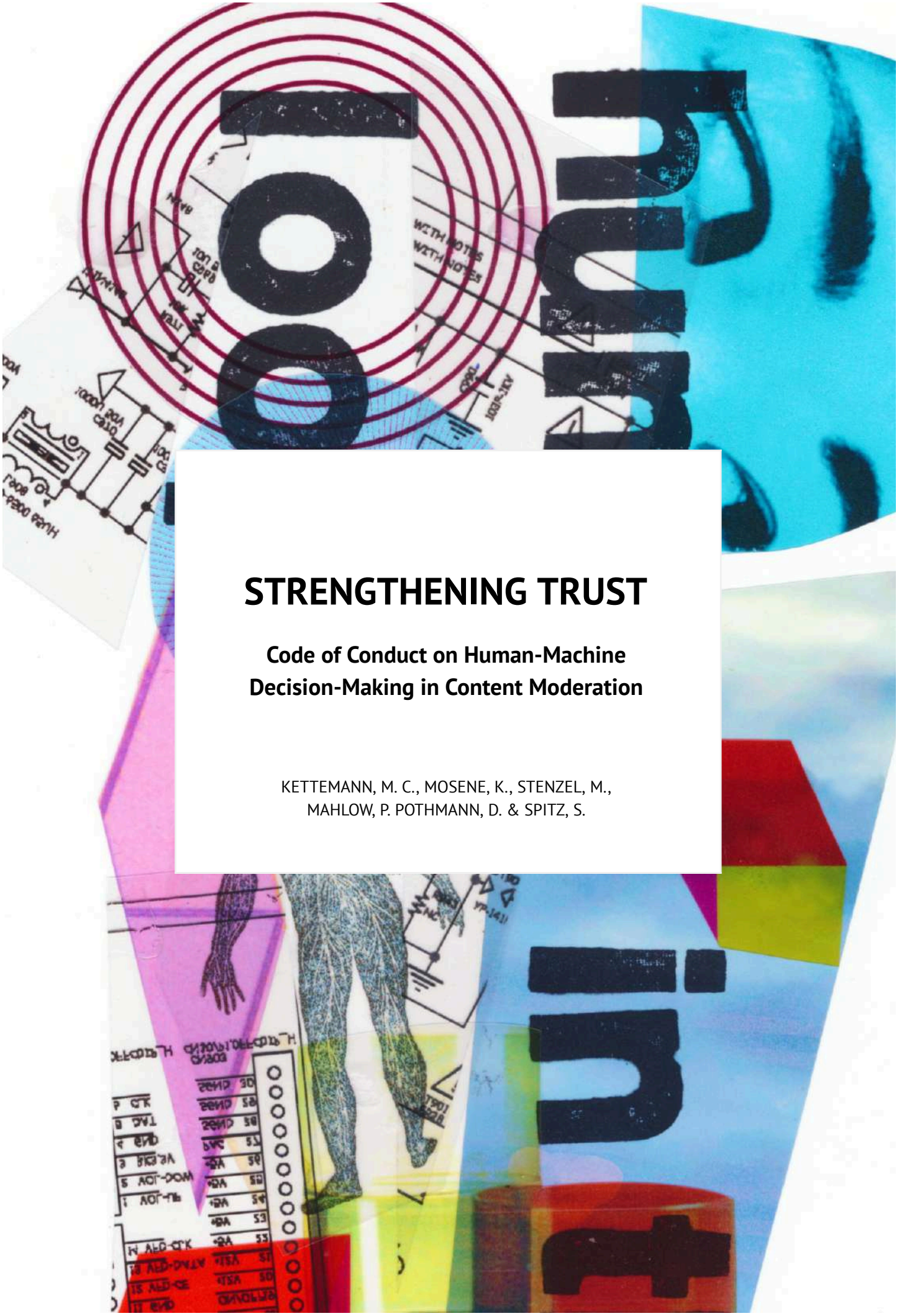


STRENGTHENING TRUST

Code of Conduct on Human-Machine Decision-Making in Content Moderation

KETTEMANN, M. C., MOSENE, K., STENZEL, M.,
MAHLOW, P. POTHMANN, D. & SPITZ, S.

STRENGTHENING TRUST

Code of Conduct on Human-Machine Decision-Making in Content Moderation

Berlin, 19. November 2025

Recommended citation: Kettemann, M. C., Mosene, K., Stenzel, M., Mahlow, P., Pothmann, D. & Spitz, S. (2025). *Strengthening Trust. Code of Conduct on Human-Machine Decision-Making in Content Moderation*. Alexander von Humboldt Institute for Internet and Society. DOI: 10.5281/zenodo.17650988

This code of conduct was developed as part of the four-year project "Human in the Loop? Autonomy and Automation in Socio-technical Systems", conducted by the Alexander von Humboldt Institute for Internet and Society (HIIG) in Berlin. The case study on content moderation examines the interplay between algorithmic systems and human judgement, situated at the intersection of private governance, public law, and societal values. The project is funded by Stiftung Mercator.

CONTENTS

Background

Providers of digital services such as Meta and TikTok have a significant influence on human-machine interaction in the governance of online content. The Code of Conduct sets out key obligations to promote the responsible, rights-preserving, and transparent design of (partially) automated systems and to share best practices. The aim is to strengthen trust in these systems and their collaboration with human moderators, minimise risks such as overblocking and underblocking, and preserve the autonomy, security, and fundamental rights of all parties involved.

The code complements existing legal frameworks such as the Digital Services Act (DSA) and the Artificial Intelligence Act (AIA), but goes beyond them: It sets industry-specific guidelines that incorporate human rights, sensitivity to discrimination, and civil society perspectives. In this way, it addresses gaps in transparency, accountability, participation, and the protection of marginalised groups. Normative principles such as the protection of fundamental rights, non-discrimination, and participation are translated into concrete measures and reinforced through cooperative voluntary commitments.

DEVELOPMENT OF THE CODE OF CONDUCT

The code of conduct addresses the crucial, yet often overlooked relationship between human moderators and technological systems. It emphasises accountability, due process, fairness, and transparency in algorithmic decision-making. Consisting of ten principles in total, it complements the DSA framework in relevant areas and aims to foster trust in semi-automated systems. Ultimately, the code seeks to (re)shape the human-machine relationship in content moderation.

The code of conduct was developed as part of the four-year project “Human in the Loop? Autonomy and Automation in Socio-technical Systems”, conducted by the Alexander von Humboldt Institute for Internet and Society (HIIG) in Berlin. The case study on content

moderation examines the interplay between algorithmic systems and human judgement, situated at the intersection of private governance, public law, and societal values.

The development of the code followed an iterative process based on scientific principles, integrating regulatory expertise and stakeholder engagement, in line with Article 45 DSA. Stakeholders included academic experts, members from NGOs, advocacy, and policy groups, representatives from German platforms, as well as German branches of very large online platforms (VLOPs) and elected representatives of the German Bundestag.

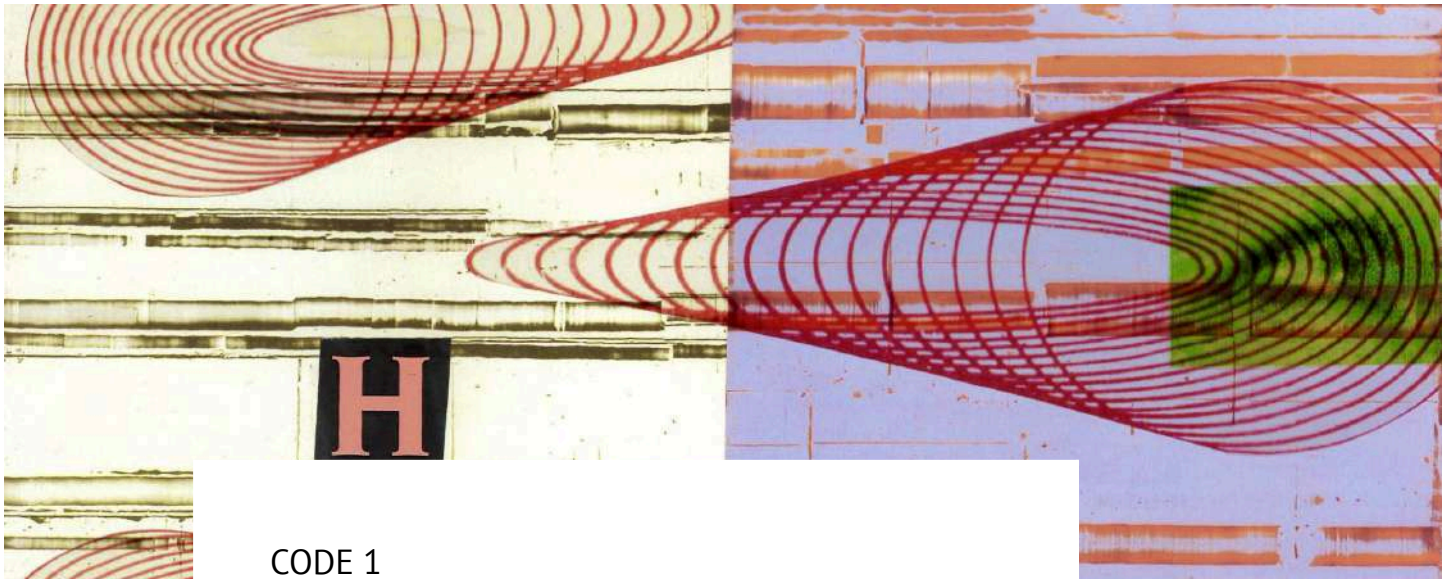
EXPERTS

Thanks to all experts who contributed their knowledge, time, and perspectives to this code of conduct process. Your expertise and critical insights have been essential to its depth and direction.

- **Svea Windwehr**
- **Josephine Ballon**, HateAid
- **Sven Winter**, gutefrage.net
- **Rainer Rehak**, Weizenbaum Institut
- **Clara Helming**, Algorithmwatch
- **Charlotte Lauber**, Landeszentrale für politische Bildung Niedersachsen
- **Johann Laux**, Oxford Internet Institute
- **Charlotte Freihse**, Bertelsmann Stiftung
- **Davy Wang**, Gesellschaft für Freiheitsrechte
- **Daniela Heinemann**, nebenan.de
- **Lisa Buckmiller**, nebenan.de
- **Philipp Kellmeyer**, Uni Mannheim
- **Markus Langer**, Abteilung Arbeits- und Organisationspsychologie, Uni Freiburg
- **Stephan Bohn**, Alexander von Humboldt Institute for Internet and Society
- **Marie Stahlhofen**, HateAid
- **Christoffer Horlitz**, Amnesty International

CREATIVE REFLECTIONS

The collages shown on this website were created during a joint project team workshop led by artist Fehmi Baumbach. They reflect the diverse topics, perspectives, and areas of focus covered by the code of conduct.



CODE 1

Recognition of various factors that influence the degree of automation

Automation technologies should be implemented in a socially responsible and context-sensitive manner. This means that decisions on the degree of automation must take into account both technical factors (e.g., risk of reproducing of content that infringes rights, scalability) and sociocultural aspects, such as societal power relations, the protection of minorities, political contexts, and the risk of instrumentalisation by the state.

In particular, it is important to

- Clearly define the degree of automation (fully automated, semiautomated, or only reviewed by humans) with reference to the above-mentioned factors
- Design systems so that they operate as independently as possible from specific user profiles. Automated decisions should only consider sensitive user characteristics to the extent that is minimally necessary in order to minimise bias
- Analyse cultural and social conditions in a differentiated manner (e.g., handling of LGBTIQA+[1] content in repressive contexts, linguistic characteristics, historically sensitive issues)
- Minimise risks to marginalised groups in automated decisions in a targeted manner.
- Consider both false positives (overblocking of legitimate content) and false negatives (failure to recognise harmful content)
- Understand social acceptance not as a uniform variable but as a pluralistic negotiation process involving the affected groups

[1] **LGBTIQA+** is an acronym that stands for lesbian, gay, bisexual, transgender, intersex, queer/questioning, and asexual, as well as other identities. The term is used to describe a wide range of sexual orientations, gender identities, and gender expressions. The plus sign (+) at the end stands for all other identities that are not explicitly included in the letters.

IMPLEMENTATION

✓ ETHICAL GUIDELINES

for the evaluation and use of automated systems that explicitly address diversity, context sensitivity, and the dangers of asymmetrical power relations. This includes the introduction of a human oversight panel within the organisation that regularly evaluates whether existing automation solutions preserve pluralistic freedom of expression and avoid discrimination.

✓ INVOLVEMENT OF AFFECTED GROUPS THROUGH PARTICIPATORY FEEDBACK PROCESSES

(e.g., through participatory consultations with affected communities, such as queer groups, linguistic minorities, etc.) to assess the social impact and acceptance of automated systems.

✓ ESTABLISHMENT OF TRANSPARENT PROCEDURES FOR EVALUATING AND AUDITING[2]

automated systems with regard to cultural, political, and societal implications, including the development of **dynamic risk assessment** models that take into account both technical maturity and potential societal conflicts (e.g., in authoritarian states).

✓ INTRODUCTION OF HUMAN CONTROL INTERVALS OR INTERVENTION POINTS

that must be planned for – both for quality assurance and error correction.

✓ REVIEW OF GEOFENCING[3] MECHANISM

where necessary, but with particular attention to the risk of geoblocking in the context of state repression.

✓ ESTABLISHMENT OF BENCHMARKS

(e.g., rate of overblocking/underblocking by user category, false positives/negatives for content from marginalised groups) in consultation with the ethics board and external institutions (research, civil society organisations) for periodic performance reviews with regard to identified deficiencies.

---

[2] **Auditing** is the systematic and independent review of processes, systems, or organizations to assess and document their compliance with established standards, regulations, or quality criteria.

[3] **Geofencing** refers to a technology that uses virtual, geographically defined boundaries to automatically trigger certain actions or notifications.

CODE 2

Responsible delegation

In order to minimise incorrect decisions and avoid inappropriate or excessive automation, tasks should only be delegated to automated systems if they are technically mature, transparently verifiable, and societally responsible – especially with regard to critical, ethically sensitive, or context-dependent decisions. The following applies: The higher the degree of delegation to automated systems, the greater the responsibility to safeguard them with robust control and fallback and feedback mechanisms. A high degree of automation always means a high degree of delegation of decision-making responsibility – this correlation must be reflected on and limited, especially in the case of content that is potentially relevant under criminal law, such as threats of violence, announcements of the intention to commit mass shootings/rampage attacks, or incitement to hatred and agitation. In such cases, automatic intervention alone is neither appropriate nor responsible – additional human evaluation remains mandatory.

Technical maturity refers to the state in which a system:

- Has been proven in independent audits[4] to make consistently accurate decisions (e.g., measured by very low false positive/negative rates, fairness metrics[5])
- Has been security tested to the latest standards (e.g., through adversarial testing[6])
- Can be continuously monitored and improved

[4] An **audit** is a systematic review process in which processes, systems, or organizations are evaluated against established criteria to verify their compliance with standards, guidelines, or quality requirements.

[5] **Fairness metrics** are mathematical measures used to evaluate and quantify whether algorithmic systems or decision-making processes treat different groups of people equally and do not exhibit systematic discrimination or bias.

[6] **Adversarial testing** is a method of specifically testing AI systems, in which manipulated or challenging inputs are used to uncover vulnerabilities, incorrect decisions, or security gaps.

- Is equipped with clear fallback/escalation mechanisms for special cases
- And has been validated against a training set of curated, human-reviewed decisions – ideally with representative, diversely annotated content to minimise bias[7]

Critical decision-making processes are those that

- Pose a risk to fundamental rights (e.g., freedom of expression, data protection, discrimination)
- Could have irreversible consequences for individuals or groups
- Take place in societally highly standardised or conflict-ridden contexts
- Or concern criminally relevant content where incorrect decisions can have serious real-world consequences (e.g., public safety, prevention of violence or hate speech)

IMPLEMENTATION

✓ TRANSPARENT THRESHOLDS AND CONTEXT DEFINITIONS

Define internal, regularly reviewed criteria catalogues that define when a decision is considered critical and when human intervention is absolutely necessary.

- **Support** efficient, more context-sensitive delegation to suitable moderation teams through **language-based systems** (e.g., large language models), limited to critical case constellations.
- Plus: Establishment of a **publicly accessible decision register** for automated interventions with transparent intervention thresholds.

✓ TRAINING-BASED SYSTEM DEVELOPMENT

Implementation of a quality-checked, annotated training dataset as a reference for automated systems. This dataset should be based on traceable, human-made moderation decisions and

[7] Algorithmic **bias** refers to systematic errors in algorithms that lead to unfair, discriminatory, or unbalanced results for certain groups of people, often caused by biased training data, incomplete data sets, or unconscious biases in algorithm development.

updated regularly; linguistic, media, and cultural diversity should be taken into account in the training material.

✓ HUMAN OVERSIGHT AS A MANDATORY COMPONENT

- Randomised checks of delegation decisions.
- Time-sensitive embedding of moderator feedback with the option of weighting (see also point 9).
- Implementation of a prioritisation system for posts with potential criminal relevance, in which automated systems flag content but are not allowed to make final decisions without human review.
- Participation of civil society actors in the development and advancement of systems.

✓ DERISKING THROUGH MONITORING

Technology impact assessment as a continuous process for evaluating the long-term effects of automated systems.

- Use of independent external audits to assess technical suitability.

✓ DEVELOPMENT OF ASSESSMENT INDICATORS

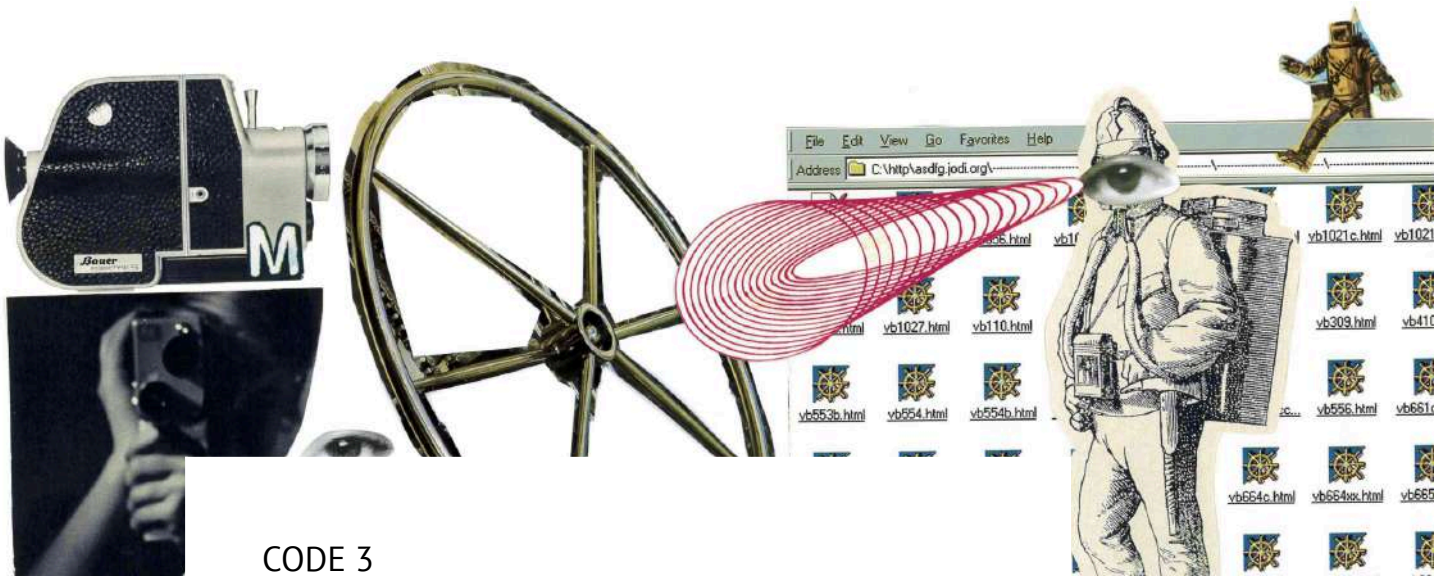
Development of quantitative and qualitative indicators for assessing the consequences of automation (accuracy[8], bias[9], user feedback[10], risk indices[11]).

[8] **Accuracy** refers to the precision of an (AI) system, i.e., the extent to which its predictions or decisions correspond to actual results.

[9] **Bias describes** systematic distortions in data or algorithms that can lead to unfair or discriminatory results.

[10] **User feedback** is feedback from users about a system or content that can contribute to improving functionality, accuracy, or fairness.

[11] **Risk indices** are metrics or ratings that assess the potential risk of a system, process, or behavior and are used for decision-making or regulation.



CODE 3

Emergency mechanisms for human intervention

In safety-critical, particularly complex, or fundamental-rights-threatening situations, human control must be strengthened. Automated systems should not continue autonomously if there are indicators of significant risks to democratic processes, public safety, fundamental rights, or user protection. Automation must be interruptible at any time – by clearly defined intervention procedures, responsible persons, and transparent documentation. The goal is risk-adequate, ethically acceptable human-machine interaction that relies on preventive and reactive emergency mechanisms.

Timely intervention is particularly necessary when

- Safety-critical situations arise (e.g., threats to users' physical or digital safety)
- Particularly complex situations arise in which the technical model logic conflicts with real-world contextualisation (e.g., through ambiguous language, cultural connotations, novel/unprecedented phenomena)
- Systemic risks as defined by the Digital Services Act or the AI Act arise (e.g., threats to democratic processes, targeted disinformation, discrimination relevant to fundamental rights)
- Significant impacts on individuals or groups are to be expected—especially for vulnerable or marginalised user groups

IMPLEMENTATION

✓ INTRODUCTION OF “OVERRIDE” FUNCTIONS

(e.g., stop button, pause mechanism) that allow human control at any time without compromising security or system performance. A system should be able to proactively and time-sensitively alert users to potential emergencies.

✓ ESTABLISHMENT OF A MULTILEVEL ESCALATION PROCEDURE THAT REGULATES

- When intervention is permissible or necessary
- Who carries the intervention out (e.g., safety teams, panel instances)
- How decisions are documented, reversible, and verifiable.

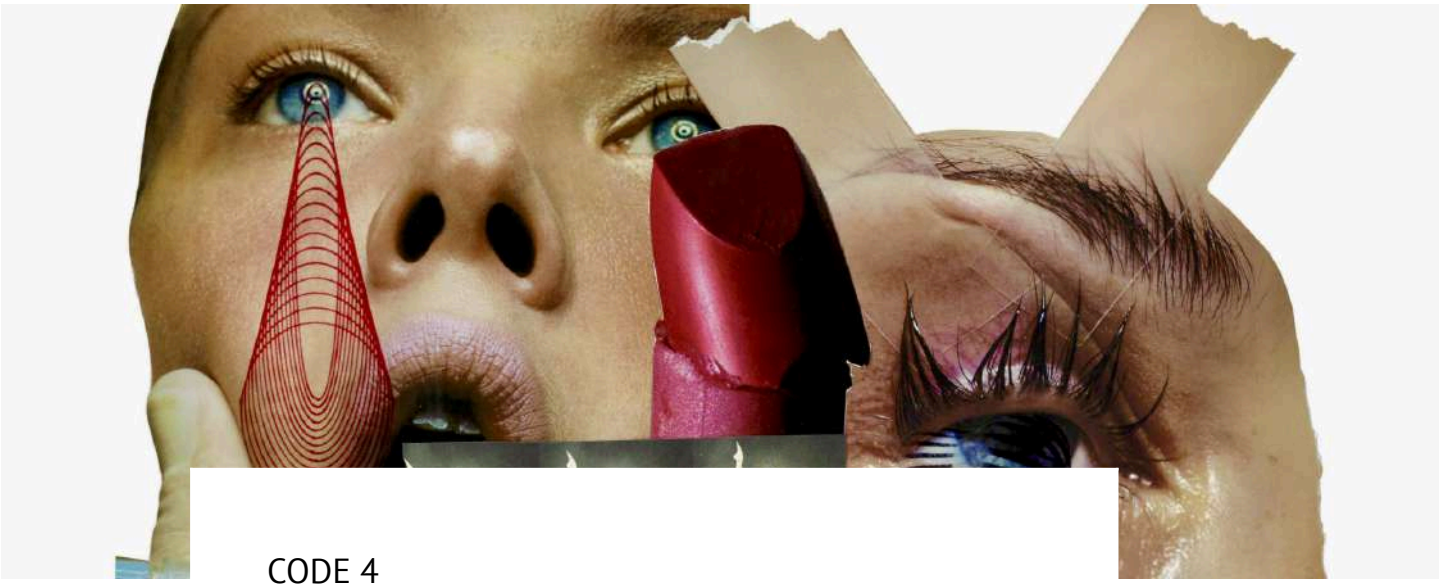
✓ BASIS

- Development of a **risk matrix** that categorises typical intervention scenarios and can be dynamically expanded. This matrix should not be a static list but rather be dynamically adaptable and capable of further development through civil society expertise. Furthermore, it should not be used as a "checkbox" solution but rather be embedded in a risk-adaptive assessment. Specifically, it should show examples, contexts, and severity levels (e.g., impact level, user group, system response[12]) in a comprehensible structure without excluding new or unexpected scenarios.
- Establishment of a **reversal procedure** for intervention decisions with an external control body (reporting and evaluation).
- Development of a **training program** for all employees involved in moderation or system supervision on the safe use of intervention mechanisms (see point 8).
- Involvement of **civil society organisations, research, authorities, affected groups, and experts** (e.g., from the fields of discrimination protection, media ethics, IT security) in the ongoing definition, evaluation, and further development of risk categories and emergency procedures.
- Establishment of **feedback loops** for a multilevel escalation process[13] between the community, moderation, and system design in order to identify and address long-term risks at an early stage.

---

[12] In the context of content moderation, **impact level** describes the extent of the potential harm caused by content (e.g., low to high risk), **user group describes** the specific target groups or communities that could be affected by certain content (e.g., minors, marginalized groups), and **system response** refers to the automated or manual measures that the moderation system takes in response to problematic content (e.g., warning, deletion, account suspension, forwarding to human moderators).

[13] **Escalation procedures** in content moderation are structured processes in which complex, borderline, or particularly serious cases are forwarded from automated systems to human moderators or specialized teams to ensure appropriate and contextual assessment and decision-making.



CODE 4

Suspension of automated moderation when complexity is indicated

Ensure that automated moderation processes are interrupted when content exhibits a high degree of cultural, ethical, or legal complexity. In such cases, human intervention (dynamic escalation for human review) must be mandatory in order to adequately consider fundamental rights, cultural contexts, and ambiguous interpretations.

Complexity indicators are characteristics that indicate that content cannot be evaluated by automated systems. These include:

- Ambiguity of linguistic expressions (irony, sarcasm, context dependency, regional idioms)
- Cultural and religious symbolism that can be interpreted differently depending on the region or group,
- Topics that strongly affect marginalised groups (e.g., queer identities, racialised perspectives, colonial or antisemitic language elements)
- Overlaps with sensitive political contexts (e.g., elections, protest, dissent, authoritarian narratives/propaganda, war).

The list of complexity indicators shall be public, dynamic, and regularly updated with input from interdisciplinary expert groups (especially in computer science, law, and social sciences). Users, civil society organisations, and moderation teams are explicitly encouraged to suggest new indicators. A defined review process shall ensure that new suggestions are evaluated and documented in a participatory manner.

Note: These indicators shall trigger a risk-based, tiered escalation process that combines automated preliminary analyses with human review.

IMPLEMENTATION

✓ EARLY DETECTION

Development of automated **early detection** that escalates content to human decision-makers based on complexity indicators and prioritises it, if necessary. This early detection should be subject to comparable audit requirements as content governance systems[14] as a whole (see point 2). Content that falls under multiple indicators is prioritised and treated with increased depth of review.

✓ COMPLEXITY INDICATORS

Option for users to explicitly refer to **complexity indicators** when reporting content.

✓ ADEQUATE TRAINING

Moderators should receive adequate training in **human rights, cultural, and contextual sensitivity** so that they can adequately assess and handle cases for escalation. This also includes **defining human resources and minimum standards for the provision** of qualified moderators, including in terms of languages, cultural knowledge, psychological resilience, and legal knowledge (see point 8).

[14] **Content governance** refers to the strategic management, regulation, and distribution of responsibility in dealing with digital content on platforms, including guidelines, moderation processes, and technical implementation.

CODE 5

Human-centered interface design & psychological support

All systems and interfaces used in the context of content governance shall be developed with a human-centered approach. The design of systems, their techno-physical interfaces, and digital user interfaces should minimise physical and psychological stress and enable natural forms of interaction for moderators – especially in demanding, highly repetitive, or potentially stressful and disturbing contexts. Mental health is not an individual responsibility but part of the employer's duty to provide a safe working environment and care for their employees.

Who are moderators (broadly defined)?

- Internal teams & outsourced service providers
- So-called high-level expert groups
- System administrators
- Community members, if applicable, when they take on moderation functions (e.g., via platform reporting systems)
- Trusted flaggers (within the meaning of Art. 22 DSA)

What does "natural interaction" mean?

- Transparent & understandable
- Barrier-free
- Psychologically relieving (e.g., options to take a break, preview blockers for distressing content)
- Intuitive

IMPLEMENTATION

✓ REGULAR USABILITY TESTS

with different user groups (e.g., based on HCI standards).

✓ EVALUATION ACCORDING TO USER-CENTERED DESIGN PRINCIPLES, SUCH AS:

- Comprehensibility
- Controllability
- Error prevention
- Promotion of emotional resilience

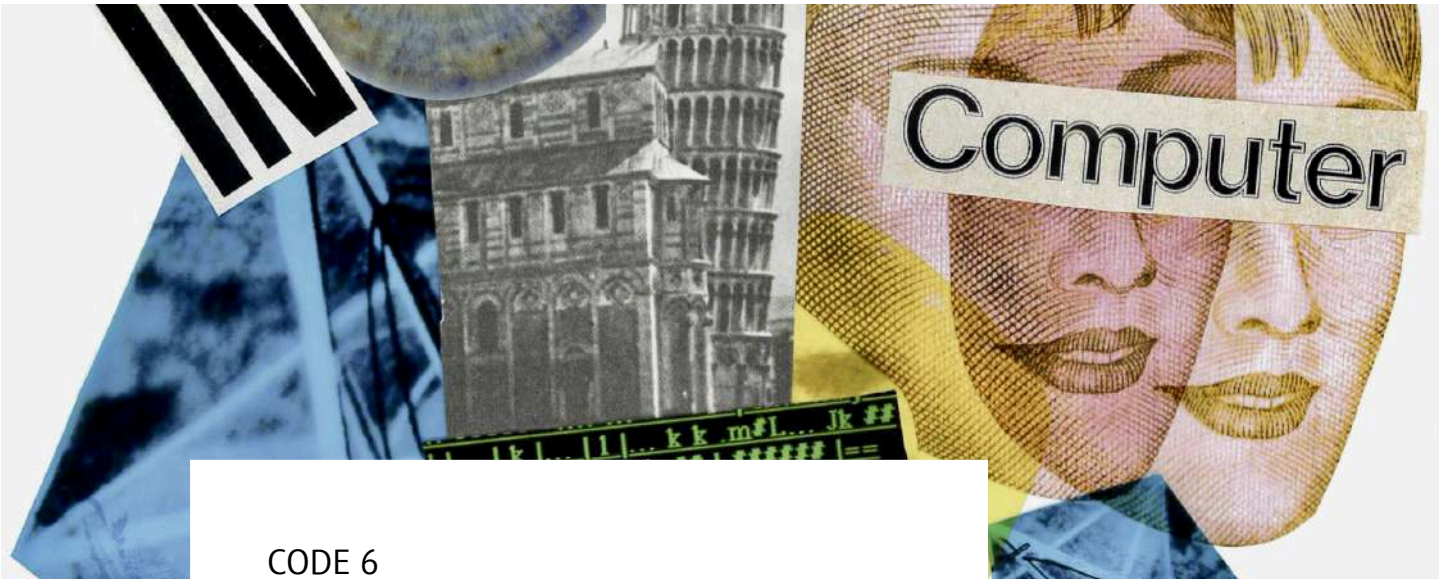
✓ TRAUMA-SENSITIVE DESIGN FOR DISTRESSING CONTENT, E.G.:

- Blurred image previews
- Staggered preview/display of sensitive content
- Visually neutral categorisation of violent material
- Automated avoidance of unnecessary repetition (e.g., through system-supported case filtering)
- Option to immediately cancel preview
- Grayscaling[15]

✓ INTRODUCTION OF REFLECTION, FEEDBACK, AND RELIEF STRUCTURES

for moderators when dealing with highly sensitive cases, in particular through regular and **acute professional psychological support services** to an appropriate extent. The availability of such support must not be restricted by daily quotas.

[15] In content moderation, **grayscaling** refers to a measure whereby content is not completely removed, but rather its visibility is restricted or visually toned down in order to protect users from potentially problematic content.



CODE 6

Balancing data protection and contextual information

Automated and semiautomated content moderation requires a careful balance between the protection of personal data and the consideration of contextual information necessary for fair, transparent, and nondiscriminatory decisions. All data processing steps – including the analysis of post content, metadata, usage contexts, and, where applicable, personal account information – shall be guided by the principles of data minimisation, purpose limitation, and contextual appropriateness.

In this context, *contextual appropriateness* means weighing the fundamental rights at stake (such as freedom of expression, data protection, protection against discrimination, or protection against violence) in a proportionate manner, taking into account the social, communicative, and technical context of a post. The collection and evaluation of so-called context data, such as visibility settings, audience targeting, communication space (public, semi-public, private), time sequence, interaction patterns, or platform architecture, may only take place if this is essential for the assessment of content.

Key principles

- Data protection in accordance with the GDPR (Art. 1,5 GDPR: Protection of natural persons with regard to the processing of personal data and on the free movement of such data)
- Principle of proportionality and balancing of rights under the DSA: Moderation decisions must be proportionate to the potential infringement of fundamental rights (Art. 14 DSA).
- Consideration of users' privacy settings (e.g., private stories, closed groups, protected profiles vs. public content) and prioritisation according to reach

IMPLEMENTATION

✓ DEVELOPMENT OF CONTEXT-PRESERVING ANALYSIS METHODS

e.g., through semantic context recognition, hierarchical discourse analysis, or space-time classification/context delimitation; without the inclusion of personal data; with a focus on the context relevant to the moderation decision.

✓ CONDUCT STANDARDISED DATA PROTECTION IMPACT ASSESSMENTS

for all systems that make automated or semiautomated decisions about content, with a particular focus on risks to marginalised groups.

✓ DEFINITION AND WEIGHTING OF NECESSARY CONTEXTUAL DATA

depending on the type of content and form of communication (e.g., irony, activism, violence prevention contexts); this includes visibility settings, target group addressing, posting time, and technical distribution mechanisms.

✓ GRADUATED RIGHTS BALANCING IN ACCORDANCE WITH THE DSA

Systems must recognise when an automated decision may have a significant impact on freedom of expression or privacy and ensure human review.

✓ DYNAMIC CONTEXT RECOGNITION

Systems must recognise whether content originates from private, temporary, or protected communication spaces and adapt their analysis accordingly.

CODE 7

Fairness and nondiscrimination

(Semi)automated systems, especially in the area of content moderation and recommendation, should be designed and regulated in such a way that they detect structural exclusion, algorithmic biases, and unintended amplification mechanisms at an early stage and effectively limit them. Fairness is understood here as equal access, nondiscrimination based on human rights, and the enabling of equal participation.

Content moderation systems (CMS) have a particular responsibility in this regard: They must not only detect and remove illegal content but also ensure that their mode of operation does not indirectly disadvantage marginalised groups, for example, through higher error rates in the detection of dialects or nondominant forms of language or through the unequal removal of legal content. These systems must be transparent, accountable, and nondiscriminatory in accordance with the requirements of the Digital Services Act (DSA). Platforms are also required to provide clear rules, transparent processes, and effective complaint mechanisms.

Recommender systems (RS) (algorithmic recommendation systems for sorting, prioritising, or controlling the visibility of content) play a central role in content governance. They largely determine what content users see and what they do not. The underlying reinforcement logic is usually based on engagement rates such as likes, shares, or watch time. However, these metrics can have discriminatory side effects if, for example, they disadvantage content that is less emotional or comes from groups whose contributions receive less feedback – such as people with disabilities, FLINTA/LGBTQIA+, BIPoC, or nondominant language communities.

Content should therefore not be prioritised solely on the basis of how polarising or emotional it is. The goal is to design recommendation algorithms that do not push problematic dynamics such as toxic discourse, hate speech, or disinformation as strongly and do not structurally disadvantage diverse, contextualised content. This means designing recommendation

algorithms in such a way that they enable democratic participation, diversity of opinion, and fair access in an unequal digital space.

In this context, **contextualised equal treatment** does not mean treating all content or users identically but rather systematically considering social inequalities, structural discrimination, and existing access restrictions. Fair algorithmic weighting requires an adaptive system design that reveals distortions and reinforcement loops, makes discriminatory effects verifiable, and enables participatory corrections.

IMPLEMENTATION

✓ MEANINGFUL STAKEHOLDER ENGAGEMENT

All relevant audit and development steps shall be carried out with the involvement of relevant external perspectives, in particular from civil society organisations, affected communities, and interdisciplinary experts with intersectional, human rights-based expertise.

✓ REGULAR, INDEPENDENT BIAS AUDITS

shall be conducted by interdisciplinary committees (see above) that integrate perspectives critical of discrimination. **Targeted system revisions shall be made after each analysis**.

✓ TRANSPARENT ANALYSIS OF TRAINING AND MODELLING DATA

for representation gaps, historical biases, and unintended exclusions.

✓ USE AND PUBLICATION OF MULTIPLE FAIRNESS METRICS

such as misclassification rates by group, visibility distributions across diverse content, and documented reinforcement mechanisms for emotionally charged, controversial, or minority content.

✓ DEVELOPMENT OF A PUBLIC FAIRNESS DASHBOARD

that presents these metrics in a comprehensible manner and is continuously updated.

✓ SYSTEMATIC EVALUATION AND CONTROL OF AMPLIFICATION MECHANISMS

Recommendation algorithms shall be continuously reviewed to ensure that they do not disproportionately amplify polarising, emotionally charged, or marginalising content. Internal feedback loops between moderation and recommendation systems, interdisciplinary impact assessments of engagement-based rankings, and long-term tests ensure that the effectiveness of these measures is monitored. For virally disseminated content, threshold-based human assessments should be used to ensure fairness, safety, and visibility of marginalised perspectives.

CODE 8

Training and continuing education for moderators

Strengthening the professional competence, ethical confidence, and psychological resilience of moderators who work with automated systems and their effects. The focus is not only on qualification, but also on care, protection, and structural relief.

- Content moderation is high-stress work and requires professional training, pay commensurate with qualifications and workload, psychological support, relief, and supervision.
- Competence building does not equal transfer of responsibility: The responsibility for fair, functional systems does not lie with individual employees but with the organisation as a whole.

IMPLEMENTATION

✓ MANDATORY TRAINING AND CONTINUING EDUCATION PROGRAMS

that combine technical, ethical, and intersectional perspectives (e.g., on algorithmic fairness[16], human rights, discrimination risks, and the functional logic of automated systems). Training courses should take into account linguistic diversity, regional contexts, and cultural codes. This applies to both content and methodological approaches (e.g., case studies in multiple languages and cultural frameworks).

✓ IMPLEMENTATION OF A MENTORING OR PEER COACHING PROGRAM

to support newcomers and promote confidence in dealing with complex automation decisions; opportunity for **supervision**; promotion of **expert groups** that can be called upon in particularly complex cases (similar to "red teams" in IT security[17]); documentation and exchange of **best practices** via internal platforms or knowledge databases.

✓ SUPERVISION

Moderation teams need regular **supervision, time for reflection, and psychosocial support** (e.g., through anonymous counseling services, external support); establishment of a clear framework for **limiting working hours during periods of high stress**; ongoing **evaluation of the stress situation** (quantitative and qualitative).

✓ DYNAMIC KNOWLEDGE TRANSFER

Moderation decisions shall be based on a continuously updated knowledge base (dynamic terms, symbols, hashtags, memes); social and political developments/in context; regular updates on platform guidelines (automatically integrated). Platforms must ensure that guideline changes and newly identified moderation risks are communicated promptly to all relevant parties via internal update systems.

---

[16] **Algorithmic fairness** refers to the goal of designing algorithmic decision-making processes in such a way that they avoid systematic discrimination and achieve results that are as fair and non-discriminatory as possible for all affected groups.

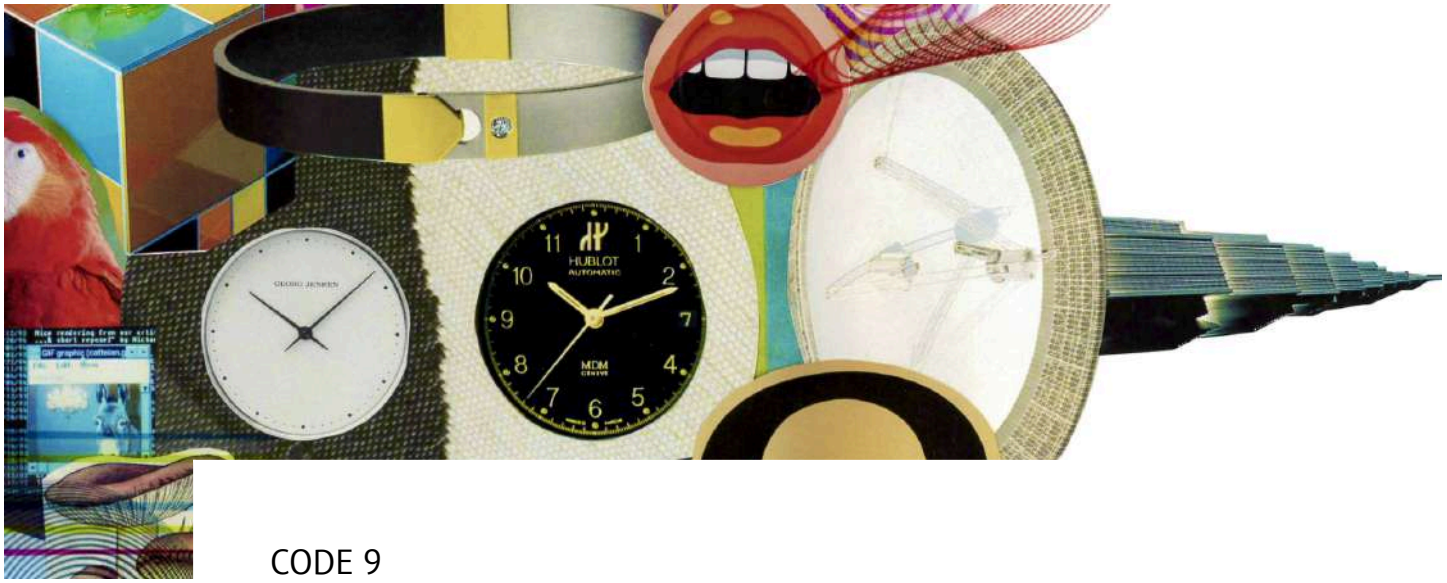
[17] **'Red teams' in IT security** are specialized groups that act from the perspective of potential attackers to uncover security vulnerabilities in systems, networks, or applications using realistic attack strategies and to test defenses against them.

✓ TRAINING COURSES SHOULD FOLLOW PARTICIPATORY, INTERACTIVE PRINCIPLES

(e.g., case studies, simulations, dialogue formats). Where appropriate, external providers should be involved, such as those specialising in discrimination-sensitive education, ethics consulting, or digital rights advocacy. This can be supplemented by cooperation with civil society organisations, research institutes, and professional associations.

✓ REGULAR EVALUATION AND FURTHER DEVELOPMENT OF CONTENT

with the involvement of external experts.



CODE 9

Continuous feedback systems

Ensuring that automated decisions are continuously reviewed and improved through human perspectives – both through internal feedback from moderators and through formalised appeal options for users. This double feedback loop should help to ensure fairness, system learning, and trust.

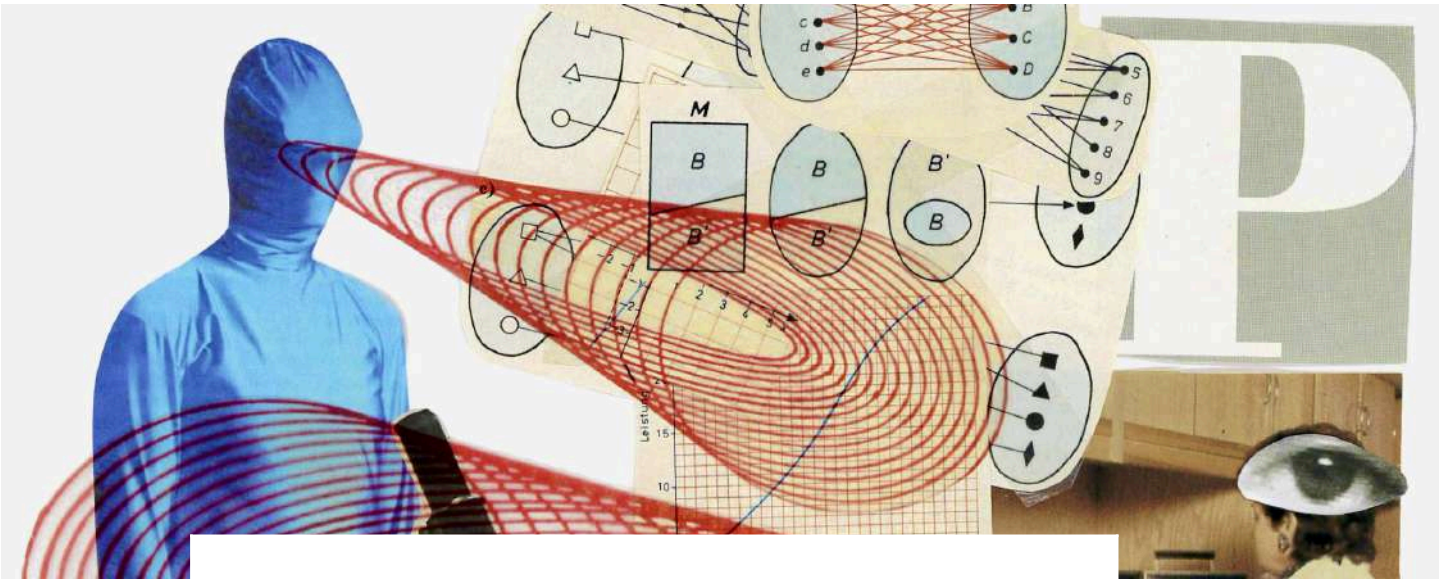
IMPLEMENTATION

✓ MODERATORS → SYSTEMS: INTERNAL FEEDBACK

- User-friendly *feedback buttons* or marking tools that allow moderators to comment on, correct, or flag system decisions for review. Feedback should be fed directly into the further development of the moderation systems via a technical interface to minimise errors systematically and in a time-sensitive manner.
- Involvement of moderators in **regular reflection and review processes**, for example through:
  - Usability workshops
  - Retrospective error analyses
  - Feedback sprints with developers
- Mandatory process evaluation by moderators at fixed intervals.

✓ USERS → PLATFORM: EXTERNAL APPEAL

- Introduction of an **easy-to-use, barrier-free appeal system** within the scope of Art. 20 DSA for moderation decisions with comprehensible justification and transparent feedback. Mandatory human review of all appeals in accordance with Art. 20 (6) DSA – no solely automated final decision. Upon request, affected users will be given **access to an overview of relevant data relating to the processing of their appeal**, such as processing time, parties involved, and outcomes, including justification.
- Users have the option of pointing out **contextual information** (e.g., irony, activism, protected groups) that may have been incorrectly classified by machines.
- Appropriate consideration of objections raised by users in the further development of moderation systems.
- **Data-minimised appeal process** for reporting or affected users (no retraumatisation by forcing them to recount violent or discriminatory experiences in detail). In the case of content that may be relevant under criminal law, **preliminary documentation** shall be securely carried out by authorised actors, in strict compliance with data protection regulations (incl. Artt. 10, 17 GDPR).
- Platforms shall maintain internal, **structured logging of all automated and hybrid moderation decisions with feedback references** (e.g., flagging, correction, objection, result).



CODE 10

Transparency, comprehensibility, and explainability

Decisions made by automated systems must be understandable to users, civil society organisations, academia, public authorities, and regulatory bodies. This includes the disclosure of relevant system information as well as the possibility of access, review of legality, and appeal.

In alignment with the protection of intellectual property and trade secrets, the following elements in particular should be disclosed:

- The core logic of algorithmic decision-making processes: e.g., filter criteria, scoring systems[18], reinforcement mechanisms, model training.
- System characteristics: rule-based systems, type of machine learning (supervised, unsupervised, reinforcement), deep learning, hash-matching, hybrid model architectures, model purpose.
- Training data used: Making them publicly available in accordance with Art. 53 (1) (d) AIA; this applies in particular to the origin of the data, data categories, and possible biases.
- Within the scope of the DSA: description of automated and human decision-making steps, including decision-making bases (terms and conditions, guidelines, legal requirements), see Art. 17 DSA.

[18] **Scoring systems** in content moderation are algorithmic evaluation methods that assign numerical values or risk scores to content based on various factors such as language patterns, context, or user behavior in order to automatically determine which content should be moderated, escalated, or approved.

- False decisions and review practices: proportion of automated moderation, withdrawal rate, appeal procedures, systematic biases (bias detection), specific effects on data subjects.

This disclosure shall be made in accordance with the requirements of the **Digital Services Act (DSA)**, in particular:

- Easily comprehensible information on algorithmic decision-making processes for users (Art. 15 DSA)
- Clear and specific obligations to provide reasons for moderation decisions (Art. 17 DSA).
- Risk assessments and their publication in the transparency report; existing risk-based assessments (e.g., within the framework of the AIA, internal risk analyses, or external audits) should be integrated in a meaningful way and communicated openly (Artt. 34, 42 DSA).
- Information to be provided for automated decisions (Artt. 13-15, 22 GDPR).

Individual case explanations

Structured, understandable, and accessible explanations shall be provided for all relevant individual case decisions, regardless of whether the decision was automated, made by a human, or arrived at through a hybrid process. This applies to:

- Users whose content has been removed, flagged, or deprioritised
- Users whose reports have not led to action
- Changes to a previously made decision

The explanations must be **clear and specific** and comply with the requirements of Art. 17 DSA, Artt. 13-15 GDPR, and Art. 86 AIA.

Decisions that have a collective impact on entire groups or subject areas should be publicly documented in aggregate form[19] and analysed regularly (e.g., regarding the visibility of queer content or political activism).

[19] **Aggregated form** refers to data obtained from raw data through statistical summarization and compression (e.g., averages, totals, frequencies), whereby individual data points are no longer recognizable and privacy is protected.

IMPLEMENTATION

✓ TRANSPARENCY DASHBOARD

Development of a **publicly accessible transparency dashboard** (based on the transparency reports within the meaning of the DSA, e.g., Artt. 15, 24 DSA) that presents core technical logic, fairness metrics, and appeal statistics.

✓ EXPLANATION FORMAT

Establishment of a standardised **explanation format for individual cases**, which will be continuously developed (as defined Art. 17 DSA).

✓ REGULAR EVALUATION

of the explanation format with the involvement of civil society organisations, community representatives, independent scientists, and moderation teams.

✓ TRANSFER OF THE EVALUATION RESULTS

into training programs, model adjustments, and further developments.

LEGAL REFERENCES

Artificial Intelligence Act (AI Act) from 2024

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679#tit_1

Digital Services Act (DSA) from 2022

Regulation (EU) No 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065#tit_1

General Data Protection Regulation (GDPR) from 2016

Regulation (EU) No 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679#tit_1