adressiert, kalkuliert, manipuliert werden können, lässt sich grundsätzlich auch sagen, dass es sie nicht «nicht» (Pias), sondern zweifach gibt: als unsichtbaren, gespeicherten Code, mit dem Computerprogramme auch jenseits von Darstellungsaufträgen rechnen können, und als visualisierte Form, die von menschlichen Wahrnehmungsleistungen als Bild erkannt und behandelt werden kann.[24] Die Überführung des ersten Zustands in den zweiten aktualisiert sich zwar kontingent und mag aus Sicht des Computers, der den visuellen Output nicht benötigt, um mit den Bilddaten rechnen zu können, einigermaßen verzichtbar erscheinen.[25] Ohne bildförmige Phänomenalisierung gehen andererseits aber auch Informationen (und Handlungsressourcen) verloren, was durch die nach wie vor begrenzte Maschinenlesbarkeit des Bildes vielfach bestätigt wird. Dass digitale Bilddaten ohne Formbezug auch informationstheoretisch nur bedingt gedacht werden können, hat William J. T. Mitchell auf eine knappe Formel gebracht: «[I]mage have always given form to information.»[26]

Digitale Bilder sind insofern Formen, die sich mittels algorithmischer Performanzen auf Screens und Displays materialisieren können und dabei perzipierbar werden. Grundsätzlich ist hier aber von einem Spannungsverhältnis zwischen Code und Form auszugehen, das sich in unterschiedlichen Anwendungskontexten unterschiedlich relaxiert.[27] So versuchen rezente Verfahren maschineller Bilderkennung auf Basis Künstlicher Neuronaler Netze (KNN) in Bitmaps Muster zu erkennen, welche als Bildinhaltsinformation operabel werden sollen: «They can tell what's in an image by finding patterns between pixels on ascending levels of abstraction, using thousands to millions of tiny computations on each level. New images are put through the process to match their patterns to learned patterns.»[28] Adrian Mackenzie hat diese maschinellen Prozesse der Bilderkennung (eigentlich genauer: Bildklassifizierung) mit Blick auf den ‹Katzenbildradar› *kittydar* detaillierter beschrieben:

> Faced with the immense accumulation of cat images on the internet, kittydar can do little. It only detects the presence of cats that face forward. It sometimes classifies people as cats. [...] [T]he software finds cats by cutting the image into smaller windows. For each window, it measures a set of gradients […] running from light and dark and then compares these measurements to the gradients of known cat images (the so called ‹training data›). The work of classification according to these simple categories of ‹cat› and ‹no cat› is given either to a neural network […], themselves working on images of cats other things taken from YouTube videos, or to a support vector machine [...].[29]

23 Zu Wortfiltertechnologien und Problemen, die sich ergeben, wenn digitale Informationen beispielsweise textförmigen *hate speech content* ‹haben›, vgl. Gillespie: *Custodians of the Internet*, 98 ff., 104 f.

24 Sarah Kember hat mit Blick auf die sozialmediale Oberflächenrealität visueller Kultur von der «endurance of photographic codes and conventions» gesprochen, dies.: Ambient Intelligent Photography, in: Martin Lister (Hg.): *The Photographic Image in Digital Culture*, London 2013, 56–76, 57.

25 Vgl. Friedrich Kittler: *Optische Medien. Berliner Vorlesung 1999*, Berlin 2011, 293 ff.

26 William J. T. Mitchell: Image, in: ders., Mark B.N. Hansen (Hg.): *Critical Terms for Media Studies*, Chicago 2010, 35–48, hier 46.

27 Johanna Drucker hat dieses Spannungsverhältnis in Form einer Kritik digitaler Ontologien adressiert: Gegen den Mythos von «pure code» (als idealisierter Vorstellung einer immateriellen «mathesis») versteht sie digitale Bilder als «material embodiments», die der oppositionellen Logik der «graphesis» («knowledge manifest in visual and graphic Form») entstammen. Vgl. dies.: *Graphesis. Visual Forms of Knowledge Production*, Cambridge 2014.

28 Dave Gershgorn: It's not about the Algorithm. The data that transforms AI research and possibly the world, in: qz.com, 26.7.2017.

29 Adrian Mackenzie: *Machine Learners. Archaeology of a Data Practice*, Cambridge 2017, 4. Die automatische Identifizierung, Klassifizierung und Annotierung von Bildinhalten wird mittlerweile auch durch komplexere *image captioning systems* ergänzt. Vgl. dazu James Walker: Googles AI can now caption images as well as humans, in: digitaljournal.com, 23.9.2016, sowie research.googleblog.com.

Den unterschiedlich eindrucksvollen Fortschritten lernalgorithmischer Bilderkennungs- bzw. Bildklassifizierungsverfahren, die an Beispieldaten – wie dem als Wettbewerbsstandard maschinellen Lernens diesbezüglich etablierten Datensatz des ImageNet-Projekts – ‹selbstständig trainieren›, stehen jedenfalls nach wie vor eher ernüchternde Gegenproben wie die für menschliche Akteur_innen völlig unproblematische Klassifikation von Tonwertumkehrbildern gegenüber.[30] Auch wenn die Klassifikation und Identifizierung von Objekten, die hier als repräsentierter Bildinhalt verstanden werden, im Fall des ImageNet-Datensatzes mittlerweile mit Fehlerquoten unter 2 % gelingt, gilt für anspruchsvollere Leistungen auch heute noch: «This doesn't mean an algorithm knows the properties of that object, where it comes from, what it's used for, who made it, or how it interacts with its surroundings. In short, it doesn't actually understand what it's seeing.»[31]

Was Rechenmaschinen in Bildern ‹sehen›, ist eigentlich schon allein deshalb bestenfalls im übertragenen Sinn zu beantworten, weil schrittweise operationalisierte Kalkulationen sich nicht wirklich mit einem menschlichen Wahrnehmungseindruck vergleichen lassen. Das Problem mit der «Semantik im Prozess der Semiose», die der ‹bedeutungsindifferente› Computer nicht kennt, weil er «rein syntaktisch [operiert]»,[32] wird durch Modelle, die Bildverarbeitung zeitlich sequenzieren und räumlich segmentieren müssen, nicht gerade kleiner: «[Z]wischen den endlosen Ziffernkolonnen und den Gestalten, die ein menschlicher Blick erkennt, gähnt eine Lücke. [...] Auf der einen Seite stehen Rohdaten, die Bilder als Felder farbiger Pixel kodieren; auf der anderen Seite eine Wahrnehmung, die nicht anders kann, als etwas zu sehen: Gesichter, Personen, Räume, Gegenstände.»[33] So wird bereits mit Blick auf gewöhnliche Formen des «vollautomatisierten öffentlichen Turing Tests» – im Original CAPTCHA genannt: *completely automated public Turing Test to tell computers and humans apart* – deutlich, dass Datenobjekte, die der menschlichen Wahrnehmung auf subjektiv unproblematische, intuitive, unmittelbare Weise als Bilder *von etwas* erscheinen, durch ihre relative Maschinenunlesbarkeit weiterhin als Sicherheitsabfragen umfunktionalisiert werden können.

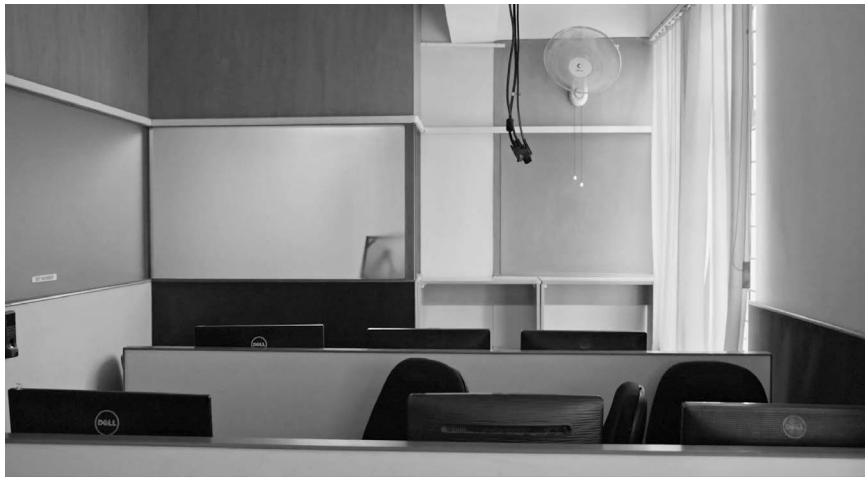
Hinzukommt, dass auch dort, wo maschinell trainierte Bildalgorithmen zumindest in instrumentellen Hinsichten bereits einigermaßen effizient und zielgerichtet zu funktionieren scheinen und entsprechende Automatismen hervorbringen, von ‹Neutralität› schon in Bezug auf die Trainingsdaten natürlich keine Rede sein kann. Robert Gehl und andere haben mit Blick auf – gerade im Kontext von Content-Moderation attraktive – CVPF-Verfahren (*computer vision-based pornography filtering*) gezeigt, dass sich bereits vor der Implementierung lernalgorithmisch ausgebildeter Filter (die ohnehin nicht das fluide kulturelle Konstrukt ‹Pornografie›, sondern lediglich typische Muster der Pixelverteilung nackter menschlicher Körper erkennen) in menschlichen Wahrnehmungskonventionen arbeitende heteronormative Skripte inskribieren,

30 Wie aus einem jüngeren Forschungsbericht zur Deep-Learning-Bilderkennung hervorgeht: «[W]e test the state-of-the-art DNNs with negative images and show that the accuracy drops to the level of random classification. This leads us to the conjecture that the DNNs, which are merely trained on raw data, do not recognize the semantics of the objects, but rather memorize the inputs.», Hossein Hosseini, Radha Poovendran: Deep Neural Networks Do Not Recognize Negative Images, in: *arxiv.org*, 20.3.2017.

31 Gershgorn: *It's not about the Algorithm*.

32 Martin Warnke: Bildersuche, in: *Zeitschrift für Medienwissenschaft*, Nr. 1, 2009, 29–37, hier 30.

33 Wolfgang Ernst, Stefan Heidenreich, Ute Holl: Editorial. Wege zu einem visuell adressierbaren Bildarchiv, in: dies. (Hg.): *Suchbilder. Visuelle Kultur zwischen Algorithmen und Archiven*, Berlin 2003, 7–15, hier 11.



Screenshots aus: *The Moderators*, Regie: Ciaran Cassidy, Adrian Chen, USA 2017 (Orig. in Farbe)

die routinemäßig *gender bias*[34] übertragen. Diese beeinflussen konkret, was Computer ‹sehen›, sofern sie die Beispieldaten vorselektieren, an denen die Musterextraktion allererst trainiert wird:

> Indeed, our analysis of CVPF shows that the computer scientists who train computers to see and filter online pornography are inscribing assumptions about pornography, human sexuality, and bodies into their academic field: namely, that pornography is limited to images of naked women; that sexuality is largely comprised of men looking at naked women; and that pornographic bodies comport to specific, predictable shapes, textures, and sizes. In other words, judging from their published works and conference articles, computer scientists appear to be training computers to see the narrow form of pornography described above while dismissing a heterogeneous array of other forms of pornography (gay, queer, trans*, hardcore, fat, bondage, hairy, and so much more) as ‹noise›.[35]

Trotz derartiger Limitierungen, die auf den strukturellen Konservatismus maschinellen Lernens verweisen – dessen Erfolge davon abhängen, «how well it makes the same distinctions that were made before»[36] –, findet eine Delegation von Handlungsmacht, die Bilder operativ an Rechenmaschinen rückbindet, gleichwohl auf verschiedenen Anwendungsfeldern längst statt. Das zeigt sich im digitalen Alltag, beispielsweise an Gesichtserkennungsalgorithmen, die in Fotoapplikationen Sammlungen vorsortieren, oder bei smarten Sicherheitskamerasystemen wie Netatmo («mit Erkennung von Menschen, Fahrzeugen und Tieren»), die ausgewählte Bilddaten mit Blick auf die Autorisierung von Zugangsberechtigung komputieren.

In klar abgegrenzten Anwendungskontexten des Internets der Dinge scheint die automatische Erfassung von Bildinhalten tatsächlich immer effizienter zu gelingen, was für die technischen Entwicklungshorizonte vor allem deshalb entscheidend ist, weil es gerade Bildsensoren sind, die immer mehr Weltausschnitte in den Datenraum ziehen.[37] Hier geht es im Kern darum, Adressierungen auch in sensortechnisch erfassten Handlungsräumen vornehmen zu können, in denen vorab keine «grammars of action»[38] präskribiert wurden. Bildanalytisch gesehen, sind die Anforderungen sozialmedialer Content-Moderation hinsichtlich der spezifischen Lektüreleistungen jedoch deutlich anspruchsvoller als die Identifizierung einer Milchtüte durch die hochaufgelösten 4K-Sensoren eines smarten Kühlschranks wie Samsungs RB7500. Auf Social-Media-Plattformen geht es nicht um mit überschaubaren Entscheidungs- und Handlungsketten verbundene Datenkreisläufe – etwa: Die Bildsensoren von Nests Videotürklingel «Hello» erfassen und erkennen per Datenbankabgleich das Muster eines autorisierten Gesichts, woraufhin die smarte Haustüre entriegelt wird –, sondern um wesentlich komplexere, kulturell voraussetzungsreichere und oftmals mit Mehrdeutigkeiten konfrontierte Erkennungsleistungen. Anders gesagt: Es geht um Informationen, von deren Bildhaftigkeit nicht ohne weiteres (und sicher nicht restlos) abstrahiert werden kann. Lukas Rosenfelder hat dazu bemerkt:

34 Zu ähnlichen Ergebnissen kommt eine Forschungsgruppe, die zu Neutralisierungsstrategien arbeitet: Jieyu Zhao u. a.: Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-level Constraints, in: *arxiv.org*, 29.7.2017. Vgl. dazu auch: Megan Garcia: How to Keep Your AI from Turning into a Racist Monster, in: *wired.com*, 13.2.2017.

35 Robert Gehl, Lucas Moyer-Horner, Sara K. Yeo: Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science, in: *Television and New Media*, 16.12.2016, 1–19, hier 2.

36 Gillespie: *Custodians of the Internet*, 107.

37 Vgl. zu Spezifik und Reichweite bildsensorischer Operationen im IoT allgemein: Simon Rothöhler: *Das verteilte Bild. Stream – Archiv – Ambiente*, Paderborn 2018, 225–274.

38 Philip E. Agre: Surveillance and Capture: Two Models of Privacy, in: *The Information Society*, Vol. 10, 1994, 101–127, hier 114.

> Bildanalyse heißt: Ich habe ein Bild, lasse einen Algorithmus über das Bild laufen, und am Schluss habe ich eine Information, die nicht mehr bildhaft ist. Bildanalyse ist sehr viel schwieriger, weil der Computer kein Konzept von Bild hat. […] Für den Computer ist ein Bild ein Haufen ungeordneter Zahlen, die aussagen: An der Position X/Y habe ich die Helligkeit Z. Das ist kein Bild, aber der Computer sieht nur das.[39]

Im gegenwärtig installierten *moderation apparatus* der Plattformbetreiber übersetzt sich diese «Konzeptlosigkeit» des Computers in Sachen Bild in eine anbieterseitig organisierte Produktionskette, bei der menschliche Akteur_innen auf verschiedenen Ebenen die wesentlichen Arbeiten übernehmen: als Mitglieder in internen *Policy*-Teams, die Regulationsvorgaben dekretieren (und eine relativ homogene Elite darstellen);[40] als Programmierer_innen, die durch *crowd worker* angehäufte Trainingsdaten konfigurieren und die Ergebnisse algorithmischer Durchmusterung bewerten; als Ausbilder_innen, die Moderator_innen bezüglich richtlinienkonformer Bildanalysen instruieren; sowie als mikroarbeitende Reviewer_innen, die Einzelbildbeurteilungen im Sekundentakt operationalisieren. Die informationstechnischen Akteur_innen übernehmen hier in erster Linie die Aufgabe, potenziell fragwürdiges Bildmaterial vorzuselektieren (*flagging*), das dann im entscheidenden Schritt aber immer noch in vielen Fällen einer interpretativen menschlichen Evaluation unterzogen werden muss.

Was der Computer aber schon jetzt effizient umsetzt, ist die Filterung von visuellem Content, dessen spezifischer «Haufen ungeordneter Zahlen» bereits als Code illegaler Bildformen identifiziert worden ist. Das bekannteste (und wegen evidenter Haftungsrisiken am weitesten verbreitete) Beispiel ist PhotoDNA – eine 2009 von Microsoft entwickelte Filtersoftware, die gegenwärtig von fast allen großen Plattformbetreibern eingesetzt wird, um tatsächlich jeden einzelnen userseitigen Upload mit einer von der NGO National Center for Missing and Exploited Children (NCMEC) verwalteten Datenbank abzugleichen, in der die Bildcodes bekannter kinderpornografischer Inhalte gespeichert sind. Über eine Hashfunktion, die den Bilddatensatz zu einer vergleichsweise aufwandslos komputierbaren numerischen Zeichenkette codiert, also als Hashwert komprimiert, sind die derzeit rund 80 Millionen Bilder der NCMEC-Datenbank «forensisch» identifizierbar[41] (und zwar auch dann, wenn durch userseitig vorgenommene Bildmanipulationen die phänomenalisierte Bildform modifiziert wurde).

Weil das von Gillespie angesprochene Problem des industriellen Maßstabs erforderlicher Content-Moderation mit der Ausbreitung diverser Live-Streaming-Applikationen – bei denen ein rechtzeitiges Interventionsregime beinahe echtzeitlich, als *real-time content-moderation* operationalisiert werden müsste – dringlicher wird, sehen sich die jüngst auch von gesetzgeberischer Seite zumindest etwas kritischer adressierten Plattformbetreiber zunehmend gezwungen, die in großem Umfang notwendige Inhaltsbeobachtung und -beurteilung nachdrücklicher über technische Aktanten abzuwickeln.[42] Das ebenfalls von Hany Farid betreute, auf extremistischen Content terroristischer Organisationen ausgerichtete PhotoDNA-Nachfolgeprojekt eGlyph ist zwar

39 Lukas Rosenthaler, zit. n.: Migration der Daten, Analyse der Bilder, persistente Archive. Rudolf Gschwind und Lukas Rosenthaler im Gespräch mit Ute Holl, in: *Zeitschrift für Medienwissenschaft*, Nr. 2, 2010, 103–111, hier 106. Martin Warnke spricht diesbezüglich von einer Übersetzungsgrenze: «Wir haben es bei Zahl, Schrift und Bild mit drei Basismedien zu tun, die zwar seitens des Codes, aber nicht seitens der kulturellen Praxis […] ineinander überführbar sind.», Warnke: *Bildersuche*, 33–35.

40 «[T]he policies of these enormous, global platforms, and the labor crucial to their moderation efforts, are overseen by a few hundred, largely white, largely young, tech-savvy Californians who occupy a small and tight social and professional circle», Gillespie: *Custodians of the Internet*, 119.

41 Vgl. Hany Farid: *Photo Forensics*, Cambridge 2016. Microsoft, YouTube, Twitter und Facebook betreiben seit 2016 auch eine gemeinsame *hashed database* zu «terroristischem Content», vgl. Sarah Perez: Facebook, Microsoft, Twitter and YouTube Collaborate to Remove ‹Terrorist Content› from Their Services, in: *TechCrunch*, dort datiert 6.12.2016, tcrn.ch/2g3Hm6n, gesehen am 28.7.2018.

42 Gillespie erwähnt als Beispiel die 2015 von Twitter übernommene Softwarefirma Madbit, die die automatische Filterung von NSFW (*not safe for work*)-Content verspricht. Vgl. ders.: *Custodians of the Internet*, 107.

ebenfalls nur bedingt automatisiert – und wie die meisten Lösungsansätze in einen «broader sociotechnical apparatus»[43] eingebettet, der für die Semantisierung und Kontextualisierung von Bildinhalten bestimmte Skripte abarbeitet –, geht aber dennoch einen Schritt über die Kalkulation bildspezifisch errechneter ‹Fingerabdrücke› hinaus. Zum einen werden hier auch umfangreichere Videodaten gehasht und durchmustert (um Moderator_innen in zeitökonomischer Optimierung direkt an automatisch markierte Stellen eines Videodatensatzes zu führen). Zum anderen wird versucht, nicht nur bereits per *crowd work* gelabelten Content zu adressieren, sondern auch neues Bildmaterial zu erfassen, bei dem nur bestimmte Bildinhaltssegmente in entsprechenden Blacklist-Datenbanken liegen. Die Gesichter notorischer Terrorist_innen oder auch nur einschlägige Logos radikaler Gruppierungen sollen als bekannte Muster auch in neugeneriertem, bislang unmoderiertem Content zuverlässig algorithmisch detektiert und gegebenenfalls ausgeflaggt werden, sind aber dann, wie in der aktuellen Praxis, weiterhin einer entscheidungsverantwortlichen menschlichen Bildprüfung zuzuführen, die abschließend zu bewerten hat, ob es sich um legitimen News-Content, eine Parodie oder eben Terrorpropaganda handelt.

Dass sich die von Plattformbetreibern aus strategischen Motiven der Verantwortungsdelegation lancierte Vorstellung einer ‹neutralen›, vom *human bias* ‹befreiten› Automatisierung[44] derzeit insbesondere auf die moderierende Verarbeitung von Bildinhalten konzentriert, reagiert auf den Umstand, dass Bilddaten immer aufwandsloser akquirierbar, immer verzögerungsfreier verteilbar werden. Mit Blick auf die entstandenen soziotechnischen Ensembles, die versuchen, die Grenzverläufe zwischen menschlicher und nichtmenschlicher Bildprozessierung zu verschieben, zeigt sich: Für die technischen Aktanten wäre Content-Moderation leichter, wenn die zirkulierenden Informationen aus Sicht menschlicher Akteur_innen keine Bilder hätten.

43 Gillespie: *Custodians of the Internet*, 101.

44 Ed Finn spricht hier mit Blick auf Content sortierende Empfehlungsalgorithmen von «algorithmic evangelism» und einem «magischen» Effizienzversprechens digitaler «Kulturmaschinen». Vgl. ders.: *What Algorithms Want. Imagination in the Age of Computing*, Cambridge, London, 22 ff.