

Vincent Conversation Export

Exported on October 29, 2026

Matter Code: N/A

Conversation Messages

 **User** (2026-10-29 08:55:34)

Would requiring compulsory identification on bots on social media be justifiable under Human Rights legislation given that a bot is an autonomous program not an human individual to whom Human Rights legislation necessarily attaches?

 **Vincent** (2026-10-29 08:55:35)

Yes—a requirement that social-media bots identify themselves as automated accounts would, in my assessment, be capable of justification under the Human Rights Act 1998. But the fact that the bot itself has no human rights is not a complete answer: the requirement may affect the rights of people or organisations using it.

The important distinction is between **identifying an account as a bot** and **identifying the person behind it**.

1. Whose rights are engaged?

Your starting point is sound: an autonomous computer program is not, merely by being autonomous, a human-rights holder. However, it does not follow that regulation of its communications falls outside human-rights scrutiny.

A bot may be the means by which a person or organisation communicates. Article 10 protects receiving and imparting information and ideas; its protection is not confined to communications

personally typed and posted by a human being. The relevant question would therefore be whether regulating that means of communication interferes with the operator's expression or the audience's receipt of information—not whether the software possesses rights. See [Human Rights Act 1998, Schedule 1, Article 10\(1\)](#).

There is also a qualification to the suggestion that rights necessarily attach only to human individuals: **organisations can enjoy applicable Convention rights**. In *Animal Defenders International*, the claimant was a non-profit company, and its Article 10 rights were accepted as engaged by restrictions on its advertising: [R \(Animal Defenders International\) v Secretary of State for Culture, Media and Sport](#) [2008] UKHL 15, [2], [26].

That does not confer rights on the bot. It identifies the potential rights-holder behind its use.

2. Compulsory labelling as a bot

Assuming a legal requirement imposed by the state, a rule requiring an account to display “**automated account**”, or an equivalent accurate description, would have a strong prospect of being justifiable.

Even if Article 10 is engaged, it expressly permits appropriate formalities and conditions which are prescribed by law and necessary in a democratic society for specified purposes, including protecting the rights of others and preventing disorder or crime: [Human Rights Act 1998, Schedule 1, Article 10\(2\)](#).

The strongest justification would be that the rule:

- enables users to distinguish human participation from automated activity;
- addresses impersonation, fraud and manufactured appearances of public support;
- preserves the ability to communicate the underlying message rather than prohibiting it; and
- applies irrespective of the political or other viewpoint expressed.

Protecting democratic debate against distortion can justify restrictions on the means of communication. In *Animal Defenders International*, the House of Lords upheld restrictions on political broadcast advertising to protect democratic rights against the distorting effects of unequal resources. **That is an analogy, not a decision about bots**, but it supports the

legitimacy of the underlying objective: [R \(Animal Defenders International\) v Secretary of State for Culture, Media and Sport](#) [2008] UKHL 15, [2], [26].

The justification would still depend on the scheme's design: a clear definition of automation, accurate and neutral labels, proportionate sanctions, and a means of challenging mistaken classification would strengthen it. A rule which indiscriminately treated ordinary scheduling or accessibility assistance as deceptive bot activity would be harder to defend.

3. Compulsory disclosure of the operator's identity

Requiring the operator's **real name to be made public** is materially more intrusive than requiring an automation label. It may engage privacy under Article 8 as well as expression under Article 10. Both rights permit justified restrictions, but public identification would require its own necessity and proportionality assessment: [Human Rights Act 1998, Schedule 1, Articles 8 and 10](#).

Anonymous expression is not excluded from protection. The Court of Appeal expressly rejected the proposition that Article 10 protects named but not anonymous speakers, recognising that protection against disclosure of identity may be legitimate: [Totalise plc v Motley Fool Ltd \[2001\]](#) EWCA Civ 1897, [25]–[26].

Accordingly, three requirements should be assessed separately:

- **Public automation label:** comparatively limited interference.
- **Confidential verification of the operator by the platform:** more intrusive, but potentially defensible with safeguards.
- **Public disclosure of the operator's identity:** substantially greater interference, particularly where it exposes lawful speakers to retaliation.

If an automation label adequately addresses the deception, that would weigh against the necessity of additionally publishing the operator's identity.

Conclusion

The better argument is not simply “bots have no human rights, so compulsory identification is permissible”. It is that a narrowly framed automation-labelling

requirement can protect human users against deception while imposing only a limited, proportionate condition on the expression of those deploying bots.

The bot's lack of rights removes a claim by the program itself; it does not remove scrutiny of the consequences for human or organisational rights-holders. I have not identified a directly on-point UK judgment deciding compulsory social-media bot labelling, so the conclusion is an application of the established framework rather than a settled ruling on that precise measure.

I can develop this into a short research memorandum if useful.

 **User** (2026-10-29 09:01:45)

Could a requirement to identify an underlying political or commercial interest on the part of bot's operator be justified?

To what extent could a Human Rights compliant requirement for certain such operators to identify themselves be justified?

 **Vincent** (2026-10-29 09:01:46)

Yes. A requirement to disclose the political or commercial interest behind a bot could be justified; a targeted requirement to identify the responsible operator could also be compatible with human rights. The strongest case is for identifying **who is paying for, directing or commissioning organised influence**, rather than requiring everyone expressing political opinions through automation to reveal their identity.

The distinction is between **transparency about the source of influence** and **compulsory exposure of the individual speaker**.

1. Disclosure of the underlying interest

A requirement to disclose an objectively ascertainable relationship—payment, sponsorship, employment, commissioning or control—is considerably easier to justify than a requirement to disclose someone’s private political beliefs.

The justification would be that audiences should be able to evaluate communications knowing whether they are encountering:

- independent expression;
- commercial promotion;
- communications commissioned by a political party or campaign;
- apparently independent accounts controlled by the same organisation.

Such disclosure can leave the message entirely available while correcting a misleading impression about its provenance. That makes it a comparatively limited interference with expression.

However, **“underlying political or commercial interest” would need careful definition.** A duty to disclose who commissioned a campaign is reasonably concrete. A duty to disclose every financial interest, political affiliation or motive which might influence an operator would be much broader, uncertain and potentially chilling.

Existing statutory precedents

There are already close precedents for both political and commercial transparency:

Political communications. The digital-imprint regime requires specified electronic material to identify the promoter and any other person on whose behalf it is published, including names and addresses. It covers defined paid-for political advertising and certain unpaid material published by specified political actors. These provisions came into force on 1 November 2023: [Elections Act 2022, ss 41–45](#).

That is particularly significant: **the law already requires public identification behind some online political communications**, without treating anonymity as an absolute entitlement. Its application turns on the statutory conditions, not whether a bot or human publishes the

material. It does **not**, however, cover all political discussion or automatically encompass every bot campaign.

Commercial communications. Consumer legislation addresses concealed commercial intent and disguised promotion. It identifies failure to disclose commercial intent, unless apparent from context, as a misleading omission; it also prohibits specified practices including undisclosed paid editorial promotion and falsely presenting a trader as a consumer: [Digital Markets, Competition and Consumers Act 2024, s 227\(1\)\(c\), Schedule 20, paras 12 and 25](#).

These are strong legislative analogies, **not conclusive judicial rulings that any proposed bot-disclosure regime would be Convention-compliant**.

2. When could operators themselves be required to identify themselves?

A targeted identification duty could extend beyond an automation label where knowing the responsible person or organisation materially advances the legitimate objective.

The following are my assessments of how the established human-rights principles would apply—not descriptions of an existing general bot-identification duty.

Strongest cases

Public identification would be most readily defensible for:

- **Businesses using bots to market products**, particularly where accounts appear to be independent consumers or reviewers.
- **Political parties, candidates and campaign organisations** using bots for organised campaigning.
- **Paid professional operators** running coordinated political or commercial influence campaigns for clients.
- **Persons commissioning or controlling networks of apparently independent accounts**, where disclosure is necessary to expose their common source.

In the last two categories, identification of the technical operator alone may be insufficient. The material information may be **the client or controlling organisation on whose behalf the**

activity occurs. Conversely, naming every programmer, employee or contractor would ordinarily contribute little to that objective.

Payment is a strong basis for drawing the line, but not the only possible one. Organisational control, coordination, reach and deliberate simulation of independent public support could also support a carefully defined duty.

Weaker cases

A blanket public-identification duty would be harder to justify for:

- individuals using automation to distribute their own lawful political views;
- small voluntary groups using scheduling or accessibility tools;
- anonymous whistleblowers or dissidents;
- users facing a credible risk of harassment or retaliation.

Political expression does not lose protection because it is automated or anonymous. The Court of Appeal has expressly recognised that Article 10 protects anonymous as well as named speakers and that protecting identity against disclosure may be legitimate: [Totalise plc v Motley Fool Ltd](#) [2001] EWCA Civ 1897, [25]–[26].

Nor does every person who benefits from a communication become its sponsor. Supporting a party, company or policy is not, without more, evidence of acting on its behalf.

3. The human-rights test

A state-imposed disclosure regime engaging expression or privacy must satisfy the relevant requirements of Articles 10 and 8: a proper legal basis, an applicable legitimate aim, and necessity and proportionality. Protecting consumers, preventing crime and protecting others' democratic rights provide potentially substantial justifications: [Human Rights Act 1998, Schedule 1, Articles 8\(2\) and 10\(2\)](#).

The proportionality inquiry asks whether the objective is sufficiently important, the measure is rationally connected to it, a less intrusive measure could achieve it without unacceptable compromise, and a fair balance is struck: [Bank Mellat v HM Treasury \(No 2\)](#) [2013] UKSC 39.

Applied here, **public identification must do something which a bot label or confidential verification would not adequately accomplish**. For example:

- Confidential verification may enable enforcement against an operator.
- Public identification of the sponsor enables the audience to assess the source and independence of the message.
- Publishing an individual's residential address may add little to either objective while creating substantial risk.

Protection against distortion of democratic debate is a legitimate concern, not merely an objection to unwelcome speech. *Animal Defenders International* supports that proposition in the different context of political broadcast advertising. It also emphasises the high standard of justification required where political expression is restricted: [R \(Animal Defenders International\) v Secretary of State for Culture, Media and Sport](#) [2008] UKHL 15, [26]–[36].

4. Features of a defensible regime

In my assessment, the strongest scheme would:

- **Define covered activity objectively:** for example, commissioned commercial promotion or coordinated electoral influence—not simply “political content”.
- **Identify the responsible principal:** distinguish the account operator, commissioner, sponsor and person exercising control.
- **Limit public disclosure:** ordinarily to the responsible organisation or professional operator and an appropriate contact address, rather than unnecessary personal details.
- **Use graduated obligations:** public sponsor disclosure where needed, with confidential verification of additional identifying information.
- **Protect legitimate anonymity:** provide a safeguarded route for cases involving credible threats or other compelling grounds.
- **Provide fair procedures:** notice, an opportunity to contest classification, independent review and proportionate penalties.
- **Apply neutrally:** the same criteria should govern competing political viewpoints and commercial interests.

These are proposed design safeguards, not a statement that each is invariably required in precisely that form.

Conclusion

A substantial, targeted identification requirement could be justified—not merely a requirement that an account say it is a bot. In particular, requiring businesses, political campaigns and professional influence operators to identify the organisation responsible and the person on whose behalf they act has a strong justification.

The defensible dividing line is **accountability for organised influence, without unnecessarily abolishing anonymous expression**. A requirement to expose objectively verifiable sponsorship or control is much stronger than a general demand that everyone disclose their identity, political beliefs or supposed motives.

 **User** (2026-10-29 09:04:48)

To what extent could requiring the geographical origin of a bot's operator be justified?

 **Vincent** (2026-10-29 09:04:49)

Requiring disclosure of the country from which a bot operation is directed could be justified, particularly for organised political campaigning or commercial promotion. Requiring an individual operator's precise location, home address or birthplace would be substantially harder to justify.

The strongest justification is **transparency about where an influence operation is controlled**, not exposure of the operator's personal geographical background.

1. What does “geographical origin” mean?

A defensible requirement would need to distinguish:

- **Country of operational control:** where the person or organisation directing the bot campaign operates.
- **Country of the commissioning or sponsoring organisation:** potentially different from the technical operator's location.
- **Country of business establishment:** relevant to commercial accountability.

- **Server or connection location:** a technical fact which does not necessarily establish who controls the communication.
- **The individual's nationality, birthplace or ancestry:** personal characteristics which do not establish where an operation originates.

These are not interchangeable. A domestic campaign may employ an overseas technical contractor; an overseas sponsor may employ a domestic operator. Disclosure of only the contractor's location could therefore obscure, rather than reveal, the source of influence.

2. Country-level disclosure: a reasonably strong case

In my assessment, a country-level disclosure requirement would have a strong prospect of justification where it addresses a demonstrable risk that audiences are being misled about the source of organised communications—for example:

- an overseas-directed network presenting itself as independent local voters;
- a commercial bot operation presenting itself as domestic consumers;
- a coordinated campaign concealing that its accounts share a controlling organisation abroad.

The relevant objectives could include protecting democratic rights, preventing fraud and protecting consumers. Articles 8 and 10 permit necessary, legally prescribed interferences for their specified legitimate purposes. But **“transparency” must be connected to such a purpose; it is not an unlimited justification in itself:** [Human Rights Act 1998, Schedule 1, Articles 8\(2\) and 10\(2\)](#).

There is also a relevant qualification: **Article 10 protects communication regardless of frontiers.** Overseas origin is not, by itself, evidence of wrongdoing or a sufficient reason to restrict lawful expression: [Human Rights Act 1998, Schedule 1, Article 10\(1\)](#).

Accordingly, a neutral factual disclosure of the country of control is more defensible than a label implying that all foreign-origin expression is disinformation.

3. How much geographical detail could be required?

The following is an application of proportionality principles, rather than an existing general rule governing bots.

Required disclosure	Likely justification
Country of establishment of a commercial operator or sponsor	Strong where it assists consumer accountability and accurate identification of the business.
Country from which an organised political bot campaign is directed	Stronger where the campaign purports to represent independent domestic participation.
Region or city	Requires an additional explanation of why country-level disclosure is insufficient—for example, a materially misleading claim of local presence.
Business address	Potentially justifiable for accountable commercial or campaigning entities, depending on the activity and applicable regime.
Individual's home address	Much harder to justify publicly where a business or service address, or confidential verification, would suffice.
Precise or live physical location	Ordinarily excessive for a public provenance label; a distinct investigative justification would be needed for compelled confidential disclosure.
Birthplace, ancestry or nationality	Generally a poor substitute for operational provenance and difficult to justify merely to reveal where a campaign is directed.

There are existing legislative analogies. Qualifying information-society service providers must make their name and geographical establishment address accessible; commercial

communications must identify the person on whose behalf they are made: [Electronic Commerce \(EC Directive\) Regulations 2002, regs 6\(1\)\(a\)–\(b\) and 7\(b\)](#). That does **not** mean every bot operator falls within those provisions.

Similarly, specified digital political material must identify the promoter and person on whose behalf it is published, including their addresses: [Elections Act 2022, s 41\(3\), read with ss 42–45](#). Neither statutory example conclusively establishes the compatibility of a new geographical-disclosure scheme, but both show that geographical accountability is not inherently inconsistent with protected expression.

4. Public disclosure and confidential verification are different

The state would need to justify separately what the platform collects and what the public sees.

For example, a scheme might require a covered operator to provide verified identifying information confidentially while publishing only the country of operational control and the responsible organisation. That could achieve much of the transparency objective without exposing an individual's precise whereabouts.

Nevertheless, even a country label can create risks for a dissident, whistleblower or member of a small, identifiable community. Anonymous expression remains protected; the Court of Appeal has recognised legitimate reasons for protecting an internet speaker's identity: [Totalise plc v Motley Fool Ltd](#) [2001] EWCA Civ 1897, [25]–[26].

Verification methods also matter. A requirement to declare a principal country of operation is not equivalent in intrusiveness to continuous location tracking. The assessment must include the means used to establish location, not merely the final label.

5. What would make the requirement proportionate?

The central question is whether the disclosure materially advances an important objective, whether a less intrusive approach would suffice, and whether its benefits outweigh its consequences: [Bank Mellat v HM Treasury \(No 2\)](#) [2013] UKSC 39.

For a proposed scheme, the strongest safeguards would be:

- **An objective definition of origin**, centred on direction, control or establishment—not ethnicity or assumed allegiance.
- **Targeting organised influence activity**, rather than every individual using automated posting.
- **Accurate attribution**, distinguishing the operator, sponsor and technical infrastructure.
- **No unsupported inference of foreign-state control** merely from an overseas location.
- **Limited public granularity**, with more detailed information disclosed confidentially only where justified.
- **Correction and appeal mechanisms**, including where a technical location estimate is unreliable.
- **Protection for credible safety risks**, with proportionate sanctions and safeguards against evasion.

Conclusion

A targeted duty to disclose the country of control or establishment of political and commercial bot operations is capable of strong human-rights justification. It may be less intrusive than naming an individual while still revealing materially important provenance.

The justification weakens as the requirement moves from **where the operation is directed** towards **where the individual lives, where they were born, or where they are at any particular moment**. The appropriate principle is to disclose enough geography to expose the source of organised influence—not more personal location information than that purpose requires.

I have not identified a UK judgment directly determining the compatibility of a bot-specific geographical-origin requirement; this assessment applies the established human-rights framework and the statutory analogies above.

 **User** (2026-10-29 09:07:26)

Putting aside bots, could a requirement to disclose the geographic origin of a post that purported to come from one country but was in fact created elsewhere be justified?

◆ Vincent (2026-10-29 09:07:27)

Yes. A targeted requirement to correct a materially misleading claim about a post's geographical origin would generally be easier to justify than a requirement that all posts disclose their origin. That conclusion does not depend on whether a bot or a human created the post.

However, **requiring correction of the false impression and requiring disclosure of the actual country are separate steps.** The latter needs justification if the former would adequately address the harm.

1. The justification is correcting misleading provenance

The strongest case would concern a post which claims, expressly or through its presentation, to be a contribution from within a particular country **and derives credibility or influence from that claim.** For example:

- purported first-hand reporting by someone claiming to be present;
- a coordinated campaign presenting overseas-created posts as independent contributions by domestic residents;
- commercial promotion falsely presented as the experience of local consumers.

The legitimate objective would not simply be to identify foreign speech. It would be to protect audiences against a materially misleading representation, connected to an applicable aim such as protecting others' rights, preventing fraud or protecting national security.

Article 10 permits proportionate conditions on expression for its specified purposes, but also protects receiving and imparting information **regardless of frontiers.** Consequently, overseas creation alone does not justify adverse treatment: [Human Rights Act 1998, Schedule 1, Article 10\(1\)–\(2\).](#)

The crucial distinction is between “this was created abroad” and “this falsely claims domestic provenance in a way that matters”.

2. What exactly is the post representing?

A requirement would need to distinguish different claims:

- **“I am posting from Country A.”** Creation or posting elsewhere may contradict that representation.
- **“I am a resident/citizen of Country A.”** Posting while abroad does not necessarily contradict it.
- **“This event occurred in Country A.”** Editing or uploading the account elsewhere does not establish that the account is false.
- **“These are independent views from people in Country A.”** Evidence that an overseas organisation fabricated those contributions may directly undermine that representation.

Similarly, the place where text was drafted, where it was uploaded, where its author lives, and where its publication was directed may differ. **A technically accurate location label could itself mislead if it purports to establish more than the evidence supports.** A server or network connection location should not automatically be treated as the author’s physical location or the campaign’s source of control.

3. Correction need not always reveal the actual country

Applying the established proportionality test, the question would be whether a less intrusive measure could achieve the objective without unacceptably compromising it, and whether the resulting balance is fair: [Bank Mellat v HM Treasury \(No 2\)](#) [2013] UKSC 39.

That suggests a graduated approach:

- **Removal or correction of a false location claim** may suffice where the harm is simply that claim.
- **Disclosure that the post was created outside the country it purports to originate from** may suffice where domestic versus overseas provenance is what matters.
- **Identification of the actual country** has a stronger justification where that country is materially relevant—for example, to understanding the source of an organised influence campaign.
- **Disclosure of a precise location or the author’s identity** ordinarily requires substantially more justification.

For example, exposing a fabricated claim to be an eyewitness may require establishing that the author was not present. It does not necessarily require telling the public exactly where the author was.

4. Commercial and political posts

Commercial posts provide a particularly strong case. Consumer legislation already addresses false or misleading information about a trader, product or other matter relevant to a transactional decision, including misleading overall presentation. Its application depends on the statutory conditions, including the likely effect on the average consumer's transactional decision: [Digital Markets, Competition and Consumers Act 2024, ss 225\(4\)\(a\) and 226\(1\)–\(3\)](#). A materially false claim of local provenance could fall within that framework; overseas creation by itself would not.

Political posts require greater care, but are not immune from provenance requirements.

Protecting democratic debate against distortion can justify regulation of communications, although political expression requires a high standard of justification: [R \(Animal Defenders International\) v Secretary of State for Culture, Media and Sport](#) [2008] UKHL 15, [26]–[36]. That case concerns broadcast advertising, not geographical labels, so the application here is by analogy.

A narrowly framed factual correction is more defensible than giving an authority broad power to label political speech “foreign” or “deceptive” merely because it disputes the speaker's perspective.

5. Necessary safeguards

In my assessment, a defensible statutory scheme would require:

- **A demonstrable geographical representation**, not an assumption based merely on language, subject matter or use of a national flag.
- **Materiality**: the false provenance must matter to the audience's understanding or to the harm being addressed.
- **Reliable evidence**, with technical uncertainty acknowledged.
- **An opportunity to contest and correct the finding.**
- **Protection for legitimate anonymity and safety**, especially for dissidents and journalists.
- **Proportionate consequences**, distinguishing mistakes from deliberate or repeated deception.

Anonymous expression remains protected, and there may be legitimate reasons not to disclose an internet speaker's identity: [Totalise plc v Motley Fool Ltd \[2001\] EWCA Civ 1897, \[25\]–\[26\]](#). A requirement ostensibly concerned only with country-level origin must therefore be assessed for its actual consequences, including whether it effectively identifies or endangers the speaker.

Conclusion

Yes: correcting a materially false claim of domestic provenance is a strong potential justification for a targeted disclosure requirement. It addresses a representation the publisher has chosen to make, rather than demanding geographical disclosure from everyone.

But **a power to require correction does not automatically justify compelling disclosure of the actual country.** The appropriate extent is the minimum geographical information reasonably needed to correct the deception and address the identified harm. I have not identified a UK judgment deciding this precise form of requirement; that is an application of the established Convention framework.

 **User** (2026-10-29 09:10:37)

Could requiring social media platforms to ensure compliance with any such legislation and imposing financial penalties for breaches and or enforcing suchj requirements by remedies such as injunction be justified?

 **Vincent** (2026-10-29 09:10:39)

Yes. Parliament could in principle impose enforceable duties on social-media platforms to secure compliance with a properly framed provenance-disclosure law, backed by financial penalties and injunctions. Human-rights compatibility would depend on both the underlying disclosure requirement **and the way platform responsibility and enforcement were designed.**

The central distinction is between requiring a platform to take **effective, proportionate compliance measures** and making it an **absolute guarantor of the geographical truth of every post**. The former is much easier to justify.

1. Platform responsibility can be justified

The justification would be that the platform controls the facilities through which the misleading representation reaches its audience and is often better placed than individual readers to obtain declarations, display corrections and respond to substantiated complaints.

There is already a legislative model for enforceable platform responsibilities. The Online Safety Act 2023 combines regulatory duties with financial penalties and court-ordered business-disruption measures. It also expressly requires regard to users' lawful expression and privacy: [Online Safety Act 2023, ss 22\(2\)–\(3\), 143–144](#).

That is a precedent for the regulatory architecture—not a statement that the Act already creates the geographical-provenance duty we are discussing, or that any such new duty would necessarily be compatible with human rights.

An appropriately targeted platform duty could, for example, require:

- facilities for making and displaying required provenance declarations;
- proportionate verification for defined high-risk activities, such as paid political or commercial campaigns;
- investigation of adequately evidenced complaints about materially false provenance;
- correction or labelling where a breach is established;
- compliance with specific regulatory or judicial orders.

These are possible features of new legislation, not existing general duties.

2. “Ensure compliance” should not mean an impossible guarantee

A platform may know where an account connected to its service without knowing where the post was drafted, who commissioned it or whether the author was temporarily abroad. **The duty must correspond to information the platform can reasonably obtain and assess.**

A requirement to prevent every false geographical representation, backed by substantial penalties for each missed post regardless of reasonable precautions, would create serious risks:

- excessive location surveillance;
- inaccurate attribution based on technical indicators;
- removal of lawful posts whenever provenance is uncertain;
- exclusion of anonymous speakers unable safely to verify their location.

Those consequences affect users' Articles 8 and 10 rights even though enforcement is formally directed at the platform. Article 10 expressly addresses restrictions and penalties affecting expression: [Human Rights Act 1998, Schedule 1, Articles 8 and 10](#).

A stronger model would distinguish **systemic compliance failures**, **failure to act on a sufficiently substantiated notice**, and **breach of a specific order**. Proactive duties need not be excluded, but should be justified by the relevant activity, risk and practical means of compliance.

3. Financial penalties

Substantial financial penalties could be justified, including penalties capable of deterring large platforms. They must nevertheless be proportionate to the platform's own breach—not simply to misconduct by a user which the platform could not reasonably detect.

Relevant considerations would include:

- seriousness and duration;
- deliberate, reckless or repeated non-compliance;
- reach and consequences of the misleading material;
- reasonable preventive measures already taken;
- cooperation, correction and mitigation;
- the platform's resources and any benefit from non-compliance.

There should also be clear statutory limits, reasons for the decision, an opportunity to make representations and effective independent challenge. The Online Safety Act illustrates a statutory penalties framework; its existence does not determine the appropriate amount for a different regime: [Online Safety Act 2023, s 143 and Schedule 13](#).

The governing proportionality inquiry includes whether less intrusive measures could achieve the objective without unacceptable compromise and whether a fair balance is struck: [Bank Mellat v HM Treasury \(No 2\)](#) [2013] UKSC 39.

The incentive structure matters: a penalty regime which makes indiscriminate removal the commercially safest response may be disproportionate even if its stated objective is merely transparency.

4. Injunctions

Targeted injunctions could also be justified—and may sometimes be preferable to penalties alone. They can specify the particular breach and the action needed to remedy it.

There is established authority for injunctions requiring internet intermediaries to assist against wrongdoing even where they are not themselves the primary wrongdoer. In *Cartier*, the Supreme Court recognised the equitable basis for website-blocking orders against innocent intermediaries: [Cartier International AG v British Telecommunications plc](#) [2018] UKSC 28, [8]–[15].

However, *Cartier* concerned intellectual-property infringements. **It does not itself establish a general power to enforce the proposed provenance regime.** New legislation should clearly identify the duties, enforcement applicants and available orders.

Depending on the established breach, an order could require the platform to:

- display an accurate correction or required disclosure;
- stop distributing specified material unless the disclosure is supplied;
- remedy an identified failure in its compliance systems.

An injunction should be clear, technically feasible and no wider than necessary. Removing an entire account—or restricting an entire service—would require substantially stronger justification than correcting a particular post.

The Online Safety Act's service-restriction provisions illustrate this distinction: they require judicial consideration of proportionality and the rights of affected parties, including UK users: [Online Safety Act 2023, s 144\(6\)–\(9\)](#).

5. Judicial safeguards remain essential

Courts are themselves public authorities for Human Rights Act purposes. A statutory enforcement power does not remove their obligation to consider Convention rights: [Human Rights Act 1998, s 6\(1\), \(3\)\(a\)](#).

Where relief may affect freedom of expression, section 12 also applies. In particular, it contains safeguards concerning unrepresented respondents and pre-trial restraints on publication, and requires particular regard to expression: [Human Rights Act 1998, s 12\(1\)–\(4\)](#).

Affected speakers should therefore have an appropriate opportunity to contest a disputed provenance finding, particularly where the proposed order would expose their location or prevent publication.

Conclusion

Platform duties, financial penalties and injunctions are all capable of human-rights justification. They can also operate together: an injunction secures compliance, while a penalty addresses an established breach. Their combined burden must remain proportionate.

The strongest scheme would impose **clear, achievable duties**, require **reliable evidence of materially misleading provenance**, prefer **correction or disclosure over unnecessary suppression**, and reserve severe sanctions for serious or persistent failures. An absolute obligation to guarantee the true geographical origin of every post, irrespective of knowledge or practical capability, would be considerably harder to defend.

Could you draft a bill that would be the basis for a statute to put the measures discussed into effect?

◆ **Vincent** (2026-10-29 11:41:06)

I've prepared the **Social Media Provenance and Transparency Bill**, with 30 proposed sections and an explanatory overview. It covers automation labels, political and commercial backing, misleading geographical provenance, platform duties, financial penalties, injunctions and appeals, with safeguards for expression, privacy and legitimate anonymity.

The document flags the principal policy choices—including OFCOM's role and proposed penalty ceilings—and the work needed before introduction; it is a legislative discussion draft, not a certification of human-rights compatibility. You can download it as Word or PDF using the document's download control.

👤 **User** (2026-10-29 12:02:38)

Could legislation also cover social media algorithms in order (1) to ensure that there is transparency as to how they operate (2) to ensure that content controls are easily accessible and easy to operate and (3) to protect vulnerable individuals? In relation to (1) I have in mind an obligation to set out potential biases in the content directed to users. In relation to (2) I have in mind something more specific than an option to opt out of targeted advertising. In relation to (3) I have in mind algorithms that could potentially direct harmful content to vulnerable people, such as suicide-related content to those with suicidal ideation and to make provision for specific requirements to carry out risk analyses in relation to specified content.

◆ **Vincent** (2026-10-29 12:02:39)

Yes. Legislation could address all three matters, and they would form a coherent additional Part of the Bill. The focus should be on **how platforms select, rank and repeatedly recommend content**, not merely advertising or the legality of individual posts.

There is already a substantial statutory foundation. The additional legislation should build on that foundation while making the duties more specific, particularly for **vulnerable adults, meaningful feed controls and disclosure of amplification biases**.

1. The existing statutory foundation

The Online Safety Act 2023 already addresses algorithms in several relevant respects:

- **Illegal-content risk assessments** must consider algorithms and how easily, quickly and widely content may spread.
- **Children's risk assessments** require separate assessment of specified harmful content, taking algorithms and different age groups into account.
- **Category 1 adult-user assessments and controls** concern specified content, including content encouraging, promoting or providing instructions for suicide, deliberate self-injury and eating disorders.
- Adult control features must be **easy to access**, and users must be offered an early opportunity to choose their settings.

See [Online Safety Act 2023, ss 9, 11 and 14–16](#).

For children, content encouraging, promoting or providing instructions for suicide, self-injury or eating disorders is expressly designated as primary priority content: [Online Safety Act 2023, s 61\(3\)–\(5\)](#).

Those provisions do not amount to a general requirement to explain every recommender system, provide comprehensive feed controls, or protect all vulnerable adults from every form of harmful recommendation. Nor should the existence of statutory provisions be confused with the precise implementation timetable for each service category.

2. Transparency about algorithms and potential biases

What could be required?

A defensible duty could require a platform to publish an intelligible **recommender-system statement**, explaining:

- its principal ranking objectives—for example, predicted engagement, viewing time, popularity, relevance or revenue;
- the principal information used to personalise recommendations;
- whether viewing, pausing on, sharing or commenting on content is treated as evidence of interest;
- how paid promotion, commercial relationships and platform preferences affect visibility;
- how negative feedback and user controls alter recommendations;
- known or reasonably foreseeable **systematic tendencies in the content selected or amplified**.

This should not ordinarily mean publishing source code. There should be **public explanation sufficient for users**, supported by **more detailed confidential access for the regulator and suitably authorised auditors**.

Your proposed bias disclosure

An obligation to identify potential biases is feasible, but “bias” should be defined by reference to demonstrable selection and amplification effects. Otherwise it risks becoming a vague obligation to declare whether an algorithm is politically “balanced”.

The duty could address whether the system tends to:

- prioritise emotionally provocative or sensational material;
- reinforce previously inferred preferences;
- increase the repetition or intensity of a subject;
- favour particular languages, geographical sources, commercial interests or classes of publisher;
- underrepresent less popular perspectives;
- mistake distressed or critical engagement for a wish to receive more similar content.

These are **matters to investigate and disclose where supported**, not assumptions that every platform exhibits those tendencies.

A meaningful duty should require providers to **test for reasonably foreseeable biases**, explain the evidence and limitations, and update their statements after material changes. Merely requiring disclosure of biases already admitted by the provider would be too weak.

The legislation should require candour about the system—not governmental determination of which opinions deserve equal prominence.

3. Content controls extending beyond advertising

Yes: the duty could expressly apply to the ordinary feed, suggested posts, video queues, search recommendations, notifications and autoplay—not just targeted advertising.

Possible statutory requirements include:

Control	Practical effect
A feed not based on behavioural profiling	Enables access without recommendations derived from inferred interests.
A following-only or chronological feed	Gives users an understandable alternative to engagement-ranked recommendations, where applicable to the service.
Topic-specific suppression	Allows users to reduce specified kinds of content across recommendation surfaces.
An effective “do not recommend this” control	Prevents the system treating further distressed engagement as overriding that instruction.
Recommendation-profile reset	Lets users clear inferred interests without deleting their account.
Controls over autoplay and repeated recommendations	Reduces involuntary continuation and cumulative exposure.
A clear explanation of individual recommendations	Enables the user to understand and change the principal reason for receiving them.

These are proposed legislative options, not a description of universal existing rights.

The important point is to regulate **usability and effect**, not merely the existence of a settings page. Requirements could include:

- access directly from the feed or recommended item;
- a short, intelligible route to changing settings;
- accessible design and language;
- settings which persist across sessions;
- no misleading prompts or repeated pressure to reverse a choice;
- no reduction in unrelated service functionality merely for exercising a control;
- testing whether the control actually changes the content received.

The existing statutory requirement that certain adult controls be easy to access provides a foundation, but more precise duties could specify these outcomes: [Online Safety Act 2023, s 15\(2\)–\(8\)](#).

4. Protection of vulnerable individuals

This is potentially the most important addition: protection should not depend entirely on a vulnerable user first recognising the risk and finding the correct setting.

A user experiencing suicidal ideation may engage with distressing content in ways that a recommender interprets as a preference. The legislative concern is therefore not just harmful material being available, but **a system repeatedly pushing it towards someone for whom that exposure presents a foreseeable serious risk.**

A proportionate duty could require providers to assess and mitigate that risk without guaranteeing that no harmful exposure will ever occur.

Mandatory content-specific risk assessments

The legislation could specify initial categories in a Schedule, including:

- suicide encouragement, promotion and instructional content;
- self-injury encouragement, promotion and instructional content;
- eating-disorder encouragement and instructional content;

- other tightly defined categories for which evidence establishes a serious recommendation-related risk.

For each category, assessments should examine:

1. **Who may be particularly susceptible**, including temporary vulnerability.
2. **How recommendations begin**, including after searches or isolated interactions.
3. **Repetition, escalation and cumulative exposure**, rather than each item in isolation.
4. **Whether the system mistakes distress, concern or compulsive viewing for preference.**
5. **Interactions between categories**, such as eating-disorder and self-injury content.
6. **Whether mitigation works**, including user controls and changes to ranking.
7. **The risk of suppressing beneficial material**, such as recovery advice, support groups and responsible reporting.

Assessments should occur **before deployment or material alteration**, periodically afterwards, and following credible evidence of serious failures. They should lead to a recorded mitigation plan, not simply a completed assessment.

The existing Act already requires category-specific assessments in relevant contexts and expressly recognises cumulative harm from repeated or combined exposure: [Online Safety Act 2023, ss 9, 11, 14 and 234\(4\)–\(5\)](#).

Avoid requiring platforms to diagnose users

The duty should not require universal mental-health profiling or a clinical diagnosis of each user. That could itself create serious privacy and discrimination risks.

A better structure would combine:

- safeguards against dangerous recommendation patterns across the service;
- confidential, voluntary safety preferences;
- carefully limited responses to reliable indications of heightened risk;
- restrictions on using inferred vulnerability to optimise engagement;
- safeguards against repurposing sensitive information for advertising.

“Suicide-related content” is too broad as a single prohibited category. **Supportive discussion, recovery accounts and help-seeking must be distinguished from encouragement, instructions and other content whose form or repeated recommendation creates a serious risk.**

5. Human-rights justification and limits

These measures could be justified as protecting health and the rights of others. But their design must preserve expression, privacy and access to beneficial information: [Human Rights Act 1998, Schedule 1, Articles 8 and 10](#).

Restricting unsolicited algorithmic amplification is different from prohibiting publication, although both can affect expression and must be justified. That distinction gives legislation scope to address harmful recommendation patterns without automatically deleting the underlying material.

The statutory approach should therefore prefer, where effective:

- safer ranking and recommendation design;
- meaningful controls;
- limits on repeated unsolicited exposure;
- access to supportive material;

over indiscriminate removal of all discussion of a sensitive subject.

The Online Safety Act itself recognises that safety measures must address expression and privacy: [Online Safety Act 2023, s 22](#).

6. How this would fit into the Bill

I would add a distinct Part entitled **“Recommender systems: transparency, user control and protection against serious harm”**, covering:

- recommender-system explanations and evidence-based bias disclosures;
- enforceable minimum feed-control standards;
- separate risk assessments for scheduled content categories;
- duties to implement and test proportionate mitigations;
- independent scrutiny and confidential regulatory access;
- public summaries which do not expose sensitive user data;
- complaints, appeals and enforcement through the Bill’s existing machinery.

The key drafting choice is to impose duties about the design and operation of recommendation systems, not merely duties to remove particular posts. That would

address your concern directly, including where individual content is lawful but its persistent, personalised amplification presents a foreseeable risk of serious harm.

I have not altered the existing draft. These additions would materially expand its scope and should be integrated with the Online Safety Act rather than left as a parallel, potentially inconsistent regime.

 **User** (2026-10-29 12:10:10)

Draft as a separate Part to the Bill measures to put those matters into effect. Identify in tabular form any amendments that would be required to the Online Safety Act 2023 and any other material information.

 **Vincent** (2026-10-29 12:10:12)

I've drafted a **separate Part 7**, with clauses 31–45 and a content-risk Schedule, leaving the existing Bill unchanged. It covers bias transparency, practical feed controls, vulnerable-user protection, content-specific assessments, audits and enforcement.

The document includes **two coordination tables** identifying Online Safety Act interfaces, necessary changes to the existing Bill, optional amendments and implementation matters. It distinguishes additional duties under this Bill from amendments that would be needed to consolidate the scheme within the Online Safety Act.

You can download the document as Word or PDF using its download control.

This conversation was exported from Vincent on October 29, 2026.

Conversation ID: 01-6465dd6df3154ef19d99b7fe1cc75197

Social Media Provenance and Transparency Bill

Discussion draft for legislative development

Drafting note

This draft proposes a United Kingdom-wide regulatory scheme administered by OFCOM. It covers automation labels, disclosure of political and commercial sponsorship and control, and correction of materially misleading geographical provenance, whether or not automation is used. It does not impose a general real-name or location-publication requirement.

The selection of OFCOM, the proposed maximum financial penalty, the appeal arrangements and the commencement periods are drafting proposals, not previously agreed instructions. Before introduction, Parliamentary Counsel should settle consequential amendments, devolved competence and legislative-consent questions, public expenditure and regulatory funding arrangements, and interaction with existing electoral, consumer, online-safety and data-protection legislation. This draft does not certify Convention compatibility.

The provisions below are proposed statutory text. This drafting note is not part of the Bill.

A BILL TO

Make provision about the identification of automated accounts on social media; the disclosure of persons responsible for political and commercial communications; the correction of materially misleading representations of geographical provenance; duties of social-media service providers; regulation and enforcement by OFCOM; financial penalties and court orders; and connected purposes.

BE IT ENACTED by the King's most Excellent Majesty, by and with the advice and consent of the Lords Spiritual and Temporal, and Commons, in this present Parliament assembled, and by the authority of the same, as follows:—

Part 1 — Preliminary

1. Purposes and governing principles

(1) The purposes of this Act are to protect persons in the United Kingdom against materially misleading representations about the automation, sponsorship, control and geographical provenance of communications on

social media, and to promote accountability for organised political and commercial influence.

(2) A person exercising functions under this Act must act compatibly with Convention rights and must have particular regard to—

(a) freedom of expression, including anonymous expression and the receipt and imparting of information regardless of frontiers;

(b) privacy and the protection of personal data;

(c) the protection of democratic participation and consumers;

(d) the importance of preserving lawful journalism, confidential sources, satire, artistic expression and accessibility assistance;

(e) the accuracy and limitations of evidence about geographical provenance; and

(f) the effectiveness, necessity and proportionality of any measure imposed.

(3) This Act does not prohibit a communication merely because it originates outside the United Kingdom or expresses a particular political opinion.

(4) A power under this Act must not be used to determine the truth or acceptability of a political opinion as such.

2. Regulated services and territorial connection

(1) A regulated service is an internet service which enables users to publish communications to other users or the public, and has a substantial connection with the United Kingdom.

(2) A service has a substantial connection with the United Kingdom if—

(a) the United Kingdom forms a material part of its intended market; or

(b) it has a significant number of United Kingdom users, having regard to its scale and nature.

(3) A regulated service does not include a service whose sole or principal function is private interpersonal communication and which does not provide a materially distinct public or mass-distribution function.

(4) Where a service has both regulated and unregulated functions, this Act applies only to its regulated functions and measures reasonably necessary to secure compliance in relation to them.

(5) The duties of a provider established outside the United Kingdom apply to the provision of its regulated service to United Kingdom users and communications directed to, or materially disseminated among, those users.

(6) Mere technical accessibility from the United Kingdom does not, without more, establish a substantial connection.

3. Interpretation

In this Act—

“account” includes a channel, page or other user-controlled publishing facility;

“automated account” means an account in which software makes substantive decisions about the selection, generation or dissemination of communications without human approval of each communication;

“commercial communication” means a communication made in the course of business to promote goods, services or a business, including a communication commissioned for that purpose;

“Convention rights” has the meaning given by section 1 of the Human Rights Act 1998;

“court” means the High Court in England and Wales, the Court of Session in Scotland, or the High Court in Northern Ireland;

“geographical provenance” means the country in which a communication is substantively created, from which its publication is principally directed, or in which its responsible principal is established, as the context requires;

“operator” means a person exercising effective control over an account or the publication of a communication, but does not include a person solely because that person supplies general-purpose software, hosting or transmission facilities;

“political influence communication” means a communication whose principal purpose, judged objectively in context, is to influence an election or referendum in the United Kingdom, support for a political party, candidate or elected office-holder, or a decision about legislation or public policy by a United Kingdom public authority;

“provider” means the person with control over the provision of a regulated service;

“publish” means make available through a regulated service to the public or a substantial audience, but does not include a genuinely private interpersonal communication;

“responsible principal” means a person who commissions or exercises effective direction over a communication or campaign, or on whose behalf it is published;

“sponsor” means a person providing money or other material consideration specifically for the publication or organised dissemination of a communication or campaign;

“United Kingdom user” means a user who is in the United Kingdom.

4. Persons and activities within scope

(1) Sections 5 to 11 apply to an operator, responsible principal or sponsor who causes or procures publication directed to, or materially disseminated among, United Kingdom users.

(2) A person is not a responsible principal or sponsor merely because the person—

(a) agrees with or benefits from a communication;

(b) purchases a product promoted by it;

(c) makes a general donation unconnected with the relevant communication or campaign and does not direct it; or

(d) supplies ordinary technical or professional services without directing its substantive content or campaign strategy.

(3) This Act does not require disclosure of an individual's beliefs, nationality, birthplace, ancestry or political associations except to the extent that an objectively established sponsorship or control relationship is required to be disclosed under section 6.

Part 2 — Transparency and correction duties

5. Identification of automated accounts

(1) An operator of an automated account must ensure that its automated character is clearly identified to persons encountering its communications.

(2) Identification must appear prominently on the account and, where an account-level indication would not reasonably inform recipients, accompany the communication or be directly accessible from it with a visible indication of automation.

(3) The identification must be factual and must not imply that automation alone makes a communication false or unlawful.

(4) An account is not automated merely because it uses scheduling, spelling correction, translation, accessibility tools or drafting assistance, where a human determines the substance and approves each communication before publication.

(5) OFCOM must issue guidance distinguishing fully automated, substantially automated and human-approved publication and specifying proportionate means of identification.

6. Disclosure of political and commercial backing

(1) This section applies to—

(a) a commercial communication published for material consideration, or by or on behalf of a business promoting its goods, services or business; and

(b) a political influence communication published for material consideration or as part of an organised campaign directed by a political party, candidate, elected office-holder, body corporate or association.

(2) The operator must ensure that the following information is clearly and readily accessible from the communication—

(a) its commercial or commissioned political character, where not already apparent;

(b) the name of each responsible principal;

(c) the name of each sponsor whose contribution is specifically attributable to the communication or campaign and is material to its publication or organised dissemination;

(d) the nature of each relationship requiring disclosure; and

(e) the country of principal establishment or operation of each person identified under paragraphs (b) and (c).

(3) A responsible principal or sponsor must provide the operator with accurate information reasonably necessary for compliance.

(4) Disclosure must identify the person exercising effective direction, and not merely an intermediary established to conceal that person's involvement.

(5) It is not necessary to identify every employee, programmer, volunteer, contractor or general donor.

(6) An individual's unpaid expression of personal political views does not fall within this section merely because it is repeated, scheduled, widely read or supportive of a party or cause.

(7) An organisation must publish an appropriate business or service address for contact about its communications. An individual may use an appropriate service address; this Act does not require publication of a residential address.

(8) This section is subject to section 11.

7. Country of direction and establishment

(1) A geographical disclosure under section 6 must distinguish, where materially different—

(a) the country in which the responsible principal has its principal establishment or centre of relevant activities; and

(b) the country from which effective direction of the campaign is exercised.

(2) Incorporation, server location, a network connection or the location of a technical contractor does not alone establish the country of effective direction.

(3) Where direction is genuinely distributed across countries, the disclosure must accurately describe that fact and identify the countries materially involved so far as reasonably ascertainable.

(4) Temporary travel does not of itself change a person's principal country of operation.

(5) The requirement is for country-level information, not precise physical or live location.

8. Materially misleading geographical representations

(1) A person publishing a communication must not knowingly or recklessly make a materially misleading representation about its geographical provenance without an effective correction accompanying that representation.

(2) A representation may be express or arise from the overall presentation of the communication or account.

(3) A representation is materially misleading only where—

(a) an ordinary reasonable recipient, considering the communication in context, would understand it to assert a particular fact about provenance;

(b) reliable evidence establishes that the asserted fact is false or misleading; and

(c) the discrepancy is likely materially to affect recipients' assessment of the communication's independence, first-hand character, commercial credibility or claimed participation in domestic democratic debate.

(4) A representation is not established merely by language, discussion of a country's affairs, use of a national symbol, or expression of solidarity with its inhabitants.

(5) A claim about residence or citizenship is not contradicted merely by creation or publication while abroad.

(6) A report about an event in one country is not misleading merely because it is edited, translated, uploaded or republished elsewhere.

(7) Clearly recognisable fiction, satire, parody and artistic performance do not contravene this section.

(8) A person who discovers an inadvertent material error must correct it within a reasonable time. A failure to do so without reasonable excuse is a breach of that duty, but does not retrospectively establish knowledge or recklessness at publication.

9. Extent of geographical correction

(1) A correction under section 8 must be sufficient to dispel the materially misleading impression and must not disclose more personal information than is necessary for that purpose.

(2) Depending on the circumstances, a correction may consist of—

(a) withdrawing or qualifying the provenance claim;

(b) stating that the communication was not created in, or directed from, the country claimed; or

(c) identifying its actual country of creation or direction.

(3) Disclosure under subsection (2)(c) may be required only where the actual country is materially relevant and a correction under paragraph (a) or (b) would not adequately protect the interest concerned.

(4) A correction must distinguish creation, publication, direction and sponsorship where failure to do so would itself mislead.

(5) An operator must not be required publicly to disclose an individual's precise location, residential address or live movements under this section.

10. Form and accessibility of disclosures

(1) A disclosure or correction must be intelligible, reasonably prominent and accessible without charge.

(2) A directly accessible disclosure may be used where including all required information in the communication would be impracticable, provided the communication visibly indicates the nature of the disclosure.

(3) Providers must, so far as reasonably practicable, preserve required disclosures when communications are shared using their services' own sharing functions.

(4) OFCOM may specify technical standards by rules following consultation, but must permit equivalent means of compliance and have regard to accessibility, service design and technical feasibility.

(5) A republication which genuinely quotes or reports on another person's communication is not an adoption of its provenance claim merely by reason of republication. Existing disclosures must not knowingly be stripped away so as materially to mislead.

11. Protected disclosure arrangements

(1) A person may apply to OFCOM for an arrangement permitting confidential verification, reduced public particulars or another effective means of compliance where ordinary disclosure would create a real and

substantial risk to safety, expose a confidential journalistic source, or cause a comparably serious interference with Convention rights.

(2) OFCOM must determine the application promptly and may make an interim arrangement pending determination.

(3) The applicant's identifying information must be protected pending determination and any timely appeal, unless disclosure is authorised by the court as necessary and proportionate.

(4) An arrangement must preserve the transparency objective as far as practicable, including by identifying an organisation or publishing a non-identifying correction where appropriate.

(5) This section does not authorise retention of a materially false provenance claim. Where safe and truthful correction is not possible, the claim must be withdrawn.

(6) OFCOM must give reasons, in a form which does not defeat the protection sought, and inform the applicant of the right of appeal under section 24.

Part 3 — Duties of regulated service providers

12. Proportionate compliance systems

(1) A provider must establish, maintain and operate effective and proportionate systems designed to secure compliance with Part 2 on its service.

(2) Those systems must include—

(a) facilities enabling users to make required disclosures;

(b) clear explanations of the requirements;

(c) risk-based measures for identifying organised evasion;

(d) procedures under sections 14 and 15; and

(e) records sufficient to demonstrate compliance without unnecessary collection of personal data.

(3) Proportionality must be assessed by reference to the provider's size, resources, service functions, access to relevant information, technical capability and the nature and scale of the relevant risks.

(4) A provider is not an absolute guarantor of the truth of every communication and does not breach this section merely because a user circumvents reasonable measures.

(5) A provider must not satisfy this duty by indiscriminately removing lawful communications or imposing unnecessary identity or location checks on all users.

13. Verification for defined activities

(1) A provider must take reasonable steps to verify the responsible principal, relevant sponsorship and country-level information for—

(a) paid political advertising;

(b) paid commercial campaigns; and

(c) coordinated account networks where reliable evidence gives reasonable grounds to suspect material concealment of a common political or commercial principal.

(2) Verification may use appropriate business records, declarations and corroborating information proportionate to the risk.

(3) The provider must record whether information is verified, self-declared, disputed or not reasonably ascertainable; it must not publicly present an inference as a verified fact.

(4) Where required information cannot reasonably be verified, the provider must seek clarification and may refuse the relevant paid promotion or campaign facility pending compliance.

(5) The inability to verify information under this section does not itself justify excluding a person from unrelated lawful use of the service.

(6) Neither this section nor section 12 requires continuous location tracking, access to private messages, weakening of encryption, or a universal requirement for government-issued identification.

14. Notices of suspected breach

(1) A provider must offer an accessible means of reporting a suspected breach of Part 2.

(2) A report must identify the communication, the representation or missing disclosure complained of, the reason it is alleged to breach this Act, and available supporting evidence.

(3) The provider must assess a sufficiently particularised report within a reasonable time, taking account of urgency, likely harm, evidential strength and the importance of the expression concerned.

(4) Unless doing so would create a demonstrable serious risk or defeat a lawful investigation, the provider must notify the publisher and allow a reasonable opportunity to respond before making a disputed public attribution or restricting publication.

(5) Where reliable evidence establishes a breach, the provider must take proportionate steps to secure correction or disclosure, or restrict the affected communication to the extent necessary pending compliance.

(6) A complaint, disputed allegation or unexplained discrepancy in connection data does not alone establish a breach.

(7) The provider must explain its decision to the affected publisher and complainant, subject to lawful confidentiality and protection against abuse of reporting procedures.

15. Review, preservation and urgent action

(1) A provider must provide prompt internal review of a decision to label, correct, restrict or remove a communication under this Act.

(2) Review must be undertaken by a person with appropriate competence who was not responsible for the original contested decision.

(3) Necessary records must be preserved while a timely review or appeal is pending, subject to section 16.

(4) An urgent temporary restriction may be imposed before representations are received only where reasonably necessary to prevent serious and imminent harm and no less restrictive effective measure is reasonably available.

(5) A restriction under subsection (4) must be reviewed without delay, with notice and an opportunity to respond as soon as practicable.

(6) Reversal of an erroneous decision must include removal of the erroneous label, restoration of the communication where practicable and notification to the affected user.

16. Privacy and information security

(1) Personal data processed for this Act must be limited to what is necessary, kept accurate, secured against misuse and retained no longer than necessary for the lawful purpose concerned.

(2) Public country-level disclosure does not authorise publication of underlying identity documents, connection records or precise location data.

(3) This Act does not disapply data-protection legislation or create a power of covert surveillance, entry, search or interception.

(4) OFCOM and providers must undertake and keep under review assessments of the effects of their measures on expression, privacy, security and vulnerable users.

Part 4 — Regulation and financial enforcement

17. Functions of OFCOM

- (1) OFCOM is the regulator for this Act.
- (2) OFCOM must publish codes of practice on compliance, evidence, verification, notices, protected disclosure and proportionate enforcement after consulting—
 - (a) the Information Commissioner and the Electoral Commission;
 - (b) providers and representatives of users, journalists and civil society;
 - (c) persons with expertise in online security, accessibility and human rights; and
 - (d) the devolved administrations.
- (3) A code may explain but must not expand the statutory duties or create an offence or penalty.
- (4) Compliance with relevant recommended measures is evidence of compliance, but a provider may demonstrate equally effective alternative measures.
- (5) OFCOM must publish an annual report on enforcement, corrected errors, protected disclosures and effects on lawful expression, without exposing protected persons.
- (6) The Secretary of State must provide OFCOM with resources reasonably sufficient to discharge its functions under this Act, subject to Parliamentary authorisation of expenditure.

18. Information notices

- (1) OFCOM may by written notice require information reasonably necessary to investigate compliance, determine an application or exercise an enforcement function under this Act.
- (2) The notice must identify the statutory purpose, information required, reasons for requiring it, means of secure production and a reasonable compliance period.
- (3) The recipient may seek variation or apply to the Tribunal under section 24 to challenge the notice before its compliance date.
- (4) A timely challenge suspends the disputed requirement unless the Tribunal directs otherwise because urgent compliance is necessary and proportionate.

(5) A notice may not require legally privileged material, defeat protection against self-incrimination so far as applicable by law, or require creation of information which cannot reasonably be obtained.

(6) Information identifying a confidential journalistic source may be compelled only by the court, on application by OFCOM, where disclosure is necessary in the interests of justice or national security or for the prevention of disorder or crime, and is proportionate having regard to source protection and Convention rights.

(7) Confidential information may be used and disclosed only for lawful functions under this Act or pursuant to an independently applicable legal obligation, and subject to appropriate safeguards.

19. Preliminary findings and compliance notices

(1) Before finding a breach or imposing a financial penalty, OFCOM must give the person concerned a notice setting out—

- (a) the alleged breach and material evidence;
- (b) proposed remedial requirements and any proposed penalty;
- (c) reasons, including the proportionality assessment; and
- (d) a reasonable period for representations, ordinarily not less than 28 days.

(2) The person must receive sufficient disclosure to challenge the case effectively, subject to lawful protections for privileged, confidential or safety-sensitive information and directions of the Tribunal.

(3) After considering representations, OFCOM may make a reasoned final decision finding a breach and requiring specified remedial steps by a reasonable date.

(4) A notice must specify achievable steps rather than require a guarantee that no future user breach will occur.

(5) A decision must explain the appeal right and any proposed publication, which must respect section 11.

20. Financial penalties

(1) OFCOM may impose a financial penalty on—

- (a) a provider for breach of a duty under Part 3 or failure without reasonable excuse to comply with a final compliance notice;
- (b) an operator, responsible principal or sponsor who intentionally or recklessly breaches Part 2, or fails without reasonable excuse to comply with a final compliance notice; or

- (c) a person who without reasonable excuse fails to comply with an enforceable information notice, or knowingly or recklessly supplies materially false information in response to it.
- (2) A user's breach does not of itself establish a provider's breach.
- (3) A penalty must be effective, proportionate and no greater than necessary, having regard to—
- (a) seriousness, duration, reach and consequences;
 - (b) intention, recklessness, repeated conduct and concealment;
 - (c) preventive measures, cooperation, correction and mitigation;
 - (d) financial benefit and the person's means; and
 - (e) sanctions already imposed for the same conduct.
- (4) The maximum penalty for a provider is the greater of £18 million and 10 per cent of its qualifying worldwide revenue for the most recent completed accounting year.
- (5) The maximum penalty for another person is £50,000; OFCOM must consider that person's actual means and the risk of suppressing lawful participation.
- (6) Qualifying worldwide revenue means revenue of the person providing the service, excluding taxes on sales and intra-person accounting transfers. Revenue of another legal person must not be attributed without express provision made by affirmative regulations identifying control criteria and safeguards against disproportionate attribution.
- (7) For continuing failure after a final compliance date, OFCOM may impose a daily penalty, but must state the daily rate, commencement and maximum duration and ensure that the aggregate does not exceed the applicable maximum for the same continuing breach.
- (8) A penalty must not be multiplied by the number of views or shares, or by artificial division of a single continuing breach.
- (9) A penalty is recoverable as a civil debt only when the appeal period has expired without appeal or the appeal has been finally determined.
- (10) Money received by OFCOM by way of penalty must be paid into the Consolidated Fund.

21. Enforcement coordination

- (1) OFCOM must consult the Electoral Commission, Information Commissioner, Competition and Markets Authority or other relevant regulator where the same conduct materially engages their functions.

(2) OFCOM must avoid unnecessary duplication of investigations and take account of other sanctions for the same conduct.

(3) OFCOM may accept and enforce a written undertaking specifying remedial steps and a compliance period, without imposing a penalty where that would adequately achieve the purposes of this Act.

(4) Neither an undertaking nor a compliance notice may require conduct that would otherwise be unlawful.

(5) This Act creates no automatic right to damages against a provider merely because a user has contravened Part 2.

Part 5 — Judicial enforcement and appeals

22. Enforcement orders and injunctions

(1) On application by OFCOM, the court may grant an injunction, interdict or other order to prevent or remedy an established or sufficiently evidenced threatened breach of this Act.

(2) An order may require a person within the court's jurisdiction to—

(a) display a specified accurate disclosure or correction;

(b) cease making a specified materially misleading provenance claim;

(c) restrict distribution of an identified communication unless and until the statutory disclosure requirement is satisfied;

(d) implement specified, achievable improvements to compliance systems; or

(e) preserve specified records for a limited period where necessary for proceedings under this Act.

(3) The court must be satisfied that—

(a) the order has a sufficient legal and evidential basis;

(b) its terms are clear and technically feasible;

(c) it is necessary and proportionate to an identified legitimate aim;

(d) an effective less restrictive measure would not adequately achieve that aim; and

(e) the rights of affected publishers, recipients and providers have been considered.

(4) OFCOM must identify and, so far as reasonably practicable, notify a publisher whose communication or protected information is directly affected, as well as the proposed respondent.

(5) The court may permit an affected person to intervene without publicly identifying that person where necessary to protect Convention rights.

(6) An order must specify its duration or a review date and provide liberty to apply to vary or discharge it.

(7) Unless the court gives specific reasons, an order must be limited to the communication, campaign or service function and territorial audience necessary to remedy the breach.

(8) This section does not authorise blocking an entire regulated service, withdrawing unrelated infrastructure services or imposing a general obligation to determine the provenance of every communication.

(9) Costs are in the discretion of the court. The court must distinguish an innocent intermediary assisting enforcement from a provider remedying its own established breach.

23. Interim relief and expression

(1) The Human Rights Act 1998, including section 12, applies to proceedings under section 22.

(2) An order without notice may be made only where the court is satisfied that there are compelling reasons for proceeding without notice and that the order is necessary and proportionate.

(3) The applicant must disclose material matters adverse to the application, including known disputes about location evidence and available less restrictive measures.

(4) An order without notice must provide for a return hearing as soon as practicable and ordinarily within seven days.

(5) An interim order must not compel irreversible public disclosure of an individual's identity or location unless the court is satisfied that there is an overriding need for immediate disclosure and that no less intrusive effective measure is available.

(6) In considering interim relief, the court must consider correction, qualification, temporary suspension of paid promotion and preservation of evidence before restricting publication more widely.

24. Appeals and challenges

(1) The First-tier Tribunal has jurisdiction under this section and is referred to in this Act as "the Tribunal".

(2) A person affected may appeal against—

(a) a determination under section 11;

(b) an information notice under section 18;

(c) a final finding or compliance notice under section 19; or

(d) a financial penalty under section 20.

(3) An appeal against a final decision must ordinarily be brought within 28 days of notification. The Tribunal may extend time where just, including where safety or confidentiality arrangements have impeded access to proceedings.

(4) On appeal the Tribunal must determine disputed facts, legal questions and proportionality for itself, and may confirm, vary, quash or remit the decision and give appropriate directions.

(5) The Tribunal may receive further evidence and must provide an effective opportunity to challenge the evidence relied upon.

(6) A financial penalty is suspended pending a timely appeal. Other requirements may be suspended or varied by the Tribunal; irreversible public disclosure of protected identifying information is suspended pending a timely appeal unless the Tribunal directs otherwise on grounds of necessity and proportionality.

(7) Tribunal procedure must provide for urgent applications, confidentiality, protected participation and effective judicial scrutiny of any restriction on access to material evidence.

(8) Appeals from the Tribunal on a point of law lie to the Upper Tribunal with permission in accordance with the Tribunals, Courts and Enforcement Act 2007.

(9) An affected publisher may apply to the Tribunal for review of a public provenance attribution or restriction which a provider states is required by a final OFCOM compliance notice, so far as that publisher had no effective opportunity to challenge the underlying finding. The Tribunal may extend the time for that application and give interim directions.

(10) This section does not exclude judicial review or the court's jurisdiction to vary or discharge an order under section 22.

Part 6 — General

25. Relationship with other legislation

(1) This Act does not disapply the Human Rights Act 1998, data-protection legislation, the law protecting confidential journalistic sources, or any applicable protection of legally privileged information.

(2) Compliance with an equivalent disclosure requirement under electoral or consumer legislation satisfies the corresponding requirement of this Act to the extent that the same information is supplied with equivalent accuracy, accessibility and prominence.

(3) No failure to disclose under this Act is established merely because information protected by a valid arrangement under section 11 is withheld from public view.

(4) An exemption from liability for user content under another enactment does not exempt a provider from its own expressly imposed duties under Part 3 or from an order lawfully made under section 22.

(5) Subsection (4) does not otherwise remove any defence or exemption applicable to liability for the user's communication.

(6) Nothing in this Act creates a general power for a private person to demand another user's identity, location records or verification documents.

26. Regulations and procedural provision

(1) The Secretary of State may by statutory instrument make provision about —

(a) procedural arrangements and secure means of supplying required information;

(b) calculation and proof of qualifying worldwide revenue within section 20;

(c) coordination between regulators; and

(d) transitional arrangements.

(2) Regulations must not expand the substantive categories of political expression requiring disclosure, require publication of precise individual location, create criminal offences, increase the maximum penalties, or authorise whole-service blocking.

(3) Regulations under subsection (1), other than transitional provision made solely for commencement, require approval by resolution of each House of Parliament.

(4) Before making regulations, the Secretary of State must consult OFCOM, the Information Commissioner, the Electoral Commission, the devolved administrations and other materially affected persons, and publish an assessment of necessity, proportionality and effects on expression and privacy.

(5) The Tribunal Procedure Committee may make procedural rules for proceedings under section 24 in accordance with its existing statutory powers.

27. Review of operation

(1) The Secretary of State must arrange an independent review of this Act's operation within three years after Part 2 comes fully into force and at intervals of not more than five years thereafter.

(2) The review must consider effectiveness, evasion, mistaken attribution, removal of lawful expression, privacy impacts, burdens on smaller services, effects on vulnerable speakers and the adequacy of appeals.

(3) The report and the Government's response must be laid before Parliament, excluding particulars whose disclosure would defeat protection under section 11.

28. Extent

This Act extends to England and Wales, Scotland and Northern Ireland.

29. Commencement and transition

(1) Sections 1 to 4 and 17, and this Part, come into force two months after Royal Assent so far as necessary for preparation, consultation and issuing codes.

(2) The remaining provisions come into force on a day or days appointed by the Secretary of State by regulations.

(3) A duty under Part 2 or 3 must not come into force until at least six months after publication of the codes materially necessary for its operation.

(4) Commencement regulations may provide proportionate transitional periods for smaller providers but must not create an indefinite exemption.

(5) A financial penalty may not be imposed for conduct occurring before the relevant duty came into force.

(6) Continued availability of a communication first published before commencement does not alone constitute a breach; a prospective requirement to correct or disclose may apply after notice and a reasonable compliance period.

30. Short title

This Act may be cited as the Social Media Provenance and Transparency Act 2026.

Explanatory overview — not part of the Bill

The Bill regulates provenance rather than the acceptability of opinions. Its three principal substantive measures are an automation label; identification of the responsible principals and material sponsors of defined political and commercial communications; and correction of materially misleading geographical claims, including claims made in human-authored posts.

Country of creation, direction and establishment are treated separately. Technical connection data does not automatically establish any of them. Actual-country disclosure is required to correct misleading provenance only

where a less intrusive correction is insufficient. There is no general public real-name requirement, residential-address duty or live-location requirement.

Platforms must operate proportionate systems and respond to substantiated cases, not guarantee that every user tells the truth. Enhanced verification attaches to defined activities, rather than every user. The draft expressly guards against indiscriminate removal, universal identity checks and continuous location tracking.

OFCOM may investigate, accept undertakings, issue compliance notices and impose civil financial penalties. The proposed provider ceiling follows the scale of the Online Safety Act's principal penalty model as a policy starting point; it is not an assessment that the same ceiling is necessary for this different subject matter. The separate proposed ceiling for other persons also requires policy and impact assessment. No criminal offence is created.

Court orders can require corrections, restrict specified non-compliant communications or remedy systems failures. They cannot under this Bill block an entire service. Orders affecting expression remain subject to section 12 of the Human Rights Act 1998. The Bill distinguishes compliance obligations of a provider in breach from the assistance of an innocent intermediary.

The First-tier Tribunal would hear merits appeals against OFCOM decisions under this Bill, with further appeals on points of law. Its procedural and institutional arrangements, including any necessary consequential amendments and judicial resourcing, require settlement before introduction.

Existing relevant frameworks include the Human Rights Act 1998, particularly sections 6 and 12 and Schedule 1 Articles 8 and 10; Elections Act 2022, Part 6; Digital Markets, Competition and Consumers Act 2024, Part 4 Chapter 1; Electronic Commerce (EC Directive) Regulations 2002; and Online Safety Act 2023. The proportionality approach and the distinction between intermediary assistance and responsibility reflected in *Bank Mellat v HM Treasury (No 2)* [2013] UKSC 39 and *Cartier International AG v British Telecommunications plc* [2018] UKSC 28 inform the safeguards. These authorities do not establish that this proposed Bill is Convention-compliant.

Recommender Systems: Additional Part to the Bill

Drafting and integration note

This is a separate proposed Part 7, comprising clauses 31 to 45 and a proposed Schedule 1, for the Social Media Provenance and Transparency Bill previously drafted. The existing Bill is unchanged. Numbering continues from clause 30 so that this draft can be considered and inserted without disturbing the existing clause numbers.

The chosen structure imposes additional duties under the new Bill, using its regulatory and judicial enforcement machinery. It does not purport to amend the Online Safety Act 2023 merely by referring to it. Existing Online Safety Act duties remain applicable. The tables following the statutory text distinguish amendments necessary for a different, consolidated approach from coordination measures appropriate to this additional-duty model.

The duties apply to the social-media services already defined in clause 2 of the Bill, irrespective of their Online Safety Act category. They do not independently extend to standalone search services or genuinely private interpersonal communications. That is a deliberate scope choice, not a statement about the coverage of the Online Safety Act.

Annual assessments, independent assurance at least every two years for designated high-risk systems, one-action access to controls and the commencement periods are proposed policy choices. They require impact assessment and technical consultation. The draft is a basis for legislative development, not a certification of Convention compatibility.

Proposed statutory text

Part 7 — Recommender systems: transparency, user control and protection against serious harm

31. Application and interpretation

(1) This Part applies to a provider of a regulated service which uses a recommender system in providing the service to United Kingdom users.

(2) In this Part—

“recommender system” means an automated or partly automated system used to select, rank, prioritise, suggest or repeatedly present content to a user, including a feed, suggested account, search suggestion, recommended

search result within the service, notification, video queue or autoplay sequence;

“behavioural personalisation” means selection or ranking based on observed or inferred information about a particular user or a group to which the user is assigned, other than preferences expressly selected by that user for that purpose;

“amplification bias” means a systematic tendency in selection, prominence, frequency or exclusion attributable to a system’s objectives, data, design or operation, whether intended or unintended;

“specified content” means content within Schedule 1;

“serious harm” means serious physical or psychological harm, including harm arising from cumulative or combined exposure;

“vulnerability” means a susceptibility to serious harm from particular content or patterns of exposure arising from age, disability, mental or physical health, distress, coercion or other circumstances, whether enduring or temporary;

“material change” means a change to a system, its objectives, data use or relevant interface which may materially alter exposure, control or risk under this Part.

(3) Vulnerability does not depend on a diagnosis, disclosure by the user or a legal finding of incapacity.

(4) This Part applies to recommendations of paid and unpaid content. Offering an advertising opt-out does not by itself satisfy any duty under this Part.

(5) A provider which procures a recommender system from another person remains responsible for compliance and must obtain the information and contractual access reasonably necessary for that purpose.

32. Public explanation of recommender systems

(1) A provider must publish and keep up to date a clear and intelligible explanation of each materially distinct recommender system used on its service.

(2) The explanation must identify—

(a) the system’s principal purposes, including the role of predicted engagement, viewing time, retention, revenue, popularity and user-selected relevance;

(b) the main classes of information used, their sources and whether they include inferred interests, sensitive inferences or activity outside the service;

(c) the main factors affecting selection and prominence, their relative importance in descriptive terms, and material interactions between those factors;

(d) the role of paid promotion, commercial arrangements and preferences for the provider's own content or services;

(e) the effects and limitations of the controls required by section 34;

(f) material amplification biases identified or reasonably foreseeable under section 33, affected groups or content types, and the supporting evidence and uncertainty; and

(g) material changes and their likely effects.

(3) The explanation must distinguish verified findings, reasonable predictions and matters not yet established; uncertainty must not be presented as proof of absence of bias or risk.

(4) A short explanation must be directly accessible from each principal recommendation interface, with access to further detail in accessible formats.

(5) A material change must be explained before it takes effect, except where an urgent security or safety correction requires immediate implementation, in which case explanation must follow without undue delay.

(6) Public disclosure need not include source code, personal data, confidential journalistic sources or detail whose publication would demonstrably enable circumvention of safety measures. An omission must be no wider than necessary and must not conceal the existence or nature of a material bias or risk.

33. Assessment and disclosure of amplification biases

(1) A provider must undertake and document proportionate testing for material amplification biases before deployment, before a material change and at least annually thereafter.

(2) Testing must consider, where relevant—

(a) reinforcement of inferred preferences and narrowing of the range of sources encountered;

(b) disproportionate amplification associated with emotional intensity, sensationalism, hostility or repeated exposure;

(c) differences associated with language, geographical source, disability, age or other relevant user characteristics;

(d) treatment of particular publishers, commercial interests and classes of viewpoint where evidence indicates a systematic selection effect;

(e) whether distressed, critical, compulsive or help-seeking engagement is treated as a preference for further similar content; and

(f) the reliability and limitations of the data and methods used.

(3) The provider must publish a meaningful summary of material findings, limitations and remedial action, integrated with the explanation under section 32.

(4) Testing must be sufficient to detect reasonably foreseeable material effects and must not be confined to effects already acknowledged by the provider.

(5) This section does not require equal prominence for every opinion, governmental approval of political views, or disclosure of the political beliefs of individual users.

34. Minimum user controls

(1) A provider must make available effective controls enabling a user to—

(a) select a recommendation mode not based on behavioural personalisation;

(b) where the service permits following or subscription, select a feed limited to followed or subscribed sources, with chronological ordering as an available option;

(c) reduce or exclude recommendations of specified topics, sources and accounts;

(d) give an instruction not to recommend a particular item or materially similar content and specify the relevant reason where the user wishes;

(e) inspect, correct, remove and reset interests inferred for recommendations, without deleting the account;

(f) disable autoplay, repeated recommendation prompts and engagement-based notifications separately; and

(g) activate a safety preference reducing unsolicited recommendations of each category in Schedule 1 without declaring a diagnosis or explaining the preference.

(2) Subsection (1)(a) permits contextual selection and preferences expressly selected by the user but not personalisation covertly reconstructed from prior behaviour or group assignment.

(3) A source-limited or chronological mode does not disapply independently applicable legal duties or a user's express safety settings. Any resulting exclusion must be explained accurately.

(4) A control must apply across the relevant feed, queues, notifications and other recommendation interfaces, subject only to a limitation which is necessary, explained and consistent with its stated purpose.

(5) User feedback expressly declining a topic must not be overridden merely because later viewing or pausing predicts engagement. The user must remain able deliberately to change the setting or search for lawful material, subject to applicable child-protection and other legal requirements.

(6) A provider need not supply a following-only feed where the service has no following or subscription functionality, or retain recommendation data which it does not otherwise hold. It must nevertheless supply equivalent meaningful control appropriate to its functions.

35. Accessibility, persistence and effectiveness of controls

(1) A provider must make the controls under section 34 easy to find, understand and operate.

(2) A clearly labelled route to the controls must be available through one user action from each principal recommendation interface. An item-specific feedback control must be available with the item or its immediate options menu.

(3) Settings must be changeable through a short and intelligible process, without unnecessary steps, misleading defaults, repeated persuasion, additional payment or loss of unrelated functionality.

(4) Choices must persist across sessions and across devices associated with the same account, so far as technically practicable. Where persistence is not possible without disproportionate data collection, the provider must explain that limitation and provide the most effective privacy-preserving alternative.

(5) A safety preference under section 34(1)(g) must take effect promptly. Delays or material limits must be explained and must not defeat the preference's protective purpose.

(6) Providers must test controls with representative users, including children where the service is likely to be accessed by them, disabled users and persons with relevant lived experience of vulnerability, using appropriate research safeguards.

(7) Testing must examine comprehension, number of steps, persistence and actual changes in exposure, not merely whether a control is displayed.

(8) OFCOM must prescribe by rules, following consultation, technical accessibility and effectiveness standards consistent with these minimum requirements. Rules must not dilute the rights in section 34 or this section.

36. Explanation of particular recommendations

- (1) A user must be able to obtain, directly from a recommended item, a concise and accurate explanation of the principal reasons for its recommendation.
- (2) The explanation must identify the principal behavioural signals or express preferences materially relied upon, whether payment or a commercial relationship influenced prominence, and the relevant controls.
- (3) A generic statement that the item may interest the user does not satisfy this duty.
- (4) An explanation must not expose another person's personal data or confidential sources, or unnecessarily reveal a sensitive health inference to another user of the device.
- (5) Where technical complexity prevents exact attribution, the provider must give the most informative reliable explanation available and state its limits; it must not manufacture an explanation.

37. Content-specific risk assessments

- (1) A provider must carry out a suitable and sufficient assessment of the risk of serious harm arising from its recommender systems in relation to each category of specified content separately.
- (2) Assessment is required—
 - (a) before first deployment;
 - (b) before a material change;
 - (c) at least annually;
 - (d) following credible evidence of a serious incident or previously unassessed risk attributable to recommendation; and
 - (e) when OFCOM gives a reasoned notice requiring reassessment of an identified material risk.
- (3) The assessment must address—
 - (a) foreseeable vulnerability, including children of different ages and vulnerable adults;
 - (b) entry points through searches, isolated interactions, inferred interests and recommendation of accounts or groups;
 - (c) repeated exposure, escalating intensity, cross-category interaction and cumulative effects;

- (d) the distinction between deliberate access and unsolicited recommendation;
 - (e) the possibility that engagement reflects distress, criticism, research, recovery or compulsive behaviour rather than a preference for further exposure;
 - (f) risks from autoplay, notification timing, popularity signals and engagement incentives;
 - (g) supportive and recovery-related content and the risk that mitigation would obstruct access to it;
 - (h) existing controls, default settings and their effectiveness in practice; and
 - (i) uncertainty, limitations and reasonably foreseeable evasion.
- (4) Lack of diagnosis or disclosure by users does not relieve the provider of assessing reasonably foreseeable risks at system and population level.
- (5) Assessments must consider evidence from appropriately qualified clinicians, researchers and persons with relevant lived experience, without exposing participants to avoidable harm.
- (6) A provider must keep the full assessment and supporting records, provide them to OFCOM on request under section 18, and publish an intelligible summary of material risks, mitigation and residual uncertainty.

38. Prevention and mitigation of serious recommendation-related harm

- (1) A provider must take effective and proportionate measures to prevent or materially reduce the foreseeable risks identified under section 37.
- (2) Measures must address system design and incentives and must not consist solely of publishing warnings or requiring users to find and activate controls.
- (3) Where relevant, measures must include—
- (a) preventing repeated unsolicited recommendation of content encouraging or instructing suicide, self-injury or disordered eating in circumstances presenting a foreseeable serious risk;
 - (b) interrupting escalating recommendation sequences;
 - (c) respecting safety preferences and negative feedback;
 - (d) excluding sensitive inferences about distress or vulnerability from engagement or advertising optimisation;
 - (e) providing unobtrusive access to appropriate support information; and

(f) testing false positives and maintaining access to lawful supportive, educational and recovery-related material.

(4) Measures must be selected by reference to risk, severity, effectiveness and interference with rights, and must be reviewed against measured outcomes.

(5) The duty is not a guarantee that no user will encounter harmful content or suffer harm. An assessment alone does not discharge it, and an isolated adverse outcome does not of itself establish a breach.

(6) A provider must record the measures adopted, reasons for rejecting reasonably available alternatives and a timetable for addressing residual risks.

(7) Protective settings must not be relaxed for a child merely because the child selects a less protective feed mode.

39. Sensitive information and protection without diagnosis

(1) This Part does not require clinical diagnosis, universal mental-health profiling, collection of medical records, weakening of encryption or examination of private interpersonal communications.

(2) A provider must, so far as reasonably practicable, prefer system-level and privacy-preserving safeguards to identification of individual vulnerability.

(3) A safety preference must be treated as an instruction about content, not as proof of illness or incapacity.

(4) An inference about vulnerability used solely for protection must not be used to target advertising, increase engagement, vary prices or eligibility, or make unrelated adverse decisions.

(5) Sensitive inferences must be appropriately secured, accessible only where necessary and deleted when no longer necessary for the lawful protective purpose. They must not be disclosed to another user, employer, insurer or advertiser by reason of this Part.

(6) Nothing in this Part supplies an exemption from data-protection legislation or independently authorises processing special-category personal data. The provider must identify and document any required lawful basis and additional processing condition.

(7) This Part does not impose automatic reporting of a user to police, health services or family members. Any such disclosure requires an independently applicable lawful basis and an assessment of necessity, proportionality and foreseeable consequences.

40. Incident response, complaints and records

(1) A provider must maintain an accessible procedure for reporting a harmful recommendation pattern, failure of a control, materially misleading algorithmic explanation or inappropriate restriction of supportive material.

(2) Reports may be made by the user, a person authorised by the user, a parent or person with parental responsibility where appropriate, or an organisation providing relevant support, without requiring disclosure of a diagnosis.

(3) The provider must triage urgent safety reports promptly, take proportionate interim action where necessary, give reasons and offer review by an appropriately trained person not responsible for the contested decision.

(4) A credible serious incident, including a reasonably suspected system-related death or serious injury, must be notified to OFCOM without undue delay after the provider has sufficient information to identify the relevant incident and plausible system connection. Notification is not an admission of causation or liability.

(5) Records relevant to an incident must be preserved for a proportionate period specified in published retention rules, subject to section 16 and any court order. Records must permit reconstruction of the relevant system version, recommendations, controls and response so far as reasonably practicable without unnecessary retention of unrelated user data.

(6) OFCOM must specify secure reporting arrangements and risk-based time limits after consultation with the Information Commissioner and relevant clinical and user representatives.

41. Regulatory access and independent assurance

(1) For functions under this Part, a notice under section 18 may require a provider to supply or generate proportionate test results, version records and explanations, or arrange a controlled demonstration of a recommender system.

(2) The notice may require confidential access to technical documentation and, where necessary and proportionate, model or code inspection by OFCOM or an independent assessor subject to enforceable confidentiality and security requirements.

(3) Public disclosure exemptions under section 32(6) do not of themselves justify withholding information from OFCOM under a lawful notice.

(4) OFCOM may designate a recommender system as high risk where evidence shows that its scale, design, user base or exposure patterns present a material risk of serious harm from specified content.

(5) Before designation, OFCOM must give reasons, disclose sufficient evidence for an effective response and allow a reasonable opportunity for

representations. Designation must be reviewed annually and withdrawn when no longer justified.

(6) A provider of a designated system must obtain independent assurance at least once in every two years and following a material change where OFCOM reasonably requires it. The assessor must be competent, independent of system design and commercial decision-making, and free from material conflicts of interest.

(7) Assurance must examine transparency, bias testing, usability, risk assessments, actual exposure and mitigation outcomes. The provider must fund reasonable costs proportionate to the risk and its resources, provide the full report to OFCOM and publish an intelligible summary.

(8) The assessor's appointment and scope must be approved by OFCOM. The provider must not prevent adverse findings from reaching OFCOM.

(9) Testing must not deliberately expose actual vulnerable users or children to harmful material where synthetic, controlled or otherwise safer methods can adequately test the relevant risk.

(10) All access and testing under this section remain subject to privilege, source protection, data-protection law and section 18 safeguards.

42. Codes, metrics and regulatory coordination

(1) OFCOM must issue codes for this Part after consulting the Information Commissioner, the devolved administrations, relevant health bodies, clinicians, researchers, providers, disabled users, children's representatives and persons with relevant lived experience.

(2) Codes must specify proportionate assessment and testing methods, accessibility standards, incident procedures and outcome measures, including recurrence after negative feedback, repeated exposure, escalation, control persistence and mistaken suppression of beneficial material.

(3) Engagement, viewing time and complaints alone must not be treated as sufficient measures of safety or informed user choice.

(4) Codes may give examples and recommend measures but must not narrow a statutory entitlement, expand a content category or substitute formal compliance for a duty expressly requiring effectiveness.

(5) OFCOM must align codes, information requests and reporting calendars with its functions under the Online Safety Act 2023. A combined assessment, audit or report may discharge overlapping procedural obligations where it separately addresses all requirements of each enactment.

(6) Compliance with a code issued under another enactment does not of itself establish compliance with this Part, or vice versa. No code or decision under this Part may authorise conduct prohibited by another enactment.

(7) OFCOM must avoid duplicate financial punishment for the same substantive failure, give reasons for the choice of enforcement route and take account of any sanction under the other regime.

43. Safeguards and assessment of compliance

(1) Measures under this Part must respect freedom of expression, privacy, equality and the protection of confidential sources.

(2) No provision requires suppression of a lawful opinion merely because it is controversial, unpopular or distressing.

(3) Restricting unsolicited amplification must be distinguished from removing a publication or preventing deliberate access. Any more restrictive measure requires its own necessity and proportionality justification.

(4) A provider must preserve access to lawful support, recovery, education, research and responsible reporting so far as consistent with effective protection, particularly the higher protection required for children.

(5) A person exercising a power must take account of the risk of discriminatory exclusion and unintended isolation of vulnerable users.

(6) In assessing a provider's breach, OFCOM must consider evidence of system performance and foreseeable risk, reasonable technical capability, resources and available alternatives. Outsourcing, commercial incentives and a failure to investigate reasonably foreseeable risks are not excuses.

(7) A measure must not be justified solely by asserting that users accepted general terms of service.

44. Schedule amendment and Parliamentary scrutiny

(1) The Secretary of State may by regulations amend Schedule 1 only to add, vary or remove a clearly defined content category where reliable evidence establishes or displaces a foreseeable risk of serious recommendation-related harm.

(2) Before making regulations the Secretary of State must obtain published advice from OFCOM, consult the persons listed in section 42(1), and publish an assessment of evidence, alternatives, effects on rights and protection of beneficial content.

(3) Regulations must not designate a political opinion as harmful merely because of its viewpoint or permit general monitoring of private communications.

(4) A draft instrument must be laid before and approved by resolution of each House of Parliament.

(5) A newly added duty must not take effect until providers have had a reasonable period, ordinarily at least six months, to assess and implement it,

subject to an earlier date justified by a published finding of serious and urgent risk.

45. Enforcement, integration and commencement

(1) Sections 1, 16 to 24 and 26 to 28 apply to this Part, with the modifications in this section. No provision of Part 2 or 3 is made a precondition to enforcement of a duty under this Part.

(2) For section 1, the purposes of this Act include meaningful algorithmic transparency, effective user control and protection against serious recommendation-related harm, and its governing principles apply to those purposes.

(3) A reference in section 20(1)(a) to a provider's duty under Part 3 includes a duty under this Part, a requirement imposed under section 35(8) or 40(6), and a lawful assurance requirement under section 41. The provider penalty ceiling and safeguards apply without an additional separate ceiling for this Part.

(4) OFCOM's functions under sections 17 to 19 and 21, and the court's powers under sections 22 and 23, extend to this Part, including orders requiring effective controls, adequate assessments, lawful audits and specified system improvements. They do not extend to whole-service blocking or a guarantee of a particular user's health outcome.

(5) Section 24 applies to a designation or requirement under section 41 as it applies to a final decision under section 19. A person directly affected may appeal within 28 days, with the extension and interim-relief powers in that section.

(6) This Part creates no automatic right to damages for a user's exposure or injury and does not exclude any independently available cause of action.

(7) A provider must describe material recommendation restrictions and controls in its terms of service before ordinary operation under this Part; urgent action lawfully required for safety must be explained promptly afterwards. This subsection does not disapply the Online Safety Act 2023.

(8) This Part and Schedule 1 extend to England and Wales, Scotland and Northern Ireland.

(9) This section and sections 31, 35(8), 40(6), 42 and 44 come into force two months after Royal Assent for preparatory functions, including making rules and issuing codes before the substantive duties take effect. The remaining provisions come into force by commencement regulations under section 29, but no earlier than six months after publication of the codes and standards materially necessary for their operation.

(10) Existing systems must be assessed, and required controls implemented, by the applicable commencement date. No penalty may be imposed for a failure occurring before that date. A material change after that date is subject to the pre-change duties.

Proposed Schedule 1 — Content requiring separate recommender-system risk assessment

1. Suicide

(1) Content encouraging, promoting, normalising or providing instructions for suicide, where the nature or manner of recommendation presents a reasonably foreseeable risk of serious harm.

(2) Graphic or method-focused depictions, accounts or representations of suicide where repeated, escalating or unsolicited recommendation presents such a risk.

2. Deliberate self-injury

(1) Content encouraging, promoting, normalising or providing instructions for deliberate self-injury, including poisoning, where the nature or manner of recommendation presents a reasonably foreseeable risk of serious harm.

(2) Graphic or technique-focused depictions, accounts or representations of self-injury where repeated, escalating or unsolicited recommendation presents such a risk.

3. Eating disorders and associated behaviours

(1) Content encouraging, promoting or providing instructions for an eating disorder or behaviours associated with an eating disorder where the nature or manner of recommendation presents a reasonably foreseeable risk of serious harm.

(2) Content promoting extreme restriction, purging or compulsive weight-control behaviour where repeated, escalating or unsolicited recommendation presents such a risk.

4. Context and beneficial content

(1) Classification must consider the content as a whole, purpose, audience, presentation and recommendation context. Mention of a topic or use of a keyword alone is insufficient.

(2) Supportive, recovery-related, educational, scientific, journalistic or artistic material is not to be treated as encouragement or instruction merely because it discusses or depicts a subject in this Schedule.

(3) Nevertheless, the assessment must consider whether the form, repetition or recommendation context of otherwise beneficial material creates a serious risk, and whether that risk can be addressed without suppressing its beneficial function.

(4) Inclusion in this Schedule creates assessment and mitigation duties, not a declaration that all content within a topic is unlawful or must be removed.

Legislative coordination tables — not part of the statutory text

A. Online Safety Act 2023: changes and coordination

The selected additional-duty model does not require wholesale rewriting of the Online Safety Act. Its duties continue alongside this Part. The table identifies where express clarification is advisable and what would require amendment if the alternative policy were to place all the new duties inside that Act. Existing provision numbers below identify the provisions to be considered; they are not assertions that those provisions already implement this draft.

Existing provision	Relationship to this draft	Amendment or action and status
Sections 9 and 10: illegal-content assessment and safety duties	Already address algorithms and dissemination of illegal content. The new Part also addresses lawful content and patterns of recommendation.	No textual amendment necessary for this model. Align assessments and mitigation through clause 42(5), without treating lawful specified content as illegal.
Sections 11 and 12: children’s assessments and safety duties	Existing child-specific duties cannot be replaced by adult choice or a less personalised feed.	No textual amendment necessary. Preserve the higher child standard and separate age-group assessment. Revise relevant codes to explain interaction. If consolidating, expressly add the new controls and assessment detail to these duties without weakening them.
Sections 14 to 16: Category 1 adult assessments and user empowerment	Already require assessments and accessible controls for defined content. New duties apply across the Bill’s regulated services and add persistent feed controls, explanation and system-level mitigation.	No amendment necessary to create the additional duties. If consolidating, amend the category-limited architecture, extend the control functions and content scope, and specify usability outcomes; amendment of section 15 alone would be insufficient.
Sections 20 and 21: content reporting and complaints	Existing procedures have specified statutory coverage; new complaints include	Procedural alignment required; textual amendment optional. A combined complaints process must

Existing provision	Relationship to this draft	Amendment or action and status
	harmful sequences and control failures, not just individual items.	separately cover clause 40. If consolidating, add these complaint types and standing expressly.
Section 22: expression and privacy	Existing cross-cutting duties relate to the Act's specified safety measures.	Do not assume automatic extension. Clauses 39 and 43 supply direct safeguards. If consolidating, extend section 22's scope to all new recommendation duties and resulting impact assessments.
Section 23: records and review	Existing record duties concern specified assessments and measures.	No amendment necessary under this model: clauses 37 to 41 impose additional records. If consolidating, extend the statutory assessment and record references to the new duties.
Sections 43 to 49 and Schedule 4: codes and their relationship to duties	The Act's code-compliance treatment must not be assumed to discharge a different Act's outcome and control duties.	Separate treatment required, supplied by clause 42(6). No repeal of existing code protection is proposed. If consolidating, settle expressly how section 49 applies to the new duties, particularly hard minimum user entitlements and effectiveness requirements.
Section 61: primary priority content harmful to children	Covers suicide, self-injury and eating-disorder encouragement, promotion and instructions. New Schedule 1 also examines risky repetition and depictions, including vulnerable adults.	No amendment necessary. Do not replace or dilute section 61. New Schedule 1 is a distinct assessment schedule, not an amendment making all suicide-related discussion prohibited content.
Section 71: restrictions in accordance with terms of service	Algorithmic intervention may restrict access, engaging the existing terms-of-service duty and exceptions.	Express clarification advisable, not necessarily indispensable. Clause 45(7) requires transparent terms; section 71(2)(b) already addresses action to avoid reasonably expected civil liability. For certainty, add an express exception in section 71(2) for action necessary and proportionate to comply with this new Part or a lawful order

Existing provision	Relationship to this draft	Amendment or action and status
Sections 77 and 78 and Schedule 8: transparency reporting	Existing reports do not automatically become comprehensive public algorithmic explanations for all services covered by this Bill.	<p>under it, while retaining explanation and complaints safeguards.</p> <p>No amendment necessary: clauses 32, 33, 37 and 41 supply direct publication duties. Align reporting calendars. If consolidating, extend both coverage and report particulars, preserving prompt change disclosures rather than relying only on annual reports.</p>
Sections 100 and 104: information and skilled-person powers	Their purposes and relevant-requirement gateways concern the Online Safety Act framework.	<p>Do not rely on automatic extension. Clauses 18 and 41 provide an independent gateway. To use the Act's powers instead, amend both purpose and relevant-requirement gateways and examine associated confidentiality, procedure, offence and enforcement provisions.</p>
Sections 131 to 143 and Schedule 13: enforcement and penalties	New duties would not automatically become enforceable requirements under that scheme.	<p>No extension selected. Clause 45 expressly engages the Bill's enforcement scheme. If consolidating, extend the enforceable-requirement definition and all necessary notice and penalty references; do not merely copy the maximum fine.</p>
Sections 144 to 148: business disruption	The existing Act permits specified service and access restriction measures subject to conditions.	<p>No extension proposed. This draft retains the Bill's prohibition on whole-service blocking. Importing these powers would be a material policy expansion requiring separate justification and drafting.</p>
Sections 167 and 168: challenges and appeals	Existing routes include Upper Tribunal review and appeals; the Bill uses First-tier Tribunal merits appeals.	<p>Keep jurisdiction distinct under this model. Clause 45(5) adds audit designation appeals to the Bill's route. Any consolidated scheme must expressly choose its appeal standard, forum, interim powers and treatment of parallel decisions.</p>

Existing provision	Relationship to this draft	Amendment or action and status
Section 234: harm and cumulative exposure	Existing concepts expressly include physical or psychological harm and repeated or combined exposure.	No amendment necessary. Clause 31 independently defines serious harm. Do not assume that the definition of harm alone creates a general vulnerable-adult safety duty.

B. Existing Bill: integration and other material information

Matter	Required step or material consequence
Placement and numbering	Insert Part 7, clauses 31 to 45, and Schedule 1 after the existing text. The original clauses 1 to 30 remain numbered as drafted. Before introduction, expand the long title after “duties of social-media service providers” by inserting “; transparency and user control of recommender systems; assessment and mitigation of serious recommendation-related harm”. This is an editorial integration instruction, not an operative power to amend the Act’s own long title.
Purposes and governing principles	Clause 45(2) extends the original purposes to algorithmic transparency, controls and harm prevention. In a consolidated print, incorporate this expansion directly into clause 1.
Provider penalties	The existing clause 20(1)(a) refers only to Part 3. Clause 45(3) expressly extends it; omission would leave direct new duties outside that penalty limb. No separate additional maximum or automatic fine per recommendation is created.
Investigation, injunctions and appeals	Clauses 41 and 45 extend information powers, technical inspection, system-improvement orders and appeals. New duties must not be left to implication. Whole-service blocking remains excluded.
Definitions and territorial scope	New Part follows the Bill’s service and UK-connection definitions, not the Online Safety Act’s full scope, exemptions or Category 1 thresholds. Cross-border enforcement, service of documents and practical recovery against overseas providers need procedural settlement before introduction.
Children and adults	Children retain stronger protection. Vulnerable adults gain systemic risk mitigation without needing to self-diagnose or activate a setting first. A non-personalised feed is not represented as necessarily safe.
Scheduled content	Initial categories are suicide, self-injury and eating disorders. This Schedule is deliberately broader for assessment than a list of illegal material, but requires serious-risk context and protects beneficial discussion.

Matter	Required step or material consequence
Bias	<p>Other subjects should be added only on evidence through affirmative regulations.</p> <p>Duty concerns evidence-based amplification effects, not ideological certification or equal airtime. Publish findings and uncertainty; test for foreseeable effects rather than relying on self-declaration.</p>
Usability	<p>One-action access to the controls hub does not mean every configuration must be completed in one action. Item-level feedback, persistence and tested exposure reduction are distinct requirements.</p>
Data protection	<p>No blanket permission to infer mental health is supplied. Special-category processing, where involved, needs an applicable condition as well as an ordinary lawful basis. Any new special statutory processing condition would require separate justification and drafting; none is assumed here.</p>
Equality and accessibility	<p>Disabled users and persons with relevant vulnerabilities must be included in design testing. A separate equality-impact assessment should examine discriminatory exclusion and accessibility; the draft does not amend equality legislation.</p>
Clinical and safeguarding boundaries	<p>Providers are not required to diagnose users or automatically contact police or relatives. Research and testing must not expose participants to avoidable harm. Crisis-support integration needs competent clinical advice and local availability checks.</p>
Independent assurance	<p>High-risk designation is evidence-based, reviewable and appealable. Provider funding is limited to reasonable, proportionate costs. Auditor competence, conflicts, secure access and publication standards require an implementing code.</p>
Individual redress	<p>Complaints and regulatory enforcement are supplied; no new automatic damages action or guarantee of health outcomes is created. A separate compensation scheme would be an additional policy choice.</p>
Commencement and costs	<p>Clause 45 provides preparatory powers and phased operation linked to publication of necessary standards. OFCOM, Tribunal, audit-market and smaller-provider resources require assessment. Current codes and service categorisation must be checked against the implementation date; their operative status is not certified by this draft.</p>
Parliamentary settlement	<p>Parliamentary Counsel should settle formal consequential amendments, devolved competence and consent, public expenditure, information-sharing gateways and procedural rules. The coordination table identifies the principal interfaces, not a completed line-by-line consequential-amendment Schedule.</p>

Sources for the coordination analysis

The existing statutory interfaces discussed in Table A are [Online Safety Act 2023, sections 9 to 16, 20 to 23, 43 to 49, 61, 71, 77 to 78, 100, 104, 131 to 148, 167 to 168 and 234, and Schedules 4, 8 and 13](#). These are existing statutory interfaces, not authority that the proposed new duties have already been enacted. The draft requires further commencement, consequential-amendment and implementation checking before introduction.