ARTICLE 19

unesco
In partnership with UNESCO

Funded by the European Union

Buku Panduan Moderasi Konten dan Kebebasan Berekspresi

Agustus 2023

SOCIAL MEDIA 4PEACE

ARTICLE 19

T: +44 20-7324 2500

F: +44 20-7490 0566

E: info@article19.org

W: www.article19.org

Tw: @article19org

Fb: facebook.com/article19org

© ARTICLE 19, 2024

Panduan ini dipublikasikan dengan dukungan pendanaan dari **Uni Eropa** dan **UNESCO**.
Penggunaan nama dan sebutan, serta penyajian materi dalam keseluruhan panduan ini tidak menyiratkan pendapat apa pun dari UNESCO dan Uni Eropa mengenai status hukum suatu negara, wilayah, kota, atau daerah tertentu, maupun otoritasnya, ataupun penetapan garis batas dan perbatasannya.

Penulis panduan ini bertanggung jawab atas pemilihan dan penyajian fakta-fakta yang dimuat di dalam buku ini, sekaligus opini yang diungkapkan di dalamnya, yang belum tentu merupakan pendapat UNESCO maupun Uni Eropa, dan tidak mengikat organisasi-organisasi tersebut.

Karya ini berada di bawah lisensi Creative Commons Attribution-Non-Commercial-ShareAlike 4.0. Anda boleh menyalin, menyebarkan, dan menampilkan karya ini serta membuat karya turunan, sepanjang:

1) menyebutkan ARTICLE 19 sebagai penyusun panduan ini;

2) tidak menggunakan karya ini untuk tujuan komersial;

3) menyebarkan karya turunan yang didasarkan atas panduan ini di bawah lisensi yang sama.

Untuk mengakses naskah legal utuh dari liesensi ini, silakan kunjungi: https://creativecommons.org/licenses/by-nc-sa/4.0/

Daftar Isi

Pengantar

Panduan ini dipublikasikan sebagai bagian dari proyek United Nations Educational, Scientific and Cultural Organization (UNESCO's) **Social Media 4 Peace** yang didanai Uni Eropa.

Kuasa perusahaan media sosial atas kebebasan berekspresi

Di awal kehadirannya, platform media sosial kerap dipandang sebagai daya yang kuat untuk tujuan kebaikan, memerdekakan kebebasan berekspresi, memungkinkan koneksi antarmanusia, dan menjadi ujung tombak revolusi demokrasi di berbagai belahan dunia. Persepsi itu kini berubah. Kini, segelintir perusahaan media sosial besar bertindak seperti penjaga gerbang yang mengendalikan apa yang dapat dilihat atau diucapkan secara daring. Perusahaan-perusahaan ini memiliki dampak langsung atas dinamika distribusi konten serta keragaman media daring dan kebebasan berekspresi.

Kekuatan dan pengaruh yang signifikan ini berpadu dengan fakta bahwa model bisnis perusahaan-perusahaan media sosial terbesar kerap didasarkan atas pengumpulan data dalam jumlah besar mengenai pengguna dan perilaku daring mereka (data perilaku) serta monetisasi data tersebut melalui iklan daring (dengan target spesifik). Model bisnis ini secara signifikan memengaruhi hak pengguna atas privasi dan dapat berdampak negatif terhadap kebebasan berekspresi. Yang menjadi kekhawatiran khusus adalah penyebaran 'ujaran kebencian' dan 'disinformasi' melalui platform daring.[1] Perusahaan media sosial dituduh telah memprioritaskan profit dengan mengorbankan keamanan pengguna melalui penggunaan algoritma yang mendorong konsumsi konten berbahaya, termasuk 'ujaran kebencian' dan 'disinformasi'. Muncul desakan yang terus meningkat agar perusahaan media sosial menggenjot upaya moderasi konten untuk melawan konten bermasalah tersebut.

Tujuan dan struktur panduan ini

Moderasi konten mencakup berbagai rangkaian tindakan dan alat yang digunakan perusahaan media sosial untuk menangani konten ilegal atau melanggar standar

komunitas perusahaan dalam platformnya. Moderasi konten dipengaruhi berbagai faktor, mulai dari penerapan model bisnis perusahaan media sosial; tekanan pemasang iklan dan regulator untuk menghindari ujaran yang secara sosial tidak diinginkan, berbahaya, atau ilegal; hingga perlunya melindungi kebebasan berekspresi secara daring.

Dengan berfokus pada beberapa platform media sosial terbesar, panduan ini memberikan gambaran umum singkat tentang situasi terkini moderasi konten, serta isu-isu terpenting yang muncul dari perspektif kebebasan berekspresi.

Struktur panduan ini adalah sebagai berikut:

- Pertama, panduan ini menguraikan standar-standar perlindungan kebebasan berekspresi daring yang berlaku untuk moderasi konten, dengan fokus utama pada 'ujaran kebencian' dan 'disinformasi'.

- Kedua, panduan ini membahas hubungan kontraktual antara pengguna dan perusahaan-perusahaan media sosial terbesar melalui syarat dan ketentuan layanan serta standar komunitas yang mengatur ujaran daring, dan isu-isu yang muncul akibat penggunaan kontrak untuk meregulasi pembicaraan.

- Ketiga, panduan ini menjelaskan kerangka regulasi moderasi konten, terutama konsep kewajiban perantara dan tren terkini yang mengarah pada regulasi yang lebih ketat terhadap perusahaan media sosial.

- Terakhir, panduan ini akan membahas proses moderasi konten yang lazim diterapkan oleh perusahaan-perusahaan media sosial terbesar—termasuk sistem terotomatisasi, peninjauan oleh manusia (*human reviewer*), dan pelaporan oleh pihak ketiga—dengan fokus utama pada kelemahan sistem moderasi konten otomatis.

Untuk analisis tantangan spesifik mengenai terlepasnya praktik moderasi konten perusahaan-perusahaan media sosial terbesar dari komunitas lokal tempat konten yang dimoderasi tersebut diproduksi dan disebarkan – berdasarkan studi atas praktik-praktik terkini di Bosnia dan Herzegovina, Kenya, dan Indonesia – silakan baca laporan ARTICLE

19 *Moderasi Konten dan Kebebasan Berekspresi: Menjembatani Media Sosial dan Masyarakat Sipil Setempat*.

Mengenai proyek ini

Panduan ini merupakan bagian dari proyek **Social Media 4 Peace** yang dilaksanakan UNESCO dan ARTICLE 19 di Bosnia dan Herzegovina, Kenya, Indonesia, dan Kolombia dengan dukungan Uni Eropa. Tujuan garis besar proyek ini adalah memperkokoh ketangguhan masyarakat terhadap konten yang disebarkan secara daring dan berpotensi membahayakan, terutama 'ujaran kebencian' dan 'disinformasi', sekaligus melindungi kebebasan berekspresi dan berkontribusi terhadap pengembangan narasi-narasi damai melalui teknologi digital, terutama media sosial. Kontribusi ARTICLE 19 dalam proyek ini difokuskan pada kekhawatiran atas praktik-praktik moderasi konten yang saat ini diterapkan platform-platform media sosial terbesar di keempat negara target.

Standar internasional hak asasi manusia (HAM) yang berlaku

Gambaran umum

Bagaimana standar hak asasi manusia (HAM) dan kebebasan berekspresi diterapkan dalam moderasi konten? Moderasi konten terutama dipengaruhi dan dibentuk oleh tindakan perusahaan media sosial di satu sisi dan negara, khususnya melalui undang-undang dan regulasi, di sisi lain. Kedua pihak memiliki tanggung jawab menurut hukum HAM internasional, termasuk dalam perlindungan terhadap kebebasan berekspresi, meskipun dalam derajat yang berbeda

Pertama, bab ini akan diawali dengan uraian tentang standar-standar yang berlaku untuk perlindungan kebebasan berekspresi daring yang seharusnya menjadi patokan tindakan apa pun yang diambil negara dan perusahaan media sosial dalam moderasi konten. Kedua, bab ini akan membahas secara singkat gambaran umum penerapan standar-standar internasional kebebasan berekspresi terhadap 'ujaran kebencian' dan 'disinformasi' – dua kategori ujaran yang mencakup beragam bentuk ungkapan, tetapi tanpa definisi yang seragam dalam hukum HAM internasional. Ketiga, dalam bab ini juga akan dijelaskan lingkup tanggung jawab HAM perusahaan media sosial, perbedaannya dengan tanggung jawab negara, dan implikasinya dalam moderasi konten.

Jaminan hak atas kebebasan berekspresi

Hak atas kebebasan berekspresi dilindungi oleh Pasal 19 Deklarasi Universal Hak Asasi Manusia (UDHR),[2] dan mendapatkan kekuatan hukum melalui Pasal 19 Kovenan Internasional tentang Hak-hak Sipil dan Politik (ICCPR)[3] serta perjanjian-perjanjian regional.[4]

Hak atas kebebasan berekspresi memiliki cakupan yang luas. Hak tersebut mewajibkan negara menjamin kebebasan semua orang untuk mencari, menerima, atau membagikan informasi dan gagasan apa pun, tanpa memandang perbatasan, melalui media apa pun yang dipilih oleh seseorang. Negara memiliki kewajiban untuk tidak turut campur tangan

dalam sirkulasi informasi dan gagasan, atau secara sewenang-wenang membatasi kebebasan berekspresi. Negara juga berkewajiban memajukan kondisi-kondisi kondusif untuk kebebasan berekspresi dan melindungi individu dari campur tangan tidak proporsional pihak-pihak swasta.[5] Dalam konteks ini, dapat dikatakan negara wajib mengambil tindakan positif untuk menjamin bahwa hak atas kebebasan berekspresi dapat secara efektif dinikmati daring, misalnya dengan memasukkan prosedur-prosedur keamanan dalam kerangka hukum penghapusan konten daring.[6]

Pada 2011, Komite HAM PBB, lembaga traktat yang memantau kepatuhan negara atas ICCPR, mengklarifikasi bahwa hak atas kebebasan berekspresi juga berlaku bagi segala bentuk ekspresi berbasis elektronik dan internet.[7]

Toolkit Global untuk Aktor Yudisial: Standar Internasional Kebebasan Berekspresi, Akses Informasi dan Keselamatan Jurnalis dari UNESCO memiliki modul spesifik mengenai tantangan-tantangan terbaru dalam perlindungan kebebasan berekspresi di internet.

Batasan-batasan hak atas kebebasan berekspresi

Berdasarkan standar HAM internasional, dalam kondisi luar biasa negara dapat membatasi hak atas kebebasan berekspresi, sepanjang pembatasan tersebut sesuai dengan Uji Tiga Bagian yang diatur dalam Pasal 19 (3) ICCPR.[8] Ketentuan ini mensyaratkan bahwa pembatasan harus:

- **Merupakan ketentuan hukum**
    - Pembatasan apa pun harus dirumuskan dengan cukup presisi agar memungkinkan individu mengatur perilakunya sendiri berdasarkan aturan tersebut. Pembatasan yang terlalu luas tidak diperbolehkan.
    - Misalnya, naskah hukum yang mengkriminalisasi ‘penyebaran rumor dengan cara yang dapat berdampak terhadap ketenteraman umum’ dapat ditafsirkan berbeda – termasuk makna ‘rumor’, ‘menyebarkan’, or ‘ketenteraman umum’ – dan tidak memenuhi standar kualitas yang dipersyaratkan Uji Tiga Bagian.

- **Memiliki tujuan yang sah**
    - Pembatasan hanya diizinkan untuk (a) menghormati hak-hak atau reputasi orang lain, dan (b) menjaga keamanan nasional, ketertiban umum, atau kesehatan dan moral masyarakat.
    - Misalnya, pihak yang berwenang tidak boleh membatasi hak kebebasan berekspresi dengan tujuan menjamin penghormatan terhadap 'teks agama yang diakui' atau melindungi agama dari olok-olok.
- **Perlu dan proporsional untuk masyarakat berdemokrasi**
    - Suatu pembatasan harus menunjukkan hubungan langsung dan segera antara ekspresi tertentu dan kepentingan yang dilindungi. Selain itu, jika suatu tujuan dapat dicapai dengan tindakan yang tidak terlalu mengekang tetapi sama efektifnya dengan pembatasan yang lebih ketat, langkah yang tidak terlalu mengekang inilah yang harus diterapkan.
    - Misalnya, vonis penjara untuk pencemaran nama baik merupakan pembatasan hak kebebasan berekspresi yang tidak proporsional. Penyelesaian kasus pencemaran nama baik seharusnya diputuskan lewat mekanisme hukum perdata atau solusi alternatif, termasuk permohonan maaf, koreksi, dan penggunaan hak jawab. Cara-cara ini mampu mengatasi masalah tercemarnya reputasi seseorang dengan efektif, tanpa perlu menimbulkan dampak yang membekukan kebebasan berekspresi. Sanksi penjara terlalu berat untuk perbuatan yang berdampak terhadap reputasi seseorang. Umumnya sanksi penjara dijatuhkan hanya untuk pelanggaran ujaran berat, seperti hasutan untuk melakukan genosida.

Uji Tiga Bagian dapat diterapkan pada semua tindakan yang diambil negara, termasuk legislasi, kebijakan, dan putusan terhadap individu. Terkait moderasi konten, Uji Tiga Bagian umumnya dapat diterapkan saat melakukan penilaian terhadap kerangka regulasi yang mengatur perusahaan media sosial atau terhadap permintaan negara untuk mengakses data pengguna atau membatasi konten.

Sebagaimana akan dibahas lebih mendetail nanti, perusahaan media sosial juga bertanggung jawab menghormati HAM, termasuk kebebasan berekspresi, dan harus menjamin bahwa produk dan layanan mereka sejalan dengan standar HAM internasional. Karena itu, Uji Tiga Bagian dapat digunakan juga untuk menganalisis apakah syarat dan ketentuan atau keputusan moderasi konten individu suatu perusahaan sejalan dengan standar internasional kebebasan berekspresi.

Kebebasan berekspresi dan 'ujaran kebencian'

Pembahasan mendetail mengenai berbagai jenis 'ujaran kebencian' dan bagaimana menanganinya dalam kerangka HAM tidak termasuk dalam cakupan buku panduan ini.[9] Karena itu, bab ini hanya akan merangkum perbedaan mendasar dan prinsip-prinsip pembatasan yang masih diperbolehkan untuk 'ujaran kebencian' berdasarkan standar internasional kebebasan berekspresi.

Tidak ada definisi yang disepakati mengenai 'ujaran kebencian' dalam hukum HAM internasional. Sederhananya, 'ujaran kebencian' adalah ekspresi kebencian apa pun yang mendiskriminasi seseorang. Karenanya, respons terhadap 'ujaran kebencian' sebagian besar didasarkan atas perlindungan persamaan bagi hak dan prinsip-prinsip non-diskriminasi yang juga diatur dalam ICCPR. Prinsip non-diskriminasi melindungi individu dari perlakuan berbeda, eksklusi, dan pembatasan berdasarkan karakteristik yang dilindungi. Sebagian karakteristik tersebut tercantum dalam Pasal 26 ICCPR; termasuk di antaranya ras, warna kulit, jenis kelamin, bahasa, agama, pendapat politik dan pendapat lainnya, asal negara atau kelompok sosial, status properti, kelahiran, dan lainnya.

Meski demikian, 'ujaran kebencian', yang didefinisikan secara luas sebagai ekspresi kebencian yang mendiskriminasi seseorang belum mencakup konsekuensi tertentu. Definisi paling sederhana ini masih mencakup banyak jenis ekspresi, termasuk yang diperbolehkan berdasarkan hukum HAM internasional. Karena itu definisi ini masih terlalu samar untuk digunakan dalam mengidentifikasi ekspresi yang mungkin dapat dibatasi berdasarkan hukum HAM internasional.

Karenanya, ARTICLE 19 mengusulkan ‘ujaran kebencian’ dibagi ke dalam tiga kategori berikut.

‘Ujaran kebencian’ yang harus dilarang

Hukum pidana internasional dan Pasal 20 (2) ICCPR mewajibkan negara melarang bentuk-bentuk ‘ujian kebencian’ parah tertentu, termasuk dengan instrumen pidana, perdata, dan administratif.

Pasal 20 (2) ICCPR mewajibkan negara menerapkan hukum yang melarang ‘segala jenis anjuran kebencian kebangsaan, rasial, atau agama yang mengandung hasutan diskriminasi, permusuhan, atau kekerasan’.

Rencana Aksi Rabat yang memberikan pedoman lengkap bagi negara untuk menjalankan kewajibannya berdasarkan Pasal 20 (2) ICCPR,[10] menguraikan uji enam bagian untuk menentukan apakah suatu ujaran termasuk pelanggaran pidana berdasarkan Pasal 20 ICCPR. Panduan tersebut mengharuskan pertimbangan atas (1) konteks sosial dan politik, (2) status pembicara, (3) niat pembicara menghasut audiens agar menjadikan kelompok tertentu sebagai sasaran, (4) konten dan bentuk ujaran, (5) keluasan dan besarnya penyebaran ujaran tersebut, dan (6) potensi bahaya, termasuk tingkat kegentingannya.

Ilustrasi Kasus

Menjelang pemilihan presiden yang diwarnai persaingan sengit, presiden petahana berpidato di serangkaian kampanye akbar. Dalam acara kampanye tersebut, presiden petahana mengangkat rumor bahwa pendukung oposisi, yang sebagian besar terdiri atas kelompok etnis berbeda, sedang mempersenjatai diri dan menjadi ancaman atas eksistensi kelompok pendukungnya. Seiring meningkatnya ketegangan, dia menggunakan bahasa rasis, bahkan mengungkit pembunuhan massal yang terjadi di negara tersebut beberapa dekade sebelumnya, serta mengajak pendukungnya untuk bertindak segera demi mengamankan kemenangan dalam pemilu tersebut.

Dalam hal ini, Presiden telah melakukan 'ujaran kebencian' dan bisa dikatakan mencapai batas definisi anjuran kebencian yang mengandung hasutan kekerasan. Presiden memahami dan mengeksploitasi ketegangan antaretnis di tengah masyarakat, dan mengerti bahwa sebagai politisi yang berpengaruh, penggunaan istilah tertentu akan dipahami dan kemungkinan besar ditindaklanjuti oleh oknum-oknum di antara peserta kampanye dengan melakukan kekerasan terhadap kelompok etnis yang diidentikkan dengan oposisi.

'Ujian kebencian' yang boleh dilarang

Negara dapat melarang bentuk 'ujaran kebencian' lainnya, sepanjang memenuhi persyaratan Pasal 19 (3) ICCPR. Termasuk di dalamnya ancaman, serangan, dan pelecehan yang termotivasi diskriminasi dan ditargetkan terhadap individu tertentu.

Ilustrasi Kasus

Pasangan sesama jenis, keduanya perempuan, mengalami konfrontasi dengan sesama penumpang kereta, yang meneriakkan makian seksis dan homofobik sehingga keduanya khawatir akan segera terjadi kekerasan fisik.

Di banyak yurisdiksi, insiden ini dapat dituntut sebagai kejahatan bermotif diskriminasi. Tindakan penumpang yang melakukan kekerasan verbal tersebut masuk ke dalam tipologi luas 'ujaran kebencian', dan merupakan pidana kekerasan. Ancaman kekerasan yang dapat dibuktikan dalam ujaran tersebut menjadikan perbuatan ini tindak pidana, dan karena bercirikan bias, konten ujaran tersebut juga merupakan bukti adanya motif diskriminasi.

'Ujaran kebencian' yang dilindungi hukum

Ujaran jenis ini harus dilindungi dari pembatasan berdasarkan Pasal 19 (2) ICCPR, walaupun menimbulkan kekhawatiran mengenai intoleransi dan diskriminasi, sehingga membutuhkan respons kritis dari negara.

Ilustrasi Kasus

Seorang remaja pria dengan jumlah pengikut sedikit mengunggah candaan yang menyinggung dan seksis, meremehkan peristiwa hilangnya seorang siswi yang kemungkinan menjadi korban pembunuhan di daerah tersebut. Twit ini memicu reaksi daring yang keras terhadap remaja tersebut, dan dia kemudian menghapus twit tersebut. Walaupun bentuk komunikasi ini tidak sopan dan mencerminkan masalah misogini yang lebih luas di masyarakat, remaja tersebut tidak bermaksud menghasut untuk melakukan tindakan berbahaya terhadap kelompok tertentu, dan jelas dia tidak memiliki pengaruh sebesar itu terhadap pengikutnya. 'Ujian kebencian' semacam ini dapat ditanggapi dengan intervensi lunak dari tokoh lokal yang berwenang, misalnya guru-guru di sekolah remaja tersebut, atau pemimpin masyarakat lainnya, tetapi tidak layak dijadikan pembenaran bagi negara untuk memberlakukan sanksi atau pembatasan lainnya.

Kategori-kategori 'ujaran kebencian' yang berbeda ini harus dipertimbangkan ketika mengukur pembatasan terhadap kebebasan berekspresi, terutama yang diberlakukan negara, termasuk dalam isu moderasi konten.

Sebagaimana dijelaskan lebih mendetail di bawah, perusahaan media sosial kerap menghapus konten yang dilindungi standar internasional kebebasan berekspresi. Penghapusan seringkali dilakukan melalui penerapan kebijakan anti-'ujaran kebencian' dan aturan lainnya di platform tersebut. Misalnya, TikTok tidak mengizinkan 'ideologi kebencian apa pun', termasuk 'misogini'. Pelarangan misogini tanpa pandang bulu semacam ini tidak akan memenuhi persyaratan Pasal 19 dan 20 ICCPR. TikTok juga tidak

mengizinkan ‘penyangkalan terhadap peristiwa sejarah yang terdokumentasikan dengan baik yang merugikan kelompok-kelompok berdasarkan atribut yang dilindungi’ dan menyatakan bahwa platform tersebut memberikan ‘beberapa perlindungan terkait usia’.[11]

Kebijakan dapat sangat berbeda dari satu platform ke platform lainnya. Misalnya, TikTok mencantumkan ‘gender’ dan ‘identitas gender’ dalam daftar karakteristik yang dilindungi, sementara Meta hanya mencantumkan ‘identitas gender’.

Meskipun perusahaan media sosial sebagai lembaga swasta berhak memberlakukan standar komunitas yang lebih ketat dibandingkan persyaratan yang diuraikan sebelumnya, syarat dan ketentuan serta keputusan moderasi konten yang mereka terapkan harus setidaknya sejalan dengan norma dan prinsip-prinsip HAM internasional.

Contoh

Dalam kasus ‘puisi Rusia’, Dewan Pengawas Meta – suatu mekanisme yang dibentuk oleh Meta untuk meninjau keputusan moderasi konten kasus-kasus pilihan dan memberi panduan kebijakan umum moderasi konten di Meta melalui masukan opini terhadap kebijakan – melakukan penilaian atas keputusan Meta menghapus postingan Facebook yang diunggah pada April 2022. Postingan tersebut diunggah setelah Rusia secara sewenang-wenang menginvasi Ukraina dan menyamakan tentara Rusia dengan Nazi, disertai kutipan puisi yang mengajak melakukan pembunuhan terhadap kaum fasis. Dewan Pengawas menyatakan bahwa ‘[D]alam mengukur risiko yang ditimbulkan konten kekerasan atau kebencian, sebagai panduan Dewan [Pengawas] umumnya menggunakan tes enam faktor yang dijelaskan dalam Rencana Aksi Rabat untuk menangani anjuran kebencian nasional, rasial dan agama yang mengandung hasutan kebencian, diskriminasi atau kekerasan’.

Dalam hal ini, Dewan Pengawas menemukan bahwa, meskipun konteks konflik bersenjata masih berlangsung antara Rusia dan Ukraina, dan terdapat referensi budaya tendensius dalam postingan pengguna tersebut, kecil kemungkinan postingan itu – yang memperingatkan potensi terjadinya lingkaran setan kekerasan – dapat menimbulkan bahaya. Dewan Pengawas berkesimpulan bahwa penghapusan konten di awal tidak diperlukan.

Kebebasan berekspresi dan ‘disinformasi’

Seperti ‘ujaran kebencian’, konsep ‘disinformasi’, ‘misinformasi’, propaganda, dan ‘informasi palsu’ tidak memiliki definisi yang disepakati dalam hukum HAM internasional dan regional.[12] Biasanya upaya mendefinisikan konsep-konsep ini dalam hukum nasional dan standar regional difokuskan pada pelarangan informasi ‘palsu’ atau ‘menyesatkan’ yang dapat mengakibatkan ‘bahaya’ atau kerugian tertentu.[13]

Karena upaya meredam ‘disinformasi’ melibatkan pembatasan kebebasan berekspresi, respons legal dan kebijakan terhadap masalah-masalah ini harus didasarkan atas standar internasional kebebasan berekspresi. Tepatnya, respons tersebut harus lolos Tes Tiga Bagian sesuai Pasal 19 (3) ICCPR. Pembatasan ‘disinformasi’ hanya diperbolehkan jika secara eksplisit terkait perlindungan tujuan-tujuan sah yang dimuat dalam Pasal 19 (3) – seperti penghormatan atas hak atau reputasi orang lain dan menjaga keamanan nasional, ketertiban umum, atau kesehatan dan moral masyarakat – atau Pasal 20 ICCPR – melarang propaganda perang dan anjuran kebencian nasional, rasial dan agama yang mengandung hasutan diskriminasi, permusuhan, dan kekerasan. Kepalsuan atau menyesatkannya suatu informasi saja belum cukup menjadi alasan untuk membatasi penyebarannya. Di samping itu, agar sejalan dengan standar kebebasan berekspresi internasional, pembatasan ‘disinformasi’ juga harus diterapkan secara proporsional.

Pada umumnya konsekuensi berbahaya ‘disinformasi’ harus ditangani dengan langkah-langkah pemberdayaan, misalnya dengan menjamin ekosistem media yang bebas, independen, dan beragam, serta mendorong terpaparnya masyarakat terhadap informasi yang paling beragam. Di sisi lain, pelarangan dan pembatasan legal lain terhadap penyebaran informasi palsu rawan disalahgunakan dan berdampak serius terhadap diskursus politik. Pemberlakuan kewajiban hukum mengenai ‘kebenaran’ menciptakan instrumen yang kuat untuk mengendalikan arus informasi dan ide, yang bisa jadi alat berbahaya di tangan otoritas publik. Dan kenyataannya, konsep ‘disinformasi’. ‘misinformasi’, propaganda, dan ‘informasi palsu’ telah digunakan dan disalahgunakan

pemegang kekuasaan untuk menekan suara-suara yang berseberangan dan mendiskreditkan informasi yang tidak disukai.

Walaupun legislasi yang mengkriminalisasikan penyebaran ‘berita palsu’ bukanlah barang baru, beberapa tahun terakhir banyak bermunculan upaya-upaya regulasi seperti ini. Pemerintahan dari seluruh dunia membuat atau memperbarui definisi pelanggaran ‘disinformasi’, termasuk melalui legislasi yang konon bertujuan melawan kejahatan siber**,** diberlakukan dalam konteks pandemi COVID-19, atau untuk menekan peliputan konflik bersenjata.[14] Legislasi semacam ini telah digunakan sebagai dasar penangkapan dan penuntutan atas bloger, jurnalis dan kritikus pemerintah.

Contoh

Pada November 2022, jurnalis asal Senegal bernama Pape Alé Niang ditangkap dan ditahan pihak berwenang Senegal. Tuduhan yang dikenakan di antaranya adalah 'penyebaran berita palsu yang dapat mendiskreditkan lembaga pemerintah'. Tuduhan tersebut muncul akibat liputan Pape tentang dugaan pemerkosaan oleh pemimpin utama oposisi Senegal, Ousmane Sonko, yang menimbulkan ketegangan politik dalam negeri. Pape Alé Niang akhirnya dibebaskan pada 10 Januari 2023 meski tetap berada dalam pengawasan pengadilan. Pape Alé Niang adalah salah satu jurnalis yang meliput isu kepentingan publik dan diadili dengan undang-undang yang mengkriminalisasi penyebaran 'informasi palsu'.

Contoh

Pada Desember 2022, salah satu tokoh oposisi kenamaan Rusia, Ilya Yashin, divonis penjara selama delapan setengah tahun karena menggunakan kanal YouTube-nya untuk mengecam pembunuhan ratusan warga sipil Ukraina oleh tentara pendudukan Rusia di Bucha. Dalam video tersebut, Ilya menunjukkan foto-foto dan berita dari lokasi pembantaian yang bersumber dari *BBC* dan media lainnya. Pengadilan memutuskan bahwa Ilya telah dengan sengaja menyebarkan informasi palsu tentang angkatan bersenjata Rusia.

Vonis tersebut didasarkan atas undang-undang yang disahkan Rusia setelah invasi ilegalnya ke Ukraina yang menjatuhkan hukuman hingga 15 tahun penjara bagi terpidana penyebaran 'informasi palsu' atau informasi yang oleh otoritas Rusia dinilai palsu dalam liputan terkait perang.

Persoalan muncul bukan saja ketika otoritas publik menjadi penentu kebenaran, tetapi juga saat lembaga swasta yang mengendalikan ruang bicara kuasi-publik dapat menentukan informasi mana yang 'benar' dan mana yang 'salah' sekaligus menekan dan

membatasi informasi yang mereka anggap tidak benar. Regulasi negara yang mewajibkan platform mengambil tindakan untuk menekan informasi palsu, atau syarat dan ketentuan yang memuat larangan terhadap segala jenis 'disinformasi', dapat dianggap sebagai campur tangan sewenang-wenang atas hak pengguna platform untuk bebas berekspresi.[15]

Tanggung jawab HAM perusahaan media sosial

Negara adalah pemegang tugas utama menurut hukum HAM internasional. Negara berkewajiban menghormati, melindungi, dan memenuhi hak asasi manusia. Kewajiban ini mencakup tugas melindungi individu dari pelanggaran HAM oleh semua aktor masyarakat, termasuk kalangan bisnis. Negara harus mencegah, menyelidiki, menghukum, dan memberi ganti rugi atas pelanggaran HAM oleh aktor swasta. Walaupun beberapa perusahaan media sosial terbesar – seperti korporasi besar lainnya – boleh dibilang memiliki kekuasaan lebih besar daripada negara tertentu dan dapat berdampak besar pada hak asasi individu dan komunitas di tempat mereka beroperasi, sebagai bisnis, kewajiban HAM perusahaan-perusahaan ini tidak setara dengan negara.

Meski demikian, *Prinsip-Prinsip Panduan tentang Bisnis dan Hak Asasi Manusia: Menerapkan Kerangka Kerja Perserikatan Bangsa-Bangsa 'Perlindungan, Penghormatan dan Pemulihan* (Prinsip-prinsip Panduan) mengakui bahwa korporasi memiliki tanggung jawab untuk menghormati HAM, terlepas dari kewajiban negara maupun penerapan kewajiban tersebut.[16]

Prinsip-prinsip Panduan secara khusus merekomendasikan bahwa korporasi harus:[17]

- membuat pernyataan publik mengenai komitmen penghormatan terhadap HAM, yang didukung manajemen senior atau eksekutif;
- melakukan uji tuntas (*due diligence*) dan evaluasi dampak terhadap HAM untuk mengidentifikasi, mencegah, dan menanggulangi potensi dampak negatif terhadap HAM dalam operasi bisnisnya;

- secara sengaja memasukkan langkah-langkah perlindungan hak asasi manusia untuk menanggulangi dampak merugikan, dan bersama-sama bertindak untuk memperkuat pengaruh mereka dalam berhadapan dengan otoritas pemerintah;
- memantau dan mengkomunikasikan kinerja, risiko dan permintaan pemerintah; serta
- menyiapkan ganti rugi jika terjadi dampak negatif terhadap HAM.

Secara spesifik, perusahaan media sosial mungkin harus mengukur dan menanggulangi dampak negatif terhadap HAM yang muncul, misalnya dari cara mereka menanggapi permintaan pemerintah untuk menghapus suatu konten atau membuka akses data, juga praktik pengumpulan data, kurasi konten, dan sistem penargetan iklan. Jika terlibat dalam pelanggaran HAM, perusahaan media sosial harus menyediakan mekanisme pemulihan yang efektif bagi komunitas yang terdampak.

Dalam praktik moderasi konten, di antara langkah-langkah yang harus dilakukan perusahaan media sosial adalah mengeluarkan ketentuan dan syarat yang jelas dan tidak ambigu sejalan dengan norma dan prinsip HAM internasional,[18] memberi laporan transparansi mengenai permintaan pemerintah,[19] memastikan sanksi yang dijatuhkan atas ketidakpatuhan terhadap syarat dan ketentuan bersifat proporsional, dan memberi mekanisme pemulihan yang efektif bagi pengguna yang terdampak jika terjadi pelanggaran.[20]

Masyarakat sipil telah menyusun rekomendasi spesifik tentang tanggung jawab perusahaan media sosial dalam menghormati standar-standar HAM. Misalnya, Prinsip-prinsip Manila tentang Kewajiban Perantara menyatakan bahwa praktik pembatasan konten oleh perusahaan harus mematuhi uji kebutuhan dan proporsionalitas berdasarkan hukum HAM (Prinsip 4) dan harus menyediakan mekanisme pengaduan bagi pengguna untuk mengajukan banding atas keputusan perusahaan (Prinsip 5(c)).

Beberapa perusahaan media sosial juga telah mempublikasikan komitmen terhadap prinsip-prinsip HAM tertentu, baik dengan menerapkan kebijakan HAM, seperti Meta, atau dengan menjadi bagian inisiatif multi-pemangku kepentingan seperti Global Network

Initiative, yang anggotanya berkomitmen untuk berkolaborasi demi memajukan hak-hak pengguna atas kebebasan berekspresi dan privasi.

Untuk memahami sejauh mana tanggung jawab penghormatan terhadap hukum HAM internasional dapat dijadikan panduan keputusan moderasi konten individu, ada baiknya membaca keputusan-keputusan Dewan Pengawas Meta, karena otoritasnya bersumber dari kebijakan konten Meta dan tanggung jawab HAM Meta sesuai Prinsip-prinsip Panduan PBB.

Contoh

Dewan Pengawas mengevaluasi kasus terkait penghapusan konten yang diunggah pengguna Instagram pada Januari 2021 yang memuat gambar Abdullah Öcalan dengan komentar "Sudah waktunya bicara tentang upaya mengakhiri isolasi Öcalan di dalam penjara." Öcalan adalah salah satu pendiri dan anggota Partai Pekerja Kurdistan (PKK), yang dilabeli sebagai organisasi teroris di Turki. Öcalan telah mendekam di penjara sejak 1999. Baik PKK maupun Öcalan disebut sebagai entitas berbahaya dalam kebijakan Facebook tentang Individu dan Organisasi Berbahaya.

ARTICLE 19 memberikan tanggapan publik dengan menyatakan bahwa pembatasan kebebasan berekspresi di Facebook harus dipandu standar-standar HAM, bahwa kebijakan yang mensyaratkan pembatasan konten berdasarkan potensi hasutan terorisme harus memiliki definisi operasi yang sangat jelas, dan terdapat perbedaan signifikan antara pendekatan HAM terhadap 'konten terorisme' dan standar komunitas Facebook yang terlalu berfokus pada pembicara atau organisasi.

Dewan Pengawas memutuskan bahwa penghapusan postingan tersebut tidak konsisten dengan komitmen Meta untuk menghormati HAM, karena ujaran tentang kondisi penahanan seseorang merupakan ujaran yang dilindungi; standar komunitas tidak cukup jelas untuk lolos dari tes legalitas; dan penghapusan tersebut tidak dibutuhkan dan tidak proporsional, mengingat konten dalam hal ini tidak mengindikasikan dukungan terhadap aksi kekerasan yang dilakukan Öcalan atau oleh PKK.

Yurisprudensi lembaga-lembaga peradilan mengenai peran HAM dalam keputusan moderasi konten masih sangat sedikit. Salah satu pengecualian yang cukup menonjol adalah putusan Pengadilan Roma dalam perkara *Facebook* melawan *CasaPound*.

Contoh

Pengadilan Roma menolak pengajuan banding Meta (saat itu masih bernama Facebook) terhadap putusan sela yang memerintahkan diaktifkannya kembali akun partai sayap ekstrem kanan Italia, CasaPound. Meta telah mendeaktivasi akun tersebut tanpa pemberitahuan atau penjelasan. Di persidangan, Meta berargumen bahwa tindakan tersebut sah karena akun itu memuat konten yang termasuk 'ujaran kebencian' dan hasutan kekerasan yang melanggar Syarat dan Ketentuan Meta.

Pengadilan berpihak pada CasaPound dan menyatakan bahwa kontrak yang dilakukan oleh Meta, walaupun merupakan kontrak perdata biasa, harus ditafsirkan sesuai Undang-Undang Dasar Italia, termasuk dalam hal hak kebebasan berekspresi, karena Meta secara *de facto* memegang peranan sistemik dalam mendorong partisipasi politik.

Pengadilan menolak argumen Meta bahwa penutupan akun tersebut adalah sanksi terhadap fakta bahwa CasaPound adalah organisasi politik yang secara intrinsik bertentangan dengan Undang-Undang Dasar dan hukum HAM. Pengadilan menyatakan bahwa penentuan apakah CasaPound merupakan entitas yang sah di mata hukum tidak berada di tangan Meta, serta menimbang bahwa CasaPound belum dilarang oleh berbagai otoritas kompeten di Italia. Lebih jauh, pengadilan memutuskan bahwa hubungan kontraktual dalam perkara ini dihentikan secara sewenang-wenang dan konten yang dibagikan CasaPound tidak mencapai derajat keseriusan yang dapat membenarkan penutupan akun.

Senada dengan kasus di atas, Mahkamah Agung Federal Jerman memutuskan pada 2021 bahwa Meta (saat itu masih bernama Facebook) terikat oleh hak-hak fundamental di Jerman (termasuk hak kebebasan berekspresi).

Contoh

Dalam perkara yang melibatkan penghapusan postingan dan penutupan akun sementara oleh Facebook akibat dugaan 'ujaran kebencian', Mahkamah Agung Federal Jerman memutus bahwa perusahaan tersebut harus menyeimbangkan hak-hak fundamental yang saling bertentangan dalam Syarat dan Ketentuannya, dalam hal ini hak pengguna atas kebebasan berekspresi dan hak Meta untuk menjalankan profesi yang dijamin Undang-Undang Dasar Jerman. Ini memberi kewajiban kepada Meta untuk memberitahu pengguna tentang penghapusan postingan selambat-lambatnya setelah tindakan tersebut dilakukan, dan memberitahukan di awal tentang pemblokiran akun, disertai alasan pengambilan tindakan tersebut, serta memberi kesempatan pengguna untuk menanggapi, dan ditindaklanjuti dengan putusan baru.

Syarat dan ketentuan Meta untuk penghapusan postingan dan pemblokiran tidak memenuhi persyaratan ini. Karenanya, perusahaan tersebut tidak berhak menghapus postingan yang diunggah pemohon atau memblokir akun mereka.

Perusahaan media sosial kerap menerima permintaan legal dari pemerintah yang mungkin tidak sejalan dengan standar HAM internasional – misalnya untuk memblokir akun tertentu, menghapus konten, atau memberi akses data pengguna. Permintaan semacam itu bisa berdasarkan undang-undang kejahatan siber, 'disinformasi', atau undang-undang lain yang tidak memenuhi standar internasional kebebasan berekspresi.

Seperti telah dijelaskan sebelumnya, tanggung jawab perusahaan untuk menghormati HAM tidak tergantung pada kesediaan pemerintah untuk memenuhi kewajiban HAM-nya. Mantan Pelapor Khusus PBB untuk Kebebasan Berekspresi (UN Special Rapporteur on Freedom of Expression) secara spesifik menyatakan bahwa tanggung jawab perusahaan

untuk menghormati kebebasan berekspresi harus setidaknya mencakup kewajiban untuk ‘sebisa mungkin memberlakukan strategi pencegahan dan penanggulangan yang menjunjung prinsip-prinsip HAM yang diakui secara internasional ketika berhadapan dengan ketentuan hukum lokal yang bertentangan’.[21] Permintaan legal harus ditafsirkan dan dijalankan sespesifik mungkin untuk meminimalkan pembatasan hak kebebasan berekspresi.[22] Pelapor Khusus lebih jauh menegaskan bahwa ketika menerima permintaan semacam itu, perusahaan harus ‘mencari klarifikasi atau modifikasi; meminta bantuan masyarakat sipil, perusahaan di bisnis yang sama, lembaga pemerintah yang relevan, institusi internasional dan regional, serta pemangku kepentingan lainnya, dan menjajaki semua opsi hukum untuk melakukan penolakan’.[23] Terakhir, perusahaan harus transparan tentang permintaan yang datang dari pemerintah, memberikan detail tipe konten yang dimuat dalam permintaan tersebut (misalnya pencemaran nama baik, ‘ujaran kebencian’, konten terkait terorisme) serta tindakan yang diambil perusahaan (misalnya penghapusan sebagian atau keseluruhan, penghapusan bersifat spesifik negara atau penghapusan global, penutupan akun sementara, penghapusan yang dilakukan berdasarkan syarat dan ketentuan).[24]

Praktik moderasi konten

Terminologi kunci

Ada baiknya kita bahas secara singkat perbedaan moderasi konten dan kurasi konten. Sistem dan proses moderasi konten dan kurasi konten boleh jadi terkait erat, tetapi memunculkan isu yang berbeda dan kerap diperlakukan berbeda oleh regulator.

Dalam buku panduan ini, kami menggunakan definisi berikut:

- **Moderasi konten** mencakup rangkaian tindakan dan perangkat yang digunakan platform media sosial untuk menangani konten ilegal dan menegakkan standar komunitas atas konten yang dibuat pengguna layanan mereka. Biasanya hal ini melibatkan penandaan (*flagging*) oleh pengguna, penanda tepercaya (*trusted flagger*), atau 'filter'; penghapusan, pemberian label, penurunan peringkat (*down ranking*) atau demonetisasi konten; dan dimatikannya fitur tertentu.

- **Kurasi konten** adalah cara platform media sosial menggunakan sistem otomatis – kerap disebut sistem rekomendasi – untuk menentukan peringkat, mempromosikan, atau mendemosikan konten di umpan berita (*news feed*), yang biasanya berdasarkan profil pengguna. Konten dapat pula dipromosikan secara berbayar oleh platform. Platform dapat pula mengkurasi konten dengan menggunakan halaman antara (*interstitial*) – peringatan yang muncul sebelum konten ditampilkan – untuk memperingatkan pengguna tentang konten sensitif, atau dengan memasang label tertentu, misalnya untuk menegaskan apakah konten tersebut berasal dari sumber yang dapat dipercaya.

Sederhananya, moderasi konten adalah persoalan memastikan konten yang dipublikasikan tidak melanggar aturan mana pun. Kurasi konten berfokus pada cara memprioritaskan dan menyajikan konten kepada pengguna, misalnya konten apa yang tampil di awal umpan berita pengguna tersebut. Kedua proses ini boleh jadi beririsan. Misalnya, penurunan peringkat konten boleh jadi merupakan suatu langkah moderasi konten, tetapi juga merupakan bagian tak terpisahkan dari kurasi konten.

Syarat dan ketentuan serta standar komunitas

Berbagi informasi atau opini di platform media sosial tidak sama sekali bebas kendali. Ketika bergabung ke Facebook, TikTok, Twitter, atau YouTube, pengguna menerima bahwa mereka harus mematuhi syarat dan ketentuan perusahaan, yang mengatur hubungan kontraktual antara perusahaan media sosial dan pengguna. Hubungan kontraktual ini menentukan parameter seperti akses dan penggunaan produk, aplikasi, dan layanan yang disediakan.

Syarat dan ketentuan ini biasanya juga mencakup standar komunitas, yang kadang disebut juga sebagai pedoman komunitas (*community guidelines*) atau kebijakan (*policies*) (sebagai contoh, lihat Standar Komunitas Facebook, Aturan dan Kebijakan Twitter, atau Pedoman Komunitas YouTube dan TikTok). Standar komunitas ini biasanya menetapkan tipe konten yang diizinkan dan dilarang perusahaan untuk ditampilkan di platformnya. Selain itu, aturan ini juga dapat memblokir kelompok atau individu yang dianggap berbahaya, atau melarang perilaku daring tertentu (misalnya menirukan orang lain atau spam).

Pengguna media sosial yang melenceng dari standar ini akan mendapati kontennya dihapus atau turun peringkat, atau akun mereka ditutup.

Terkait ‘ujaran kebencian’ dan ‘disinformasi’ yang menjadi fokus buku panduan ini, sebagian besar standar perusahaan media sosial menangani konten semacam itu dengan langkah-langkah tertentu, walaupun dengan tingkat presisi bervariasi dan dengan pemberian label yang berbeda-beda.

Misalnya, untuk menangani informasi ‘palsu’ dalam arti luas, pendekatan yang diambil sangat berbeda dari satu platform ke platform lainnya.

- Aturan TikTok mengenai ‘integritas dan otentisitas’ menyatakan platform tersebut akan menghapus konten atau akun yang memuat informasi menyesatkan yang menimbulkan bahaya signifikan. Lebih lanjut, TikTok mendefinisikan ‘misinformasi berbahaya’ yang harus dihapus sebagai konten tidak akurat atau salah yang

menimbulkan bahaya signifikan terhadap individu, komunitas, atau publik secara luas, terlepas dari niat. Bahaya signifikan dapat mencakup 'melemahkan kepercayaan publik terhadap institusi dan proses-proses kenegaraan seperti pemerintah, pemilihan umum, dan lembaga-lembaga ilmiah', tetapi tidak berlaku untuk 'informasi yang jelas-jelas tidak akurat, mitos, atau berbahaya terhadap kepentingan komersial dan reputasional'.

- Tindakan yang diambil Twitter – mulai dari pembatasan amplifikasi hingga penghapusan atau pemberian konteks – difokuskan pada informasi 'salah' atau 'menyesatkan' yang dapat membahayakan populasi yang terdampak krisis (misalnya dalam konflik bersenjata, darurat kesehatan publik dan bencana alam berskala besar); media menyesatkan (didefinisikan sebagai media sintetis, dimanipulasi, atau di luar konteks, yang berpotensi menipu atau membingungkan); atau konten yang ditujukan untuk memanipulasi atau mengganggu pemilihan umum atau proses kenegaraan lainnya.

- Meta menyatakan bahwa platformnya menghapus 'misinformasi jika diduga dapat secara langsung berkontribusi terhadap risiko bahaya fisik yang bersifat segera'. Selain itu Meta juga menghapus 'konten yang dapat secara langsung berkontribusi menggangu jalannya proses-proses politik, dan media hasil manipulasi tertentu yang sangat mengelabui'.

Platform-platform ini juga menerapkan kebijakan terhadap perilaku di wilayah-wilayah lain yang beririsan dengan penyebaran 'misinformasi', misalnya akun palsu, penipuan, dan perilaku tidak otentik yang terkoordinasi.

Masing-masing standar komunitas ini dapat ditelaah secara mendetail dan memunculkan serangkaian isu dan kekhawatiran yang berbeda-beda dari perspektif kebebasan berekspresi (sebagai rujukan, lihat analisis ARTICLE 19 pada 2018 atas syarat dan ketentuan serta standar komunitas Facebook, Twitter, dan YouTube). Walaupun analisis semacam itu berada di luar cakupan buku panduan ini, kami menguraikan kekhawatiran umum terkait sistem yang berlaku sekarang dalam bab berikutnya.

Kekhawatiran yang muncul akibat regulasi ujaran berdasarkan kontrak

Privatisasi regulasi ujaran (yaitu regulasi ujaran dengan menggunakan kontrak) memunculkan kekhawatiran serius bagi perlindungan kebebasan berekspresi. Kekhawatiran ini diperparah dengan kenyataan bahwa segelintir perusahaan media sosial menggenggam kekuasaan besar atas apa yang orang lihat dan bagikan secara daring dan minim akuntabilitas publik.

Menurunkan standar kebebasan berekspresi

Standar komunitas yang tercantum dalam syarat dan ketentuan biasanya memiliki standar pembatasan kebebasan berekspresi yang lebih rendah dibandingkan pembatasan yang diizinkan berdasarkan hukum HAM internasional. Misalnya, seperti disebutkan dalam konteks kebijakan 'ujaran kebencian', kebijakan TikTok tidak mengizinkan misogini. Tetapi, pelarangan misogini tanpa pandang bulu tidak memenuhi syarat Pasal 19 dan Pasal 20 ICCPR.

Selain itu, beberapa perusahaan media sosial mencantumkan dalam syarat dan ketentuannya bahwa mereka berhak menghapus konten apa pun berdasarkan pertimbangan sepihak tanpa alasan apa pun. Misalnya:

- Syarat dan ketentuan Snapchat menyatakan: "Kami dapat menutup atau membekukan sementara akses Anda ke Layanan jika Anda tidak mematuhi syarat-syarat ini, Pedoman Komunitas atau perundangan-undangan, dengan alasan apa pun di luar kendali kami, atau untuk alasan apa pun, dan tanpa peringatan di awal'.

- Syarat dan Ketentuan TikTok menyatakan: 'Kapan pun dan tanpa pemberitahuan sebelumnya, kami berhak menghapus atau menutup akses terhadap konten berdasarkan kebijakan kami atas alasan apa pun atau tanpa alasan sekalipun'.

ARTICLE 19 mengakui bahwa meskipun pada prinsipnya perusahaan media sosial bebas membatasi konten berdasarkan kebebasan berkontrak, merekai tetap harus menghormati HAM, termasuk hak kebebasan berekspresi, privasi dan proses hukum sejalan dengan

Prinsip-prinsip Panduan. Hak menghapus konten dengan ‘alasan apa pun atau tanpa alasan’ jelas tidak memenuhi tanggung jawab menghormati HAM.

Walaupun perusahaan media sosial secara legal memiliki lebih banyak kelonggaran untuk membatasi ujaran di platformnya dibandingkan dengan yang boleh dibatasi oleh negara berdasarkan hukum HAM internasional, permasalahan muncul karena aturan moderasi konten di ruang-ruang kuasi-publik ini tidak dipandu prinsip-prinsip kebutuhan dan proporsionalitas. Kenyataannya, aturan moderasi konten justru kerap didikte keinginan perusahaan untuk mendongkrak profit dan memenuhi kebutuhan industri periklanan, yang tidak mau diidentikkan dengan konten yang menimbulkan ketersinggungan, mengguncang perasaan atau mengganggu, walaupun konten semacam itu tetap dilindungi standar kebebasan berekspresi. Pada praktiknya, standar kebebasan berbicara yang rendah kerap muncul akibat perusahaan media sosial yang menyesuaikan standar komunitasnya dengan aturan hukum setempat yang berada di bawah standar internasional kebebasan berekspresi.[25]

Kurangnya transparansi dan akuntabilitas

Penerapan standar komunitas sangat minim transparansi. Hal ini berdampak negatif terhadap kemampuan untuk menuntut tanggung jawab perusahaan atas penghapusan konten yang salah, sewenang-wenang dan diskriminatif.

Kadang-kadang kurangnya transparansi ini dapat berpengaruh terhadap muatan dari suatu kebijakan. Contoh yang mencolok adalah kebijakan Organisasi dan Individu Berbahaya dari Meta. Dalam kebijakan ini, Meta melabeli individu, organisasi dan jaringan tertentu yang dianggap perusahaan tersebut ‘mendeklarasikan misi berbahaya’ atau ‘terlibat kekerasan’, dan akan menghapus ‘pujian’ atau ‘dukungan’ untuk entitas-entitas tersebut. Meski demikian, walaupun kebijakan ini bertanggung jawab atas penghapusan banyak konten di platform Meta, daftar entitas yang mendapat label berbahaya tersebut tidak dipublikasikan, sehingga kontennya mustahil dianalisis.

Problem lebih lanjut muncul karena standar komunitas tidak bisa diakses di banyak bahasa, sehingga banyak pengguna sulit memahami aturan yang berlaku atas ujaran daring mereka.[26]

Dalam hal transparansi laporan, beberapa perusahaan media sosial telah berupaya memperbaiki praktiknya dalam beberapa tahun terakhir. Namun, laporan terkini tentang penegakan standar komunitas media sosial masih terbatas dari segi detail dan kualitas sehingga sulit mendapatkan wawasan yang bermakna.

Sebagai gambaran, dalam laporan transparansinya Meta menerbitkan informasi penghapusan konten berdasarkan syarat dan ketentuan, dibagi berdasarkan bidang kebijakan. Tetapi, data agregat tentang jumlah konten yang dihapus berdasarkan kategori yang kompleks dan luas seperti 'ujaran kebencian' tidaklah terlalu informatif. Tidak ada informasi tentang bagaimana Meta menerapkan definisi perusahaan untuk 'ujaran kebencian' (dan data apa yang dimasukkan ke alat moderasi konten otomatis yang digunakan); jumlah kasus penghapusan yang diperinci berdasarkan kategori (misalnya berdasarkan karakteristik-karakteristik yang dilindungi) dan berdasarkan negara (untuk memahami apakah standar komunitas diterapkan berbeda dari satu negara ke negara lainnya); dan berapa persen penghapusan berdasarkan kebijakan 'ujaran kebencian' itu terjadi setelah pemberitahuan dari institusi pemerintah di negara-negara tersebut. Kebanyakan laporan transparansi juga lebih berfokus pada penghapusan konten, sementara tindakan-tindakan lain seperti penurunan peringkat tidak disebutkan.[27]

Akibat kurangnya transparansi, umumnya sulit mengetahui apakah standar komunitas benar-benar diterapkan secara wajar dan konsisten, atau justru sewenang-wenang dan diskriminatif, kecuali jika ada liputan pers mengenai kasus spesifik atau kampanye publik oleh individu atau kelompok yang terdampak.

Pada 2021, UNESCO menyusun rangkuman mengenai akuntabilitas dan transparansi di era digital bertajuk *Letting the Sun Shine In*. Di dalamnya tercantum sederet prinsip-prinsip transparansi tingkat tinggi yang dapat meningkatkan transparansi platform daring, dengan

fokus pada isu-isu seperti transparansi konten dan proses, pengumpulan dan penggunaan data pribadi, serta uji tuntas dan ganti rugi.

Kurangnya prosedur pengamanan dan pemulihan

Prosedur pengamanan yang berlaku dalam penghapusan konten media sosial kurang memadai. Transparansi lagi-lagi menjadi masalah. Tidak selalu jelas apakah perusahaan memberi tahu pengguna bahwa kontennya telah dihapus atau ditandai, atau akunnya telah dijatuhi sanksi, serta alasan di balik tindakan-tindakan tersebut. Bahkan jika disampaikan kepada pengguna, seringkali pemberitahuan tersebut hanya berisi referensi kebijakan yang menjadi dasar tuduhan pelanggaran, tanpa penjelasan memadai agar pengguna memahami alasan pemberlakuan pembatasan tersebut. Seperti dijelaskan dengan lebih mendetail di bawah, masalah ini diperparah oleh penggunaan alat moderasi konten otomatis.

Walaupun kebanyakan perusahaan media sosial besar memperbolehkan pengguna mengajukan banding atas penghapusan konten atau pembekuan akun sementara, hal ini hanya akan bermakna jika pengguna menerima pemberitahuan dan memahami alasan yang berujung pada sanksi pembatasan konten mereka. Dalam konteks ini, masalah timbul terutama karena perusahaan media sosial mempraktikkan '[shadowbanning]', yang dideskripsikan sebagai keadaan di mana konten pengguna disembunyikan atau mengalami penurunan visibilitas tanpa pemberitahuan dari platform.

Kekurangan lainnya yang tak kalah kritis adalah individu yang kontennya dihapus biasanya tidak mendapatkan mekanisme pemulihan legal yang memadai. Syarat dan ketentuan perusahaan seringkali tidak memberikan dasar untuk mengajukan klaim terkait pembatasan konten, dan pengguna di banyak yurisdiksi juga tidak bisa menggunakan jalur klaim non-kontraktual. Penyelesaian sengketa yang problematik dan pilihan ketentuan hukum dalam syarat dan ketentuan perusahaan media sosial dapat menjadi penghalang tambahan bagi pengguna untuk mengakses keadilan. Dalam beberapa kasus, hal ini dapat menghalangi pengguna untuk mengajukan klaim ke pengadilan setempat di negara domisili mereka atau memberlakukan hukum setempat terhadap syarat dan ketentuan

tersebut. Ini membuat kebanyakan pengguna enggan mengajukan tuntutan hukum karena ketiadaan sumber daya.

Misalnya untuk pengguna yang berbasis di Inggris – dan ketentuan serupa bisa jadi juga berlaku di yurisdiksi lainnya – Twitter mensyaratkan semua sengketa terkait syarat dan ketentuan dibawa ke pengadilan di San Francisco, sementara hukum yang berlaku adalah hukum Negara Bagian California. Syarat dan ketentuan Snapchat memungkinkan perjanjian arbitrasi dengan syarat arbitrasi tersebut dilakukan di Amerika Serikat disertai pernyataan pengabaian (*waiver*) untuk *class action* (kecuali untuk prosedur gugatan sederhana dengan opsi batal dalam batas waktu 30 hari yang luput dari perhatian kebanyakan pengguna). Ketentuan Meta lebih masuk akal, karena menyatakan bahwa sengketa konsumen berada di bawah yurisdiksi pengadilan negara domisili utama pengguna dan hukum di negara tersebut berlaku. Syarat dan ketentuan TikTok juga menyatakan bahwa sengketa terkait syarat dan ketentuan berada di bawah yurisdiksi pengadilan di mana pengguna berada serta pengadilan Republik Irlandia dan pengadilan Inggris dan Wales.

Praktik-praktik di luar jalur hukum

Terakhir, otoritas publik, terutama institusi penegakan hukum, kerap meminta kerja sama platform media sosial dalam kerangka pemberantasan aktivitas kriminal (misalnya penyebaran materi kekerasan seksual terhadap anak) atau bahaya sosial lainnya (misalnya 'ekstremisme daring') tanpa melalui jalur hukum. Terutama karena pihak-pihak berwenang ini tidak selalu memiliki kewenangan untuk memerintahkan penghapusan konten bermasalah, mereka kadang menghubungi perusahaan media sosial secara informal dan meminta penghapusan konten atas dasar syarat dan ketentuan. Meskipun tidak memiliki kewajiban hukum untuk memenuhi permintaan tersebut, perusahaan berada di posisi sulit, terutama dalam keadaan manakala konten yang dimaksud menyerempet ilegalitas. Alhasil, perusahaan media sosial kerap menjadi perpanjangan tangan hukum tanpa memberi peluang pada pengguna untuk mengajukan banding atas legalitas pembatasan tersebut di pengadilan.[28]

Contoh

Dalam kasus musik *UK drill*, Dewan Pengawas membatalkan keputusan Meta untuk menghapus klip video musik bergenre *UK drill* dengan judul *Secrets Not Safe* oleh Chinx (OS) dari Instagram. Awalnya Meta menghapus konten tersebut mengkuti permintaan Kepolisian Metro Inggris. Kepolisian mengirimkan surat elektronik kepada Meta untuk meminta perusahaan tersebut meninjau semua konten yang memuat lagu *Secrets Not Safe* dan memberikan konteks tambahan meliputi kekerasan antargeng di London, termasuk pembunuhan, serta kekhawatiran kepolisian bahwa lagu tersebut dapat memicu kekerasan balas dendam lebih jauh. Meta menghapus konten tersebut dari akun yang ditinjau atas dasar pelanggaran kebijakan antikekerasan dan hasutan.

Temuan Dewan Pengawas di antaranya menunjukkan bahwa '[j]alur yang digunakan penegak hukum untuk mengajukan permintaan kepada Meta tidak menentu dan kabur. Institusi penegak hukum tidak dituntut memenuhi kriteria minimum untuk menjustifikasi permintaan mereka, dan karena itu interaksi yang terjadi tidak konsisten. Data yang dipublikasikan Meta mengenai permintaan pemerintah juga tidak lengkap.'

Peran kerangka regulasi

Proses moderasi konten tidak melulu berkutat dengan penegakan standar-standar syarat dan ketentuan perusahaan media sosial, tetapi juga dengan ketentuan regulasi penghapusan konten ilegal dan yang mengundang keberatan. Perusahaan menerima tekanan yang semakin besar dari pemerintah selama beberapa tahun terakhir untuk menghapus lebih banyak konten dari platform – mulai dari konten 'ujaran kebencian' dan 'ekstremisme' hingga 'disinformasi'.

Fokus tradisional dalam pengaturan kewajiban perantara

Secara tradisional kerangka regulasi berfokus pada pengaturan kewajiban entitas yang disebut sebagai perantara internet – satu istilah bermakna luas yang mencakup perusahaan *hosting* situs web, penyedia layanan internet, mesin pencari, dan platform

media sosial.[29] Undang-undang kewajiban perantara meregulasi sejauh mana perantara internet dapat dimintai pertanggungjawaban hukum atas konten yang disebarkan atau dibuat oleh penggunanya (konten pihak ketiga) dan apakah mereka juga memiliki kewajiban, misalnya, membayar ganti rugi berupa uang kepada pihak yang dirugikan.

Umumnya, rezim tanggung jawab terentang dalam spektrum antara pertanggungjawaban mutlak atau pertanggungjawaban tanpa kesalahan *(strict liability)* di ujung yang satu dan imunitas di ujung yang lain. Dalam rezim pertanggungjawaban mutlak, perantara internet dapat diadili di pengadilan akibat kesalahan perilaku pengguna, tanpa perantara harus bersalah atau mengetahui pelanggaran tersebut. Praktis perantara harus memantau konten dan bertindak jika relevan dengan kepatuhan hukum. Model ini sudah diterapkan antara lain di Thailand.

Contoh

Chiranuch Premchaiporn, editor Prachatai, yaitu situs berita daring di Thailand, diadili dan divonis berdasarkan ketentuan Undang-Undang Kejahatan Komputer Thailand 2007 karena tidak segera menghapus komentar anonim yang dianggap menghina raja Thailand.[30] Selain memberi hukuman terhadap 'data palsu' yang merugikan pihak ketiga, mengakibatkan kepanikan publik, atau merongrong keamanan negara', Undang-Undang Kejahatan Komputer juga berlaku terhadap 'pemberi layanan yang sengaja mendukung ' data yang salah tersebut. Hukum pidana Thailand (*lèse-majesté*) menyatakan bahwa siapa pun yang 'mencemarkan nama baik, menghina, atau mengancam raja, ratu, pewaris takhta, atau wali kerajaan' akan dihukum penjara.

Chiranuch Premchaiporn divonis satu tahun penjara dan denda 30.000 baht, yang kemudian dikurangi menjadi delapan bulan penjara dengan penangguhan dan denda 20.000 baht setelah banding.[31]

Imunitas total terhadap kewajiban atas konten buatan pengguna – yang berarti segala klaim yang diajukan kepada perantara mengenai konten buatan pengguna akan ditolak –

bukan suatu hal yang lazim. Contoh paling menonjol yang terutama mendorong pendekatan ini adalah Section 230 dari Undang-Undang 1996 tentang Kesusilaan dalam Komunikasi, yang berlaku di Amerika Serikat. Section 230 memberi imunitas legal kepada platform daring atas konten yang dipublikasikan oleh pihak ketiga (walaupun imunitas tersebut tidak berlaku dalam hal pelanggaran hukum pidana federal, undang-undang hak cipta, atau undang-undang privasi dalam komunikasi elektronik).

Banyak sistem hukum menempatkan diri di antara pertanggungjawaban mutlak dan imunitas total dengan memberlakukan sistem imunitas bersyarat. Perantara internet bebas dari kewajiban sepanjang menghapus konten begitu mengetahui ilegalitas konten tersebut. Sistem kewajiban berdasarkan pengetahuan ini biasanya berjalan melalui prosedur yang dikenal dengan 'pemberitahuan dan penghapusan' (*notice and take down*).

Cara kerja prosedur 'pemberitahuan dan penghapusan' berbeda dalam setiap yurisdiksi. Biasanya perantara internet dianggap memperoleh pengetahuan tentang ilegalnya suatu konten ketika mendapat notifikasi dari pihak ketiga. Jika tidak menghapus konten ilegal tersebut setelah mendapat notifikasi, perantara internet dapat dituntut pertanggungjawabannya atas kerugian yang mungkin terjadi. Misalnya, Undang-Undang Layanan Digital Uni Eropa yang baru-baru ini diberlakukan memberikan kewajiban atas konten yang menjadi subjek pelaporan berbasis bukti dari pengguna.

Dari perspektif kebebasan berekspresi, secara luas diakui – termasuk oleh mandat khusus kebebasan berekspresi – bahwa imunitas yang luas terhadap kewajiban perantara internet adalah salah satu cara paling efektif untuk melindungi kebebasan berbicara daring. Jika perusahaan dapat dimintai tanggung jawab atas konten yang dipubikasikan pengguna, praktis platform harus memonitor semua konten buatan pengguna – suatu invasi masif terhadap hak-hak privasi. Rezim seperti ini juga memberi insentif kuat bagi perusahaan untuk menyensor penggunanya secara berlebihan dan menghapus materi yang sah menurut hukum, demi menghindari risiko melanggar hukum.

Pengalaman menunjukkan bahwa bahkan rezim imunitas bersyarat yang bekerja lewat prosedur 'pemberitahuan dan penghapusan' memberi insentif untuk menghapus konten

dengan segera atas dasar tuduhan pihak swasta atau badan publik, tanpa penentuan pengadilan mengenai ilegalitas konten tersebut. Selain itu, orang yang mempublikasikan konten bermasalah tersebut biasanya tidak diberi peluang untuk mempertimbangkan pengaduan yang dilayangkan atas konten tersebut.

Pada 2011, mantan Pelapor Khusus PBB untuk Kebebasan Berekspresi Frank La Rue menyatakan bahwa sensor tidak boleh didelegasikan ke entitas swasta, dan negara tidak boleh menggunakan atau memaksa perantara untuk melakukan sensor atas nama negara.[32] Dia juga mencatat bahwa rezim 'pemberitahuan dan penghapusan' rawan disalahgunakan oleh negara maupun aktor swasta, dan kurangnya transparansi dalam pengambilan keputusan oleh perantara kerap mengaburkan praktik-praktik diskriminatif atau tekanan politik yang memengaruhi keputusan perusahaan.[33]

Tren ke arah meningkatnya regulasi platform daring

Beberapa tahun terakhir, perusahaan media sosial semakin banyak dikritik karena meraup keuntungan melalui algoritma yang mendorong interaksi adiktif dengan konten 'ekstremis' dan konten 'berbahaya' lainnya. Ini menimbulkan pertanyaan tentang perlunya regulasi yang lebih ketat untuk menjinakkan kekuatan perusahaan-perusahaan media sosial terbesar, menanggulangi konten ilegal dan berbahaya, serta meningkatkan akuntabilitas demokratis kepada publik yang lebih luas atas keputusan yang diambil perusahaan. Beberapa pemerintahan menanggapi dengan proposal yang membebani platform dengan '*duty of care*', kewajiban bertindak hati-hati, atas penggunanya untuk mencegah 'bahaya' akibat ujaran pengguna lain di platform tersebut.[34]

Banyak proposal terkini tentang kerangka regulasi nyatanya bermasalah dari perspektif kebebasan berekspresi. Meskipun kesannya disusun untuk meningkatkan akuntabilitas perusahaan media sosial, proposal regulasi ini justru kerap berfokus pada regulasi 'konten'. Artinya proposal-proposal ini seringkali mengatur ujaran pengguna dan bukan produk, sistem, dan proses yang diterapkan perusahaan media sosial. Negara pada praktiknya menuntut perusahaan menjadi polisi komunikasi antarmanusia dan

menentukan ujaran mana yang 'ilegal' dan 'berbahaya',[35] padahal tanggung jawab penentuan tersebut terletak pada otoritas peradilan independen.

Selain kekhawatiran terkait keabsahan mensubkontrakkan (*outsourcing*) keputusan mengenai legalitas ujaran pengguna kepada aktor swasta, dalam banyak kasus, penilaian ini kelewat kompleks, tergantung pada konteks, dan karenanya harus dilakukan oleh individu terlatih. Kenyataannya, perusahaan media sosial menggunakan sistem moderasi algoritmik, seperti pencocokan tagar secara otomatis (*automated hash-matching*) dan alat berbasis prediksi pemelajaran mesin (*machine learning*), untuk melakukan moderasi konten. Karena saat ini teknologi-teknologi tersebut belum cukup canggih (dan mungkin tidak akan pernah cukup canggih) untuk membedakan konten legal dari konten ilegal dengan cara yang handal, seringkali konten legal diidentifikasi sebagai konten ilegal yang berujung pada penghapusan banyak konten legal.

Regulasi harus menggunakan pendekatan HAM dan merangkul pasar digital

Daripada memberikan mandat kepada platform untuk membatasi jenis ujaran yang tidak diinginkan, proposal regulasi seharusnya memastikan bahwa HAM menjadi inti regulasi platform. Prinsip-prinsip legalitas, legitimasi, kebutuhan, dan proporsionalitas yang diatur dalam Pasal 19 (3) ICCPR harus diterapkan sepenuhnya. Seperti kerangka mana pun yang memberlakukan pembatasan kebebasan berekspresi, regulasi yang mengatur perusahaan media sosial harus berakar pada bukti-bukti kuat dan memprioritaskan minimalisasi sensor dan pembatasan untuk menanggulangi bahaya daring.

Bagaimana ini diwujudkan dalam praktik? Daripada meminta platform lebih mengontrol ujaran pengguna dengan menyaring dan mengevaluasi semua konten yang dibuat, regulator harus berfokus pada metode-metode yang tidak terlalu mengganggu dan spesifik dirancang untuk menanggulangi efek negatif model bisnis perusahaan media sosial, termasuk sistem rekomendasinya. Misalnya, solusi regulasi harus mewajibkan perusahaan lebih transparan kapada regulator, peneliti, dan pengguna tentang cara kerja sistem rekomendasinya; menetapkan batasan yang jelas tentang jumlah data pengguna yang boleh dikumpulkan; dan mengharuskan adanya uji tuntas kinerja HAM. Regulator

juga harus berfokus pada transparansi keputusan moderasi konten dan perbaikan sistem penyelesaian sengketa yang muncul dari keputusan tersebut.

Selama bertahun-tahun ARTICLE 19 juga telah mengadvokasikan bahwa solusi regulasi harus juga menangani posisi dominan platform daring, menggunakan alat-alat regulasi yang meningkatkan kompetisi pasar dan pilihan pengguna atas konten yang dapat mereka lihat secara daring.

Sebagian solusi regulasi ini telah diadopsi Uni Eropa pada 2022 dalam Undang-Undang Layanan Digital dan Undang-Undang Pasar Digital Uni Eropa. Meskipun kerangka regulasi ini seharusnya bisa lebih ambisius dalam hal perlindungan HAM daring (misalnya dengan menetapkan hak eksplisit bagi pengguna atas enkripsi dan anonimitas), produk hukum ini berfokus pada upaya menyeimbangkan kembali pasar digital dan meregulasi moderasi konten dan sistem kurasi yang digunakan perusahaan media sosial.

Di samping itu, ketentuan moderasi konten dalam Undang-Undang Layanan Digital Uni Eropa menjawab kekhawatiran yang timbul akibat 'regulasi ujaran berbasis kontrak'. Ketentuan tersebut di antaranya:

- pengguna harus menerima notifikasi dan pernyataan alasan jika konten mengalami penurunan visibilitas (termasuk penghapusan dan demosi);
- pembentukan mekanisme pengaduan internal yang memungkinkan pengguna mengajukan pengaduan atas keputusan moderasi konten kepada perusahaan secara elektronik dan bebas biaya;
- negara anggota membentuk mekanisme penyelesaian sengketa luar pengadilan untuk keputusan moderasi konten, sehingga pengguna berhak memilih lembaga penyelesaian sengketa luar pengadilan untuk sengketa keputusan moderasi konten, termasuk pengaduan yang tidak bisa diselesaikan melalui sistem penanganan pengaduan internal; dan

- laporan transparansi yang relatif mendetail mengenai aktivitas moderasi konten di perusahaan media sosial tersebut, misalnya jumlah permintaan dari pihak berwenang negara anggota untuk menindak konten ilegal, atau moderasi konten yang dilakukan atas inisiatif perusahaan, termasuk jumlah dan jenis tindakan yang diambil yang berpengaruh terhadap ketersediaan, visibilitas, serta aksesibilitas konten yang dipublikasikan.

Sampai titik tertentu, Undang-Undang Layanan Digital dapat menjadi model proposal regulasi di seluruh dunia. Contoh yang paling jelas adalah rancangan Undang-Undang Kebebasan, Tanggung Jawab dan Transparansi yang sedang dibahas di Brasil saat buku panduan ini ditulis. Sebelumnya, Undang-Undang Penegakan Jaringan Jerman (NetzDG), yang mewajibkan platform menghapus konten yang 'jelas terbukti melanggar hukum' dalam 24 jam setelah notifikasi, dengan ancaman denda berat, telah menginspirasi beberapa negara, termasuk Kenya, Malaysia, dan India, untuk memperketat undang-undang kewajiban perantara.[36]

Pembahasan tentang penyusunan regulasi platform dengan pendekatan yang menghormati HAM serta bagaimana menangani peluang dan risiko HAM yang muncul dari teknologi digital secara lebih luas juga berlangsung di forum-forum internasional. Misalnya, ada upaya yang sedang dilakukan di PBB untuk mencapai kesepakatan mengenai Global Digital Compact yang di antaranya mengatur isu-isu seperti 'memajukan internet yang dapat dipercaya dengan memasukkan kriteria akuntabilitas untuk konten diskriminatif dan menyesatkan'. UNESCO telah merumuskan panduan global untuk meregulasi platform internet, dengan tujuan 'membekali proses-proses regulasi digital platform yang tengah dikembangkan atau dikaji kembali, dengan cara yang konsisten dengan standar internasional HAM'.

Media berita dan moderasi konten

Perlindungan pluralisme dan keragaman media dalam ekosistem digital telah menjadi kekhawatiran bagi regulator. Media berita bergantung pada platform untuk mengakses audiens, mendapatkan pemasukan iklan, dan pendanaan, sehingga platform memiliki

pengaruh besar terhadap pilihan-pilihan editorial, organisasional, dan bisnis media. Lebih spesifik, muncul kekhawatiran akan pengaruh sistem moderasi konten terhadap visibilitas, monetisasi, jangkauan konten dan akun, serta dampaknya terhadap independensi editorial jurnalis dan keberlangsungan finansial jurnalisme.

Kekurangan sistem moderasi konten otomatis –yang dibahas secara lebih mendetail pada bab berikutnya dalam buku panduan ini– membuat keadaan semakin rumit bagi pelaku media. Karena tidak mampu mempertimbangkan konteks secara semestinya, alat-alat moderasi konten otomatis juga rawan melakukan penghapusan liputan isu kepentingan publik, misalnya liputan tentang kelompok ekstremis dan pelanggaran HAM.

Contoh

Sebuah media di Sarajevo dihalangi saat menerbitkan konten di Facebook berupa liputan putusan Mahkamah Pidana Internasional untuk Yugoslavia (ICTY) dalam perkara Ratko Mladić yang dipidana atas sejumlah kejahatan internasional, di antaranya terkait pembantaian di Srebrenica. Hal ini terjadi karena Facebook telah secara keliru melabeli ICTY sebagai organisasi kriminal. Sistem moderasi konten otomatis kemudian menyimpulkan bahwa artikel mengenai persidangan Mladić berupaya mendukung 'organisasi kriminal'.[37]

Beberapa proposal kerangka regulasi mencantumkan ketentuan yang dapat membentengi 'konten editorial' dari aturan moderasi konten (sementara proposal regulasi lainnya, misalnya di Australia atau Kanada, lebih berfokus pada upaya memberi tekanan kepada perusahaan media sosial untuk membayar berita yang mereka gunakan). Misalnya RUU Keamanan Daring Inggris – yang masih dinegosiasikan saat buku panduan ini disusun – membebaskan konten media berita dari aturan moderasi konten, sebagai pengecualian dari lingkup kewajiban yang diberlakukan RUU tersebut terhadap perusahaan media sosial. Pengecualian media lainnya juga dibahas dalam negosiasi Undang-Undang Layanan Digital Uni Eropa tetapi pada akhirnya ditolak. Undang-Undang Kebebasan Media Eropa, yang juga tengah dinegosiasikan pada saat penyusunan buku panduan ini, mengusulkan

satu proses yang memungkinkan penyedia layanan media meminta perlakuan khusus dari platform manakala ada sangkut pautnya dengan cara konten mereka dimoderasi.[38] Lebih spesifik lagi, proposal-proposal terkini mensyaratkan bahwa sebelum konten ditutup sementara, perusahaan media sosial harus mengkomunikasikan alasan keputusan tersebut kepada pihak media dan menjamin bahwa semua pengaduan media 'akan diproses dan diputuskan sebagai prioritas tanpa penundaan tak beralasan'.[39]

Yang juga jadi bahan perdebatan adalah apakah tepat menyusun aturan moderasi konten untuk media berita. Ada yang berargumen bahwa syarat dan ketentuan perusahaan media sosial tidak boleh diletakkan di atas standar editorial media, dan pelaku media pada umumnya sudah akuntabel berdasarkan hukum, kode etik dan keanggotaan dalam organisasi profesi.

Meski demikian, sejumlah organisasi advokasi kebebasan berekspresi dan hak-hak digital, termasuk ARTICLE 19, menegaskan bahwa meskipun tak bisa dibantah bahwa pelaku media menghadapi tantangan dalam menyikapi aturan moderasi konten –dan faktanya relasi kuasa asimetris antara perusahaan media sosial dan pelaku media membutuhkan solusi regulasi– dari kacamata kebebasan berekspresi, memberi perlakuan khusus kepada pelaku media tertentu dalam moderasi konten justru akan bermasalah. Pada prinsipnya, ujaran sebagian aktor publik tidak boleh dihargai lebih dari yang lain atas dasar siapa aktor tersebut dan bukan apa yang dia katakan.

Apalagi jika mengingat bahwa dalam standar kebebasan berekspresi untuk 'ujaran kebencian', sebagaimana dijelaskan sebelumnya, posisi dan pengaruh pembicara merupakan faktor kunci yang harus dipertimbangkan pengadilan ketika menilai apakah suatu ujaran telah mencapai tingkat hasutan kebencian yang dilarang Pasal 20 (2) ICCPR. Meminta perusahaan media sosial untuk menghapus konten yang diposting pengguna biasa tetapi melindungi konten yang sama jika dipublikasikan pelaku media yang mendapat pengecualian tentunya tidak sesuai dengan standar-standar ini.

Keistimewaan untuk media juga dapat memperkokoh kekuasaan media petahana dengan mengorbankan jurnalis warga, bloger kecil, atau aktivis yang tidak memenuhi kriteria

mendapatkan pengecualian regulasi, termasuk mereka yang melakukan aktivitas jurnalisme untuk tujuan nirlaba. Karena itu, organisasi advokasi kebebasan berekspresi umumnya menolak hak istimewa media dalam hal moderasi konten.

Proses-proses moderasi konten

Perusahaan media sosial menggunakan serangkaian pendekatan untuk moderasi konten, dan menggunakan berbagai alat untuk menerapkan kebijakan konten dan membatasi atau menghapus konten dan akun ilegal atau yang mengundang keberatan. Mengingat volume konten yang dihasilkan pengguna – angka-angka terbaru misalnya menunjukkan bahwa setiap menit setidaknya 350.000 twit diposting ke Twitter, 500 jam video diunggah ke YouTube, dan lebih dari 510.000 komentar dan 136.000 foto diposting di Facebook – mayoritas perusahaan media sosial besar memilih untuk banyak mengandalkan alat otomatis demi mengurangi kebutuhan moderasi oleh manusia yang memakan banyak waktu dan karena itu hanya dilakukan pada peninjauan konten dalam kondisi spesifik.

Kelemahan otomatisasi

Moderasi konten otomatis melibatkan pendeteksian, pemfilteran, dan penggunaan alat moderasi otomatis untuk menandai, memisahkan, dan menghapus konten atau akun tertentu. Ada sejumlah alat otomatis, banyak di antaranya dimotori kecerdasan buatan dan pemelajaran mesin (*machine learning*) yang bisa diaktifkan selama proses moderasi konten. Alat-alat ini dapat dikerahkan pada berbagai kategori konten dan format media dalam setiap tahapan siklus hidup konten, untuk mengidentifkasi, menyortir, dan menghapus. Alat dan metode otomatis yang paling banyak digunakan di antaranya ialah teknologi tagar digital,[40] pengenalan gambar,[41] atau pemrosesan bahasa alami *(natural language processing/ NLP)*.[42]

Platform kerap menerapkan apa yang biasa disebut moderasi konten hibrida, umumnya menggabungkan penggunaan alat otomatis untuk menandai dan memprioritaskan kasus-kasus spesifik bagi peninjau manusia (*human reviewer*) yang membuat keputusan akhir.[43]

Platform hanya bisa melakukan moderasi konten dalam skala besar dengan sedikit banyak mengandalkan moderasi konten otomatis, karena moderasi oleh manusia tidak akan sanggup mengolah banyaknya informasi yang dihasilkan pengguna. Di saat yang sama, alat-alat ini dapat menimbulkan risiko serius dari perspektif kebebasan berekspresi, terutama ketika diterapkan terhadap kategori ujaran yang kompleks seperti konten 'ujaran kebencian' atau 'teroris'. Kategori semacam ini membutuhkan pemahaman kontekstual dan nuansa tingkat tinggi yang tidak dimiliki alat-alat tersebut.

Kurangnya akurasi dan keandalan

Akurasi alat moderasi konten dalam mendeteksi dan menghapus sangat tergantung pada tipe konten yang harus ditangani. Misalnya alat otomatis dapat secara efektif mengidentifikasi konten yang digolongkan sebagai materi kekerasan seksual terhadap anak. Dalam hal ini, ada konsensus internasional yang jelas bahwa konten seperti itu ilegal, ada parameter yang jelas tentang apa yang harus ditandai, dan model-model telah dilatih dengan menggunakan cukup data untuk menghasilkan tingkat akurasi yang tinggi.[44]

Tetapi tidak demikian halnya dengan kategori seperti 'ujaran kebencian' atau 'konten teroris'. Agar pengklasifikasian NLP dapat dilatih untuk beroperasi secara akurat, perlu diberikan parameter dan definisi ujaran yang jelas.[45] Sudah banyak diketahui bahwa definisi konten ekstremis atau teroris –belum lagi apa yang bisa dikategorikan sebagai pujian atau dukungan untuk organisasi teroris– tidak tersedia atau kabur. Umumnya alat NLP kerap tidak dapat memahami unsur nuansa dan konteks dalam suatu ujaran atau mengidentifikasi satir atau konten yang dipublikasikan sebagai liputan.[46] Sangat penting bagi jurnalis dan organisasi HAM untuk meningkatkan kesadaran tentang kekejaman teroris. Menyaring dan menghapus 'konten teroris' tanpa mengapresiasi konteksnya berisiko menginterupsi kerja jurnalistik resmi dan pendokumentasian pelanggaran HAM. Banyak kejadian di mana moderasi otomatis berujung pada penghapusan massal konten yang mengandung kepentingan publik.

Contoh

Syrian Archive, sebuah proyek pengamanan bukti pelanggaran HAM dan kejahatan lain yang terjadi selama konflik di Suriah untuk tujuan advokasi, keadilan, dan akuntabilitas, mendapati video-video mereka yang mendokumentasikan kejahatan perang dihapus dari YouTube. Ini dapat mengakibatkan banyak dokumentasi yang mungkin merupakan bukti penting kejahatan perang hilang, kadang untuk selamanya.

Contoh

Pada Mei 2021, Serikat Kebebasan Sipil Amerika (ACLU) melaporkan belasan aktivis dan jurnalis yang meliput pelanggaran HAM dan serangan udara terhadap warga sipil mengeluh bahwa Facebook telah menutup akun mereka berdasarkan kebijakan yang melarang 'pujian, dukungan, atau representasi' 'kelompok teroris', dan menuding isu 'sistem deteksi otomatis yang tidak presisi' dan ketiadaan definisi yang disepakati tentang istilah-istilah seperti 'terrorisme', 'ekstremisme dengan kekerasan', atau 'ekstremisme', apalagi 'dukungan' bagi hal-hal tersebut.

Akurasi sistem penggolongan NLP akan turun terutama ketika diberlakukan terhadap bahasa-bahasa dan konteks berbeda. Alat-alat otomatis memiliki kemampuan terbatas untuk menganalisis dan memahami varian bahasa dan perilaku yang mungkin muncul akibat faktor demografi dan kedaerahan.[47] Di samping itu, kebanyakan alat NLP memiliki akurasi rendah ketika menganalisis teks yang bukan dalam bahasa Inggris karena kurangnya sumber daya dalam bahasa lainnya.[48]

Amplifikasi bias

Di luar isu melemahnya akurasi ketika menganalisis bahasa selain bahasa Inggris, adanya bias dalam alat otomatis menimbulkan risiko lebih memarginalkan dan menyensor kelompok yang sudah menghadapi prasangka dan diskriminasi tidak proporsional di ranah

daring maupun luring. Risiko ini berakar dari berbagai bias manusia yang diumpankan ke dalam data untuk melatih alat-alat otomatis dan karenanya dapat teramplifikasi melalui penggunaannya.[49] Jika dataset yang digunakan tidak mencakup representasi yang memadai, timbul risiko kecerdasan buatan *(artificial intelligence/ AI)* sistem tersebut mempelajari dan melanggengkan bias-bias yang mendasari data tersebut.[50]

Contoh

ACLU melaporkan bahwa selama bertahun-tahun, Meta memperlakukan ujaran perempuan dan orang-orang dengan warna kulit berwarna dengan cara berbeda dibandingkan ujaran laki-laki dan orang kulit putih – termasuk ketika menggambarkan kekerasan seksual dan rasisme yang dialami. Pada 2017, ketika perempuan kulit berwarna dan orang kulit putih memposting konten yang persis sama, hanya akun perempuan kulit berwarna yang dibekukan sementara.

Kurangnya transparansi dan akuntabilitas

Perlu ada transparansi dan akuntabilitas yang lebih besar dalam penggunaan alat otomatis. Transparansi atas cara pengumpulan data, tingkat akurasi alat moderasi konten otomatis, dan jumlah konten yang dihapuskan dengan tepat ataupun keliru masih belum memadai.[51] Ini menimbulkan kekhawatiran tentang hak-hak kebebasan berekspresi individu-individu yang kontennya keliru ditandai atau yang akunnya dihapus secara tidak tepat. Kekhawatiran ini diperparah apabila terdapat filter unggahan sehingga konten yang ditandai akan hilang dari platform bahkan sebelum diunggah, sehingga sulit mengetahui apakah penghapusan tersebut keliru atau tidak.[52]

Laporan pengguna dan 'penanda tepercaya (trusted flaggers)'

Biasanya perusahaan media sosial mengizinkan penggunanya melaporkan konten yang mereka anggap ilegal, melanggar standar komunitas, atau jelas-jelas berbahaya. Seperti dijelaskan sebelumnya, atas dasar kewajiban perantara, beberapa sistem

regulasi mematok konsekuensi legal terhadap laporan tersebut melalui prosedur ‘pemberitahuan dan penghapusan’ yang bermasalah dari sudut pandang kebebasan berekspresi dan meningkatkan risiko penandaan yang menimbulkan kejengkelan atau disalahgunakan.

Beberapa perusahaan media sosial juga mengandalkan sistem ‘penanda tepercaya’ di mana laporan yang diajukan penanda tepercaya akan disegerakan untuk peninjauan. Biasanya penanda tepercaya ini adalah individu atau entitas dengan kepakaran spesifik untuk mengidentifikasi dan menandai konten ilegal. Mereka juga bisa berasal dari kalangan masyarakat sipil.

Saat ini informasi tentang sistem penanda tepercaya dari perusahaan media sosial umumnya masih kurang memadai. Informasi ini termasuk proses seleksi penanda tepercaya dan sejauh mana konten yang mereka tandai akan ditinjau secara layak atau otomatis dihapus. Walaupun turut memperbaiki kualitas penandaan konten, sistem penanda tepercaya ini tidak setara dengan penilaian konten bermasalah secara imparsial dan independen. Penanda terpercaya kerap dipilih atas dasar keahlian mereka tentang dampak konten tertentu, baik hak cipta, konten terkait terorisme, atau ‘ujaran kebencian’, di samping kedekatan mereka dengan korban ujaran-ujaran tersebut, bukan berdasarkan kepakaran di bidang kebebasan berekspresi. Karena itu, para penanda tepercaya ini belum tentu tepat untuk melakukan penilaian imparsial mengenai apakah pembatasan suatu konten konsisten dengan hukum HAM internasional.[53]

Sistem ini juga sudah diatur dalam Undang-Undang Layanan Digital Uni Eropa, yang memberikan hak istimewa kepada penanda tepercaya dan menyatakan bahwa begitu suatu konten ditandai penanda tepercaya, platform harus menghapus konten ilegal tersebut ‘secepatnya’. Masyarakat sipil telah mengkritik undang-undang ini karena mengizinkan pemerintah atau penegak hukum menyandang status penanda tepercaya, yang dapat membuka peluang terjadinya penyalahgunaan.

Perlunya pemahaman mendalam mengenai konteks

Kelemahan sistem moderasi konten otomatis membuat perusahaan media sosial tidak bisa tidak harus mempekerjakan peninjau manusia dalam jumlah yang cukup, dan menginvestasikan sumber daya untuk lebih memahami konteks dari bentuk-bentuk ekspresi tertentu. Hal ini penting untuk meninjau dan menilai secara memadai kategori ujaran yang tidak cocok dievaluasi menggunakan moderasi otomatis, juga agar keputusan moderasi konten yang mengalami banding dapat ditindaklanjuti dengan lebih baik. Lebih penting lagi, peninjau manusia harus diperiksa latar belakangnya dan dilatih dengan benar untuk mengurangi risiko bias dalam pengambilan keputusan. Mereka juga harus merupakan penutur asli bahasa yang mereka tangani dan memahami konteks lokal yang berlaku. Saat ini jumlah moderator konten manusia yang bekerja untuk masing-masing perusahaan media sosial, bagaimana mereka dilatih, dan lokasi mereka masih minim transparansi. Kurangnya transparansi ini mengisyaratkan adanya isu yang lebih besar tentang kurangnya sumber daya yang dialokasikan perusahaan-perusahaan media sosial terbesar untuk memahami konten di banyak negara tempat mereka beroperasi, yang tidak mereka lihat sebagai kepentingan strategis.

Kesimpulan

Moderasi konten adalah bidang yang cepat berubah. Ekosistem regulasi terus berganti, pemain baru bermunculan, perusahaan media sosial secara rutin memperbarui kebijakannya, dan organisasi internasional terus meluncurkan inisiatif demi menemukan solusi global dan regional untuk konten daring 'berbahaya'. Moderasi konten juga perlu terus merespons tantangan-tantangan nyata seperti pandemi COVID-19, konflik bersenjata internasional dan non-internasional, serta produk dan teknologi baru seperti *metaverse*.

Meskipun perubahaan-perubahan terjadi sangat cepat, aspek-aspek tertentu dari moderasi konten tidak akan segera berubah. HAM harus menjadi inti setiap upaya organisasi internasional dan negara untuk meregulasi perusahaan media sosial. Perusahaan media sosial harus lebih serius mengemban tanggung jawab menghormati HAM, termasuk dalam hal moderasi konten, dan memahami konteks penerapannya.

Sejauh ini, mereka belum berhasil melakukan hal tersebut. Minimnya investasi untuk memahami konteks lokal secara mendalam juga mencakup kegagalan berinteraksi secara memadai dengan aktor-aktor masyarakat sipil untuk mendapatkan pemahaman tersebut. Koalisi lokal untuk kebebasan berekspresi dan moderasi konten dapat menjembatani kesenjangan antara pemangku kepentingan lokal dan perusahaan media sosial, untuk menjamin bahwa HAM dan konteks lokal yang relevan diintegrasikan dengan benar dalam keputusan moderasi media sosial. Namun ini membutuhkan komitmen serius dari perusahaan media sosial untuk menghormati HAM dan menanggulangi dampak negatif terhadap HAM yang muncul dari lemahnya sistem dan proses moderasi konten mereka.

Laporan ARTICLE 19 *Moderasi Konten dan Kebebasan Berekspresi* menguraikan isu-isu ini dengan lebih mendetail dan merangkum apa yang menjadi pertaruhan: 'Jika perusahaan-perusahaan ini gagal mempertimbangkan berbagai dimensi konteks lokal (linguistik, politik, sosial, budaya, dan ekonomi), proses moderasi konten dapat berdampak dramatis terhadap masyarakat yang menjadi korban, misalnya mempertajam polarisasi dan meningkatkan risiko kekerasan.'

DAFTAR PUSTAKA

ARTICLE 19, *Hate Speech Explained: A Toolkit*, 2015.

ARTICLE 19, *Online Harassment and Abuse against Women Journalists and Major Social Media Platforms*, 2020.

ARTICLE 19, *Side-Stepping Rights: Regulating Speech by Contract. Policy Brief*, 2018.

ARTICLE 19, *Watching the Watchmen: Content Moderation, Governance, and Freedom of Expression. Policy Brief*, 2021.

UNESCO, *Addressing Hate Speech on Social Media: Contemporary Challenges*, 2021.

UNESCO, 'Countering Hate Speech'.

UNESCO, *Finding the Funds for Journalism to Thrive: Policy Options to Support Media Viability*, 2022.

UNESCO, *Guidelines for the governance of digital platforms*, 2023

UNESCO, 'How to Address Online #HateSpeech with a Human-Rights Based Approach?', 2022.

UNESCO, 'The Rabat Plan of Action on the Prohibition of Incitement to Hatred', YouTube video, 2022.

UNESCO, *Safeguarding Freedom of Expression and Access to Information: Guidelines for a Multistakeholder Approach in the Context of Regulating Digital Platforms*, 2023.

UNESCO, 'Towards Guidelines for Regulating Digital Platforms for Information as a Public Good', YouTube video, 2023.

UNESCO, *Windhoek+30 Declaration: Information as a Public Good*, World Press Freedom Day International Conference, November 2021.

Catatan Akhir

[1] Istilah 'ujaran kebencian' dan 'disinformasi' tidak didefinisikan dalam hukum HAM internasional. Atas alasan tersebut, ARTICLE 19 menggunakan kedua istilah tersebut di antara dua tanda kutip dalam keseluruhan buku panduan ini.

[2] Setelah diadopsi dalam resolusi Sidang Umum PBB, UDHR tidak secara tegas mengikat negara-negara anggota. Hanya saja, banyak di antara ketentuan di dalamnya dipandang memiliki kekuatan hukum sebagai kebiasaan internasional sejak diadopsi pada 1948; lihat *Filartiga v. Pena-Irala*, 630 F. 2d 876 (1980) (US Circuit Court of Appeals, 2nd circuit).

[3] UN General Assembly, International Covenant on Civil and Political Rights, 16 Desember 1966, UN Treaty Series, vol. 999, hlm. 171.

[4] Pasal 10 Konvensi Eropa untuk Perlindungan Hak Asasi Manusia dan Kebebasan Fundamental, 4 September 1950; Pasal 9 Piagam Afrika mengenai Hak Asasi Manusia dan Hak Penduduk (Piagam Banjul), 27 Juni 1981; Pasal 13 Konvensi Amerika tentang Hak Asasi Manusia, 22 November 1969.

[5] Lihat European Human Rights Court, *Dink v. Turkey*, paras. 106 dan 137 (Applications no. 2668/07, 6102/08, 30079/08, 7072/09, dan 7124/09), 14 September 2010.

[6] A. Kuczerawy (2017) 'The Power of Positive Thinking, Intermediary Liability and the Effective Enjoyment of the Right to Freedom of Expression', JIPITEC, hlm. 226.

[7] UN Human Rights Committee, General Comment No. 34 on Article 19: Freedoms of Opinion and Expression, CCPR/C/GC/34, 12 September 2011, paras. 12, 17, dan 39.

[8] Pasal 10 of the Konvensi Eropa tentang Hak Asasi Manusia dan Kebebasan Fundamental; Pasal 9 Piagam Banjul, 27 Juni 1981; Pasal 13 of the Konvensi Amerika tentang Hak Asasi Manusia, 22 November 1969. Untuk penjelasan tentang Tes Tiga Bagian, lihat UNESCO, 'The Legitimate Limits to Freedom of Expression: The Three-Part Test', video YouTube.

[9] Toolkit "Hate Speech" dari ARTICLE 19 memberikan panduan identifikasi 'ujaran kebencian' dan cara efektif merespons sambil tetap melindungi kebebasan berekspresi dan kesetaraan. Lihat juga UNESCO, *Addressing Hate Speech on Social Media: Contemporary Challenges*, dan UNESCO, 'How to Address Online #HateSpeech with a Human Rights-Based Approach?', video YouTube.

[10] Lihat juga video UNESCO, 'The Rabat Plan of Action on the Prohibition of Incitement to Hatred'.

[11] Dalam laporan ini, kami merujuk dan mengutip panduan komunitas, syarat dan ketentuan, dan dokumen serupa dari berbagai platform media sosial sebagaimana adanya pada saat penulisan. Ini adalah dokumen-dokumen hidup yang sering diubah atau diperbarui, dan mungkin tidak lagi memuat materi yang dibahas.

[12] Untuk jenis-jenis informasi keliru, lihat UNESCO, *Journalism, Fake News & Disinformation*.

[13] Laporan Council of Europe *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making* mengajukan perbedaan antara 'misinformasi', 'disinformasi', dan 'malinformasi' sebagai berikut: misinformasi terjadi ketika informasi keliru disebarkan tanpa maksud menimbulkan kerugian; disinformasi terjadi ketika informasi keliru sengaja disebarkan untuk menimbulkan kerugian; malinformasi

terjadi ketika informasi yang benar disebarkan dengan tujuan menyebabkan kerugian, kerap dengan memindahkan informasi yang dirancang untuk tetap privat ke ruang publik.

[14] Contoh terbaru termasuk Turki, Tunisia, Sudan, dan Inggris.

[15] Dalam konteks ini, penting dicatat perlunya menciptakan lingkungan yang memberdayakan hak atas kebebasan berekspresi dan kesetaraan serta menjawab penyebab pokok 'disinformasi' dan 'ujaran kebencian'. Negara, misalnya, harus berfokus pada kewajiban positif untuk memajukan lingkungan komunikasi yang bebas, independen, dan beragam, mencakup keragaman media dan literasi media dan digital, yang menjadi kunci penanganan 'disinformasi' dan 'ujaran kebencian'. Untuk detail lebih lanjut, lihat UNESCO, 'Countering Hate Speech'; ARTICLE 19, *'Hate Speech' Explained: A Toolkit*; ARTICLE 19, 'Submission to UN Special Rapporteur on Freedom of Expression and "Disinformation"'; dan UNESCO, *Media and Information Literate Citizens: Think Critically, Click Wisely!*

[16] *Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*, dikembangkan oleh Perwakilan Khusus Sekretaris Jenderal untuk isu hak asasi manusia dan perusahaan transnasional dan entitas bisnis lainnya. Komisi HAM PBB mendukung *Guiding Principles* dalam resolusi 17/4 16 Juni 2011.

[17] *UN Guiding Principles*, Prinsip 15.

[18] UN Human Rights Council, Report of the Special Rapporteur on Freedom of Expression, 6 April 2018, A/HRC/38/35, paras. 45–46.

[19] UN Doc., A/HRC/38/35, paragraf 52.

[20] UN Doc., A/HRC/38/35, paras. 28 (untuk proporsionalitas sanksi) dan 59 (untuk keefektifan berbagai metode pemulihan).

[21] UN Doc., A/HRC/38/35, paragraf 11.

[22] UN Doc., A/HRC/38/35, paragraf 50.

[23] UN Doc., A/HRC/38/35, paragraf 51.

[24] UN Doc., A/HRC/38/35, paragraf 52.

[25] Lihat ARTICLE 19, *Watching the Watchmen*, hlm. 16.

[26] ARTICLE 19, *Content Moderation and Freedom of Expression*, hlm. 18.

[27] Untuk praktik terbaik, lihat Spandana Singh dan Kevin Bankston, The Transparency Reporting Toolkit: Content Takedown Reporting.

[28] ARTICLE 19, Side-Stepping Rights, hlm. 16-17.

[29] ARTICLE 19, *Internet Intermediaries: Dilemma of Liability*, hlm. 3.

[30] Pengadilan Tingkat Pertama memutuskan bahwa, sebagai administrator ,Chiranuch Premchaiporn bertanggung jawab memantau konten di forum dengan ketat, karena dapat berdampak negatif terhadap

keamanan nasional serta hak-hak dan kebebasan orang lain. Putusan ini kemudian dikukuhkan di pengadilan yang lebih tinggi.

[31] Untuk ringkasan putusan, lihat Columbia University Global Freedom of Expression, *Prosecutor v. Chiranuch Premchaiporn*.

[32] Special Rapporteur on the Protection and Promotion of Freedom of Opinion and Expression, Report of 16 May 2011, A/HRC/17/27, paragraf 43.

[33] UN Doc., A/HRC/17/27, paragraf 42.

[34] Lihat, misalnya, Online Safety Bill Inggris atau Digital Services Act Uni Eropa.

[35] Lihat, misalnya, RUU Keamanan Daring Inggris (walaupun 'legal tapi berbahaya' untuk ketentuan dewasa telah digantikan dengan kewajiban perusahaan menegakkan Syarat dan Ketentuan) atau French Draft Bill on Countering Online Hatred (dikenal sebagai Loi Avia atau RUU Avia), yang kemudian dinyatakan inkonstitusional oleh Pengadilan Konstitusional Prancis (Conseil d'État).

[36] J. Mchangama dan J. Fiss, *The Digital Berlin Wall: How Germany (Accidentally) Created A Prototype for Global Online Censorship*; lihat komentar tambahan dari penulis di 'The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two'.

[37] ARTICLE 19, *Content Moderation and Local Stakeholders in Bosnia and Herzegovina*, hlm. 39.

[38] Pasal 17 proposal European Media Freedom Act.

[39] Pasal 17 proposal European Media Freedom Act.

[40] Pencocokan tagar memberi 'sidik jari' digital unik pada gambar dan video yang sebelumnya telah terdeteksi berbahaya. Konten buatan pengguna yang baru teridentifikasi berbahaya dapat otomatis dihapus apabila tagar yang terkomputasi sesuai dengan tagar yang tersimpan dalam database konten berbahaya yang diketahui. Lihat Ofcom, *Use of AI in Online Content Moderation*, hlm. 48.

[41] Walaupun teknologi tagar digital menggunakan pengenalan gambar, teknik ini juga bisa digunakan secara lebih luas – misalnya untuk mengidentifikasi objek spesifik dalam sebuah gambar, misalnya senjata. Lihat S. Singh (2019) 'Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content', *New America*, Juli, hlm. 14.

[42] NLP adalah teknik di mana teks diuraikan untuk memprediksi maknanya – misalnya, apakah teks tersebut mengungkapkan opini positif atau negatif. Lihat Center for Democracy and Technology, 'Mixed Messages? The Limits of Automated Social Media Content Analysis', hlm. 9.

[43] Singh, 'Everything in Moderation', hlm. 7.

[44] Singh, 'Everything in Moderation', hlm. 7.

[45] Center for Democracy and Technology, 'Mixed Messages?', hlm. 5.

[46] Singh, 'Everything in Moderation', hlm. 13.

[47] Singh, 'Everything in Moderation', hlm. 18.

[48] ARTICLE 19, *Content Moderation*, hlm. 17; Brennan Center for Justice, *Double Standars in Social Media Content Moderation*, hlm. 18.

[49] Center for Democracy and Technology, 'Mixed Messages?' hlm. 6.

[50] Center for Democracy and Technology, 'Mixed Messages?' hlm. 14.

[51] Singh, 'Everything in Moderation', hlm. 16.

[52] B. Heller (2019) 'Combating Terrorist-Related Content through AI and Information Sharing', Institute for Information Law, 26 April, hlm. 3.

[53] ARTICLE 19, *Side-Stepping Rights*, hlm. 32.