

Trusted Flaggers under Article 22 DSA

Athena Forum Submission on the Draft Guidelines for Trusted Flaggers

Executive Summary

This submission by Athena Forum responds to the European Commission's draft Guidelines on the implementation of Article 22 of the Digital Services Act (DSA), which establishes the trusted flagger mechanism.

Athena Forum welcomes efforts to provide operational clarity concerning the designation, independence, expertise, accuracy, transparency and oversight of trusted flaggers. However, the draft Guidelines insufficiently address the risks that arise when priority flagging mechanisms operate in areas of intense legal and political contestation, particularly debates concerning sex-based rights, women's rights, the provision of single-sex spaces and services, the safeguarding of children and adolescents, and the rights of lesbians and gay men.

Across Europe, individuals and organisations defending sex-based rights have faced content restrictions, professional sanctions, threats, harassment and legal proceedings. The language necessary to articulate sex-based rights has itself become contested. Statements distinguishing sex from gender and gender identity, referring to a person's sex, or defending single-sex spaces and services repeatedly have been characterised as hateful, harmful, "transphobic" or "extremist", despite forming part of legitimate and lawful political, legal and scientific debate. Equality between women and men is a fundamental value of the European Union, enshrined in the Treaties and reflected in the Union's obligation to eliminate inequalities and promote equality between women and men throughout its activities. The EU equality framework also recognises the legitimacy of specific measures addressing sex-based disadvantage and differences in treatment in the provision of goods and services, including forms of single-sex provision.

At the same time, women expressing such views have themselves faced threats, intimidation and coordinated harassment, including hostile targeting as "TERFs" or "extremists", without equivalent recognition as victims of online abuse. This creates a risk of asymmetric enforcement: lawful advocacy may be over-restricted while aggression against those engaging in it is under-recognised.

The trusted flagger mechanism can amplify these problems if organisations with strong advocacy positions are granted privileged procedural status without sufficient scrutiny of their independence, objectivity, methodologies and track records. Expertise in a field does not in itself guarantee impartial application of the law. Priority processing must not become a route through which contested political positions are transformed into de facto content moderation standards.

Athena Forum therefore recommends stronger safeguards concerning designation and review; scrutiny of applicants' prior conduct; precise legal substantiation of notices; independent assessment by platforms; protection of lawful sex-based expression; meaningful transparency; pluralistic consultation and training; and attention to the cumulative effects of repeated flagging and automated enforcement on freedom of expression.

1. The Digital Services Act and Trusted Flaggers under Article 22

The Digital Services Act establishes a harmonised EU framework for addressing illegal content online while safeguarding fundamental rights, including freedom of expression and information. Article 22 provides for the designation of trusted flaggers whose notices must be given priority and processed without undue delay by providers of online platforms.

The mechanism gives designated entities a position of considerable practical influence. Although trusted flaggers do not determine illegality and platforms retain responsibility for decisions concerning notified content, priority status may shape the speed, visibility and practical consequences of enforcement. Over time, repeated notices may also influence platform moderation practices, internal taxonomies, risk assessments and automated systems.

This makes the quality and integrity of trusted flagger activity essential. Speech is complex, contextual and evolving. Laws concerning hate speech, harassment, threats, defamation, insult and related offences differ between Member States and require careful assessment of language, intent, context, target, severity and applicable law. These questions are particularly difficult where speech forms part of political, legal, scientific or social controversy.

Such laws are ordinarily interpreted through legal processes in which competing arguments and fundamental rights can be weighed. They should not be reduced to automated assumptions or keyword-based classifications. Nor should priority flagging become a mechanism for escalating speech that is controversial, offensive to some audiences or merely alleged to be unlawful.

Trusted flaggers should therefore submit priority notices only where they can identify a sufficiently clear and substantiated legal basis for considering content illegal. Notices should identify the relevant legal provision and provide the contextual information necessary for independent assessment. Platforms must retain responsibility for determining whether notified content is illegal and should not treat trusted flagger status as creating a presumption of illegality.

The same safeguards should inform the wider DSA ecosystem. Digital Services Coordinators (DSCs) play a central role in designation and oversight. Out-of-court dispute settlement bodies may influence the treatment of contested moderation decisions. Very large online platforms develop systemic-risk mitigation strategies with potentially far-reaching effects on public discourse. The European Commission exercises important supervisory functions. Across these structures, the protection of fundamental rights must remain central.

2. Trusted Flaggers and Sex-Based Rights

2.1 Sex-based rights and the language necessary to articulate them

Equality between women and men is a fundamental value of the European Union and a binding obligation. Sex remains a central legal and material category in areas including equality, discrimination, violence, healthcare, data collection, sport, privacy and the provision of certain spaces and services.

In recent years, however, the relationship between sex and gender identity has become one of Europe's most contested political and legal questions. The term "sex" has increasingly been displaced in policies and public communications by "gender", while references to women may be reformulated through language intended to include persons on the basis of gender identity. This has led to serious consequences for the legal clarity and integrity of sex-based protections.

Conflating sex and gender identity obscures the material basis of women's rights and weakens the ability to maintain sex-based distinctions where these are lawful, necessary and proportionate.

These disputes have practical consequences. They concern access to single-sex spaces and services, women's sport, female providers of intimate care, women-only refuges, women-only associations and events, dating services, prisons and other settings in which sex may be relevant to privacy, dignity, safety or equality.

Athena Forum's position is that sex-based rights cannot be effectively defended without the freedom to use clear sex-based language. A woman arguing that a male should not enter a female-only space must be able to describe the material basis of that objection. Requiring speakers to adopt language that presupposes the very proposition they contest makes coherent advocacy impossible.

Sex-based language may conflict with the preferences or political positions of gender-identity advocates. That conflict does not make such language inherently hateful, harassing or illegal.

2.2 The growing contest over sex-based expression

The inclusion of gender identity in national anti-discrimination, hate-speech or related legal frameworks has created difficult questions concerning the boundary between protection from unlawful conduct and the freedom to express sex-based views. Concepts such as "misgendering" are not uniformly defined in law, yet the deliberate use of sex-based language may become the subject of complaints, litigation or demands for sanction.

The current European context illustrates this tension.

- In France, feminist campaigner [Dora Moutot](#) was convicted of insult on grounds of gender identity after stating that women need to be wary of "people with penises".
- In Denmark, psychotherapist [Lotte Ingerslev](#) was taken to court for alleged harassment and defamation after using male pronouns for a male person identifying as a woman while raising concerns about the individual's sexually explicit online content.

- In Germany, the owner of a women-only fitness studio, [Doris Lange](#), became involved in a discrimination dispute after refusing access to a male identifying as a woman.

In the United Kingdom, [Maya Forstater](#)'s successful litigation established that gender-critical beliefs are protected under equality law as beliefs worthy of respect in a democratic society. The UK Supreme Court's 2025 judgment concerning the meaning of "woman" and "sex" under the Equality Act further demonstrated that the legal relationship between sex and gender identity is neither trivial nor settled by activist terminology.

These cases arise under different national laws and should not be collapsed into a single legal category. Taken together, however, they demonstrate that speech concerning sex and gender identity is deeply contested and highly context-dependent. This is precisely why trusted flagger mechanisms require caution.

2.3 Platform moderation and the risk of asymmetric enforcement

Platform rules may go beyond the requirements of national law. Some major platforms have revised their policies concerning contested speech, while TikTok introduced an explicit [prohibition](#) on "misgendering" and "deadnaming" in its Community Guidelines in 2022.

Rules framed in this way may capture lawful statements about sex, advocacy for single-sex spaces and services or criticism of gender-identity policies. A moderation system that treats the use of sex-based language as inherently abusive risks embedding one side of a political dispute into platform governance.

Debates concerning women's rights illustrate the broader asymmetry. Advocacy relating to sex-based rights, single-sex spaces and services, the commodification of women's bodies, prostitution, pornography and equality between women and men may be removed, restricted or made less visible online. Women (and men) articulating sex-based positions, insisting on clarity in language and law, or challenging gender-identity policies may be characterised as hateful, "transphobic" or "extremist".

At the same time, these women (and men) may themselves face threats, intimidation and coordinated harassment, including hostile targeting with labels such as "TERF", without being recognised or protected as victims of online abuse.

Women have faced cancellation, professional consequences, public vilification and physical threats for expressing concerns about sex-based rights, women's rights, the provision of single-sex spaces and services, the safeguarding of children and adolescents, and the rights of lesbians and gay men. This can silence not only prominent campaigners but ordinary women who are concerned about changes affecting their rights, services or families.

The DSA framework should therefore guard against both under-enforcement and over-enforcement. Women must be protected from threats, incitement, harassment and other illegal content. At the same time, lawful sex-based expression must not be treated as illegal merely because it conflicts with gender-identity claims.

3. Key Concerns Regarding the Draft Guidelines

3.1 Independence, objectivity and advocacy conflicts

The draft Guidelines should more clearly address the relationship between subject-matter expertise and advocacy.

Civil society organisations frequently combine expertise with campaigning. This does not automatically disqualify them from trusted flagger status. However, expertise does not itself guarantee independence or objectivity. Where an applicant has taken a sustained position in a contested political dispute, DSCs should examine whether its methodologies and prior conduct demonstrate the capacity to distinguish unlawful content from lawful opposition to its own advocacy objectives.

Relevant scrutiny should include whether an applicant has systematically characterised sex-based advocacy as hateful or extremist; treated the use of sex-based language as inherently abusive; promoted gender-identity policies that conflict with sex-based rights; or pursued action against women (and men) expressing lawful sex-based views through police reporting, assistance in legal proceedings, media campaigns or policy advocacy seeking restrictions on such expression.

The same scrutiny should apply to whether organisations claiming to support victims of online hatred have done so even-handedly, including where women defending sex-based rights are targeted by gender-identity activists. A demonstrably one-sided record should be relevant to designation, periodic review and, where appropriate, revocation of trusted flagger status.

This concern is not hypothetical. [HateAid](#), a designated trusted flagger in Germany, publicly presents so-called “TERFs” as a “special case of anti-queer hostility” and states that their arguments, “under the guise of feminism”, purportedly legitimise “hate and incitement” against trans people. It further characterises gender-critical positions as an “ideologised worldview” and presents the claim that there are more than two sexes as a “biological fact”. These statements concern precisely the contested questions of sex and gender identity on which women’s rights advocates, researchers and others hold opposing views.

Sonderfall der Queerfeindlichkeit: TERFS

TERFs, **trans-exclusionary radical feminists** (trans-ausschließende radikale Feminist*innen), sind sehr präsent auf Social-Media-Plattformen. TERFs verstricken **unterschiedlichste ideologische Argumente**, die unter dem Deckmantel des Feminismus **Hass und Hetze gegen trans Personen** vermeintlich legitimieren. Die Personen hinter dieser Ideologie **sprechen** trans Personen nicht nur ihre **Geschlechtlichkeit ab**, sondern gehen davon aus, dass die **Kategorie Frau durch die queere Community aufgelöst** werden soll.

TERFs glauben an ein **binäres Geschlechtermodell**, das sich anhand der Genitalien ablesen lässt. Sie nehmen an, dass ein Mensch, der mit einem Penis geboren wird, **von Geburt an patriarchale Unterdrückung** in sich trägt. Deswegen sehen sie in jeder Person, die einen Penis hat oder mit einem geboren wurde, eine Gefahr für cis Frauen. TERFs bezeichnen sich selbst als „gender-kritisch“. Sie sprechen inter Personen einfach ihre Existenz ab. TERFs machen andere **Lebenswege, Einstellungen, Identitäten und biologische Tatsachen**, wie die, dass es **mehr als zwei Geschlechter gibt, unsichtbar** und **lehnen sie kategorisch ab**, um ihr **ideologisiertes Weltbild** abzusichern.

Doch, und dafür steht **Feminismus eigentlich**, auf dem Weg für **Gleichberechtigung und Antidiskriminierung** geht es darum, **strukturelle Barrieren aufzuzeigen und aufzubrechen**. Das schließt **alle marginalisierten Gruppen** ein. Bei der Diskussion um Menschenrechte, sowohl im **Analogen** als auch im **Digitalen**, geht es darum, wie wir **gemeinsam für uns alle** eine **sichere und gerechte Welt** schaffen können.

Das **erkennen TERFS nicht** und **erschweren** den Weg zu Gleichberechtigung, Toleranz und Selbstbestimmung. Alle Mitglieder der Gesellschaft müssen **gemeinsam**, in **Allyship**, dafür sorgen, dass der analoge sowie der digitale Raum **keinen Platz** für **Rassismus, Sexismus**, Queerfeindlichkeit, **Ableismus** und andere Formen von Diskriminierung bieten.

Source: HateAid, “[LGBTQIA+: Queerfeindlichkeit und Empowerment im Netz](#)”, section “Sonderfall der Queerfeindlichkeit: TERFS”. Screenshot accessed 10 July 2026.

3.2 Contested concepts must not be treated as settled law

Trusted flaggers should not convert contested political concepts into presumptions of illegality. Terms such as “misgendering”, “deadnaming”, “transphobia” and “hate” may be used differently across advocacy, platform policy and national law.

The Guidelines should require a clear distinction between these categories. A breach of a platform’s terms and conditions is not necessarily illegal content under the DSA. Content described by an advocacy organisation as hateful is not necessarily unlawful. Disagreement with gender-identity claims is not, by itself, evidence of hatred or harassment.

This distinction is particularly important where sex-based and transactivist language directly conflict. The trusted flagger mechanism should not privilege one vocabulary by treating the other as presumptively suspect.

For example, the following definitions of ‘hateful conduct/content’ are frequently applied by gender identity advocacy groups as a proof of violations of applicable laws on hate crimes and hate speech.

Activist definition	Content
Misgendering	Referring to a person according to their biological sex rather than their self-reported gender identity.
Deadnaming	Referring to a trans-identified person by a previous name rather than their current chosen name.
Denial of gender identity	Referring to a person according to their biological sex or declining to affirm through speech their self-declared gender identity.
Transphobic misinformation	A label that gender identity activists frequently apply to statements describing biological sex as immutable or discussing the medical aspects or consequences of “gender transition”.
Anti-trans hate speech	A label that gender identity activists frequently apply to statements affirming biological sex or advocating for sex-based rights for women.
Incitement to hatred based on gender identity	A label that gender identity activists frequently apply to statements affirming sex-based rights and lawful sex-based exemptions on the basis that such advocacy encourages discrimination against people who identify as trans.

3.3 Accuracy and evidentiary standards

Priority status should carry a correspondingly high standard of care.

Trusted flaggers should identify the specific legal basis relied upon and provide sufficient factual and contextual reasoning. Article 22 should not be used merely to draw enhanced platform attention to content whose illegality is uncertain.

Where content forms part of a wider pattern, relevant context may properly be supplied. But contextual analysis must operate in both directions. A single statement should not be stripped of the political, scientific or legal debate in which it occurs to portray it as hateful. Equally, campaigns of threats or harassment should not escape scrutiny merely because each item is viewed separately.

Platforms must continue to conduct an independent assessment and should not treat a notice from a trusted flagger as determinative.

3.4 Inconsistent national laws and cross-border effects

National laws differ significantly. Content may be alleged to be illegal in one Member State but lawful in another. Cross-border platforms, trusted flaggers and DSCs must avoid turning the most restrictive national interpretation into a de facto EU-wide speech standard.

This is particularly important because equality between women and men is a fundamental EU value. The EU equality framework also recognises specific measures addressing sex-based disadvantage and, under defined conditions, differences in treatment in access to goods and services. These commitments should remain a reference point when the Guidelines are implemented.

Restrictive national approaches to speech critical of gender-identity claims should not be allowed, through platform systems or trusted flagger practice, to undermine lawful advocacy for sex-based rights, including single-sex spaces and services, across the Union.

3.5 Transparency and accountability

Transparency obligations should allow meaningful scrutiny of trusted flagger activity in contested areas.

Trusted flaggers should publish sufficiently detailed information on the legal bases invoked, categories of content flagged and outcomes of notices, subject to privacy and data-protection requirements. Attention should be paid to patterns involving sex-based language, disputes concerning sex and gender identity, and allegations of hate speech or harassment arising from political advocacy.

DSCs should examine not only notice volumes and formal accuracy rates, but patterns of over-flagging, selective flagging and under-recognition of illegal content directed at disfavoured groups.

3.6 Training and pluralistic consultation

Training on independence, objectivity, legality, EU values and obligations and fundamental rights should include a plurality of perspectives, and should remain grounded in existing EU legal frameworks, including the scope of EU competences under Article 19 TFEU and the guarantees of the EU Charter of Fundamental Rights. It should not present contested policy concepts or advocacy positions as settled requirements of EU law.

DSCs, trusted flaggers, platforms and other relevant DSA bodies should consult women's rights and sex-based rights organisations when developing methodologies, guidance and training concerning disputes over sex and gender identity.

This is not a request to replace one ideological position with another. It is a safeguard against closed systems in which actors sharing the same assumptions train, designate and validate one another.

3.7 Automation and systemic risks

A further concern is the interaction between trusted flagger activity and automated moderation.

Repeated notices may contribute to scalable enforcement patterns. Contested assumptions embedded in notices, taxonomies or moderation guidance may migrate into automated systems that are unable to assess context adequately. A disputed interpretation can thereby move from an individual notice to a platform-wide rule affecting large numbers of users.

Systemic risk assessments concerning freedom of expression should therefore explicitly consider the risk of restricting lawful sex-based expression. Platforms should examine whether their mitigation strategies disproportionately restrict lawful advocacy concerning sex-based rights.

The European Commission should monitor these effects. Moderation policies that categorically suppress lawful sex-based expression should be scrutinised for compatibility with the DSA's fundamental-rights framework.

4. Recommendations

To ensure that the trusted flagger mechanism achieves its objectives without reinforcing viewpoint discrimination or restricting lawful expression, Athena Forum recommends the following:

4.1 Strengthen designation and review

Digital Services Coordinators should assess not only formal expertise but applicants' independence, objectivity, methodologies and prior conduct. This should include scrutiny of whether an applicant has:

- systematically treated lawful sex-based expression as hateful or illegal;
- supported action against individuals for lawful advocacy concerning sex-based rights;
- demonstrated one-sidedness in identifying victims of online hatred and harassment;
- conflated breaches of platform policy with illegality;
- treated contested political concepts as settled legal standards.

Such evidence should be relevant to designation, periodic review and possible revocation.

4.2 Require precise legal substantiation

Trusted flagger notices should:

- identify the specific legal provision allegedly breached;
- explain the applicable jurisdiction;
- provide sufficient contextual reasoning;

- distinguish illegality from offensiveness, harmfulness or disagreement;
- avoid using Article 22 merely to escalate uncertain cases for enhanced scrutiny.

4.3 Protect lawful sex-based expression

The Guidelines should explicitly recognise the risk that lawful expression concerning sex-based rights, the safeguarding of children and adolescents, and the rights of lesbians and gay men may be misclassified as hateful or illegal.

4.4 Ensure independent platform assessment

Priority processing must not create a presumption of illegality. Platforms should retain and exercise independent responsibility for assessing notices.

Where platform policies themselves categorically prohibit forms of lawful sex-based expression, these policies and their systemic effects should be scrutinised under the DSA's fundamental-rights framework.

4.5 Strengthen transparency

Trusted flaggers should provide meaningful reporting on:

- legal bases invoked;
- categories of content flagged;
- outcomes of notices;
- patterns of disputed or unsuccessful notices;
- relevant areas of contested speech.

DSCs should monitor patterns of both over-flagging and under-enforcement.

4.6 Ensure pluralistic training and consultation

Training for DSCs, trusted flaggers and platform moderators should include:

- the EU's foundational values and Treaty obligations concerning equality between women and men which requires the Union in all its activities to aim to eliminate inequalities and promote equality between women and men;
- freedom of expression and other fundamental rights;
- the distinction between illegal content and lawful but contested speech;
- the relevance of sex-based rights across equality law, policy and public debate;

- the legal and political contestation surrounding sex and gender identity;
- risks of viewpoint discrimination and advocacy capture;
- contextual assessment and the limitations of automated moderation.

Women's rights and sex-based rights organisations should be included in relevant consultation and training processes at national and European level.

4.7 Address systemic and automated effects

The European Commission, DSCs and platforms should examine whether repeated flagging practices, moderation policies or automated systems disproportionately suppress lawful sex-based expression. Systemic-risk assessment and mitigation should consider:

- repeated erroneous flagging;
- visibility restrictions;
- chilling effects on lawful participation;
- migration of contested assumptions into automated systems;
- failure to recognise threats and harassment directed at those defending sex-based rights.

4.8 Ensure coherence across the wider DSA ecosystem

The same fundamental-rights safeguards should inform the work of trusted flaggers, DSCs, out-of-court dispute settlement bodies, platforms and the European Commission.

Out-of-court dispute settlement bodies should not treat sex-based language as inherently violative of platform rules and should identify cases in which moderation decisions raise broader freedom-of-expression concerns.

Very large online platforms should assess whether their moderation and mitigation strategies protect lawful expression concerning sex-based rights while addressing illegal threats and harassment directed at women (and men).

Member States should ensure that the implementation of national laws affecting speech does not erase the distinction between gender identity and sex or prevent advocacy for sex-based rights. Relevant authorities and courts should be equipped to assess these disputes with due regard to freedom of expression, equality between women and men and the legitimate role of sex-based language in public debate.

Conclusion

The trusted flagger mechanism is intended to improve the identification and processing of illegal content. Its legitimacy depends on maintaining a clear boundary between illegality and political or personal disagreement.

Debates concerning sex and gender identity are among the most contested in contemporary Europe. They engage fundamental questions of equality between women and men, women's rights, privacy, dignity, safeguarding, language and freedom of expression. The DSA should not resolve these debates by granting privileged procedural influence to one side or by allowing contested advocacy concepts to become de facto standards of illegality.

Athena Forum therefore calls for a trusted flagger framework grounded in legal precision, independence, objectivity, transparency and fundamental rights. Women must be protected from threats, harassment and violence. They must also remain free to articulate sex-based rights in clear language, to defend lawful single-sex spaces and services, and to participate fully in democratic debate.

Athena Forum is a European advocacy group working to safeguard and advance sex-based rights across law, policy and civil society, and confront their misrepresentation in public discourse. Our work engages policymakers, journalists, researchers, educators, legal and medical professionals, academics, civil society actors and feminist advocates to ensure that sex-based rights remain central to democratic policy and legal frameworks. We work across advocacy, legal and policy analysis, public and media engagement and institutional monitoring, building a resourceful and impactful non-partisan platform across Europe.