

ANNEX A: GAPS BETWEEN OFCOM'S ANALYSIS OF CAUSES AND IMPACTS OF ONLINE HARM PROPOSED MITIGATIONS: ILLEGAL HARMS AND PROTECTION OF CHILDREN COMBINED

In our response to Ofcom's illegal harms consultation, [we provided a table](#) analysing how far harm arising from the functionalities that it identified in its risk register (volume 2) were mitigated by specific measures in the codes (annex 7). The approach Ofcom takes in its protection of children's consultation is broadly similar to that proposed in the illegal harms consultation - though this caveated by many references throughout the documents that the responses to the latter have not yet been taken into account and further updates will follow. It is not clear, however, whether these will have a material impact on the approach to both sets of codes.

We have carried out the same analysis on the children's consultation as we did previously and updated our table to combine the results from both for ease of reference. As we set out in the introduction to the previous document, we would expect that Ofcom's decisions on which measures to include in their codes of practice would reflect the level of risk threat that the functionalities identified in the risk register pose. We would also reiterate here our acknowledgement that the work that has gone into the risk registers themselves - [volume 3](#) in the children's consultation, [volume 2](#) in the illegal harms - is thorough and analytical. But - with specific reference to the new material - this assessment (again) does not flow through to the mitigation measures in the codes of practice for user-to-user services ([annex 7](#)) and search ([annex 8](#)), which as previously focus primarily on content takedown and measures to deal, ex-post, with primary priority content (PPC), priority content (PC) or non-designated content (NDC). The exception to this is the measures - much publicised in [Ofcom's press material](#) and communications around the launch of the consultation - relating to recommender systems.

This is welcome and goes some way to addressing the scale and impact of harm caused by the recommendation and promotion of PPC, PC or NDC content to children - whether they have searched for, or engaged with, it previously - that is evidenced by Ofcom in its consultation. But we remain concerned that the rules-based nature of the Codes (which is NOT required by the definition of "measures" in the Act¹) - specifying narrow recommended measures rather than describing desired outcomes - and the fact that the Codes are designed as a "safe harbour" (eg if companies follow the measures they will be judged to have complied with their duties under the Act²), means that there is no incentive for

¹Section 236(1) Online Safety Act

² "Services that choose to implement the measures we recommend in Ofcom's Children's Safety Codes will be treated as complying with the relevant children's safety as well as their reporting and complaints duties. This means that Ofcom will not take enforcement action against them for breach of that duty if those measures have been implemented. This is sometimes described as a "safe harbour." However, the Act does not require that service providers adopt

companies to implement mitigating measures to protect children beyond those described in the codes, even if their risk assessment has flagged that their service poses particular risks from other ex ante functionalities (such as design choices). This is particularly notable in relation to the omission of any measures relating to livestreaming - which is mentioned in relation to seven out of the nine types of content in the children's risk profiles; and in relation to two new functions that are covered in the children's consultation: stranger pairing and ephemeral messaging. Furthermore, smaller companies are in many instances exempt from implementing particular mitigating measures due to Ofcom's proportionality analysis.

The following tables provide detailed analysis on the individual functionalities, the number of offences (for the illegal harms codes) or types of content (for the children's codes) where Ofcom identifies that particular functionality is a contributory factor, and the appearance (or not) of mitigating measures relating to this functionality in the codes of practice for user to user and search services for both duties. A summary "at a glance" table is provided for U2U (pages 3-6) and search (p7-8). Supporting tables for user-to-user services (from p9) and search services (pp21-25) provide more detail and extracts from Ofcom's consultation materials. We have divided the measures in both sets of codes into "ex ante" and "ex post", the latter largely applying to measures relating to content moderation and takedown when either illegal content or PPC, PC or NDC has been identified on a service. While we have used the term "ex ante" in relation (generally speaking) to the non-takedown measures, the measures identified are focused on the presence of specific content (either illegal or designated) on the service (or the search functionality enabling users to find it) so are not what we would term "safety by design" measures. These we would classify as biting at a systemic level separate to the nature of the particular types of content (e.g. business model, default settings or measures that are not directed to a particular type of content for eg rebalancing weighting in recommender tools).

the measures set out in the Children's Safety Codes, and service providers may choose to comply with their duties in an alternative way that is proportionate to their circumstances. (Vol 5, para 13.4)

COMPARISON OF RISK REGISTER FUNCTIONALITIES WITH USER-CODE OF PRACTICE MITIGATIONS (Annex 7): SUMMARY TABLE

Functionality	Illegal harms offences	Children's PPC, PC or NDC	Code of practice: ex ante mitigations		Code of practice: ex post mitigations	
	15 in total	9 in total	Illegal harms	Children	Illegal harms	Children
Content: posting, commenting, hyperlinks, including images and video	15	9	Limited to user controls measures (eg muting, blocking): 9A, 9B	Limited to user controls measures (eg muting, blocking, disabling comments): US2, US3	Content moderation & takedown: 4A-F	Content moderation & takedown: CM1-CM7 Limited: Signposting children to support when they a) report content (all services); b) post or repost content (large, risky services); US3, US4
Reposting or forwarding content	5	4	None	None	Limited: reference to "limiting time"	None
Livestream & live audio	9	7	None	None	None	None
Use of hashtags	5	8	None	None	None	None
Editing visual content	9	4	None	None	None	None
Screen capturing or recording	1	2	None	None	None	None
User tagging	5	3	None	None	None	None
User profiles	10	4	Limited to user controls: 9A, 9B	Limited to user controls: US2, US3	None	None

Functionality	Illegal harms offences	Children's PPC, PC or NDC	Code of practice: ex ante mitigations		Code of practice: ex post mitigations	
	15 in total	9 in total	Illegal harms	Children	Illegal harms	Children
User connections	8	8	Limited to default settings, user controls: 9A, 9B	Limited to default settings, user controls: US2, US3	None	None
Stranger pairing	N/A	1	N/A	N/A	None	None
User search	2	1	None	None	None	None
User groups	9	4	None	None	None	
User base profile	3	7	None	Significant measures via age assurance (AA1-6) though no differentiation for age ranges within this	Limited: references in 4E, 5B	None
Recommender systems	11	8	None	Significant new measure (RS1-3) covering PPC and PC, and feedback	Limited: A6 ("limited time"), A9 safety metrics	Not applicable: ex-ante design choice
Group messaging	6	6	None	US1: option to accept or decline an invite to a group chat	None	None
Encrypted messaging	10	3	None	None	None	

Functionality	Illegal harms offences	Children's PPC, PC or NDC	Code of practice: ex ante mitigations		Code of practice: ex post mitigations	
	15 in total	9 in total	Illegal harms	Children	Illegal harms	Children
Direct messaging	15	6	Limited to user controls: 9A, 9B Plus 7A: Default settings for child users where services are high risk for CSAM	Limited to user controls: US2, US3	None	
Ephemeral messaging	N/A	2	N/A	None	N/A	None
Anonymous user profiles	15	5	9C has recommendations re user labelling schemes, but this is only limited to services at risk of fraud or the foreign interference offence	None	None	None
Fake user profiles	13	4	As above 9C	None	None	None
Business model - inc small, fast-growing services; ad revenue	5	3	None	None	None	None
Payment facility	2	0	None		None	
User location	4	1	Included in A7 default settings measures, but only limited to services at high risk of grooming		None	
UGC search facility	3	3	None		None	Limited: Signpost children to support

Functionality	Illegal harms offences	Children's PPC, PC or NDC	Code of practice: ex ante mitigations		Code of practice: ex post mitigations	
	15 in total	9 in total	Illegal harms	Children	Illegal harms	Children
						services when they search for harmful content (high or medium risk): US5
Posting goods or services for sale	7	0	None		None	
Building lists or directories	2	0	None		None	

COMPARISON OF FUNCTIONALITIES WITH SEARCH CODE OF PRACTICE MITIGATIONS (ANNEX 8): SUMMARY TABLE

NB the analysis in of the search functionalities that cause harm is less detailed and presented in a different way to the evidence in the user-to-user sections of both consultations.

Functionality	Illegal harms	Children's PPC, PC or NDC	Code of practice: ex ante mitigations		Code of practice: ex post mitigations	
			Illegal harms	Children	Illegal harms	Children
Typing in searches for illegal / specified content	8	Not defined	Limited: provision of warnings for CSAM searches; and provision of suicide prevention information in relation to suicide/self-harm searches	None	Search moderation & takedown: 4A-F - these measures largely replicate the user-to-user content moderation measures but with 4A applying to deindexing or deranking illegal content. An additional deindexing measure applies to CSAM URLs (4G)	Equivalent as for illegal harms: Measures SM1-7
Ranking	-	N/A	None	None	As above	As above.
Reverse image search	1	Not defined	None	N/A	None	N/A
Search prediction or personalisation	3	Not defined	None	N/A	Limited: requires action when there is a user report that predictive	Limited: offer users means to easily report predictive search

Functionality	Illegal harms	Children's PPC, PC or NDC	Code of practice: ex ante mitigations		Code of practice: ex post mitigations	
					search suggestions are directing users to priority illegal content	suggestions relating to PPC and PC (SD1); provide crisis information in response to searches relating to suicide, self-harm and eating disorders (SD2)
Revenue models	2	Not defined	None	None	None	None
Commercial profile/size	-	Not defined	None	None	None	None
Gen AI/chat bots	-	Not defined	None	None	None	None

COMPARISON OF VOLUME 3 FUNCTIONALITIES WITH CODE OF PRACTICE MITIGATIONS ([ANNEX 7](#)) - USER TO USER SERVICES - FULL TABLE

These tables apply only to the children's consultation measures, rather than providing material to compare with the illegal harms consultation measures (as per the covering RAG-rated table)

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
CONTENT FUNCTIONALITIES			
Posting content	Pornography Suicide and self-harm content Eating disorder content Abuse and hate content Bullying content Violent content Harmful substances content Dangerous stunts and challenge content	Limited US2 (user support) sets out that all U2U services that have user profiles, <u>and</u> certain user interaction functionalities (user connections, posting content and are medium to high-risk of one or more of bullying content, abuse and hate content and violent content allow blocking or muting of other users' accounts. The Government produced its own "best practice" guide for safety by design for platforms that enabled private or public interaction in 2021: https://www.gov.uk/guidance/private-and-public-channels-improve-the-safety-of-your-online-platform	Extensive Content is primarily dealt with in the codes via moderation, with the measures mirroring those in the illegal harms consultation: <ul style="list-style-type: none"> • CM1: swift action • CM2: internal content policies (only for large and multi-risk services) • CM3: performance targets (ditto) • CM4: prioritisation of review of content (ditto) • CM5: resourcing • CM6: moderator training There is an additional, new measure relating to providing materials for volunteer moderators for their roles.

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
			P62: The definition table at the end of the codes says re “content”; “For the avoidance of doubt, comments, titles and descriptions are considered to be ‘content’ within this definition, as are livestreaming videos or audio, and hyperlinks”
Commenting on content	Suicide and self-harm content Eating disorder content Abuse and hate content Bullying content Violent content Dangerous stunts and challenge content	Limited US3 also sets out (for services that meet the same condition as above) that users should be able to disable comments.	Extensive (as per content above)
Hyperlinks - eg use to direct users to more extreme content	Pornography Suicide and self-harm content Eating disorder content Violent content Harmful substances content	None recommended	Extensive (as per content above)
Reposting or forwarding content	Pornography Suicide and self-harm content Bullying content Violent content Dangerous stunts and challenge content	None recommended.	Limited CM4 (prioritisation) refers to the fact that “In setting the policy, the provider should have regard to at least the following: a) the virality of content: <i>the provider should prioritise content for review in a</i>

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
			<i>way which minimises circumstances in which the number of child users encountering a particular item of content that is harmful to children increases exponentially over a period of time” (Annex 7 page 23)</i>
Posting images or videos	Pornography Eating disorder content	None recommended.	Extensive (as per content above) P62: The definition table at the end of the codes says re “content”; “For the avoidance of doubt, comments, titles and descriptions are considered to be ‘content’ within this definition, as are livestreaming videos or audio, and hyperlinks”
Livestream	Suicide and self-harm content Abuse and hate content Violent content Harmful substances content	None recommended NB the government produced its own “best practice” guide to “safety by design” for livestreaming in 2021: https://www.gov.uk/guidance/live-streaming-improve-the-safety-of-your-online-platform	Limited (except as type of “content”) P62: The definition table at the end of the codes says re “content”; “For the avoidance of doubt, comments, titles and descriptions are considered to be ‘content’ within this definition, as are livestreaming videos or audio, and hyperlinks” This does not consider how the functionality of livestreaming is used to

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
			facilitate, encourage or promote the occurrence of the harm (eg particularly in relation to the types of PPC and PC identified) in the first place.
Livestream - Sending messages via livestream	N/A	None recommended	<p>Limited (except as type of “content”)</p> <p>P62: The definition table at the end of the codes says re “content”; “For the avoidance of doubt, comments, titles and descriptions are considered to be ‘content’ within this definition, as are livestreaming videos or audio, and hyperlinks”</p> <p>This does not consider how the functionality of livestreaming is used to facilitate the offences in the first place.</p>
Live audio	N/A	None recommended	<p>Limited (except as type of “content”)</p> <p>P62: The definition table at the end of the codes says re “content”; “For the avoidance of doubt, comments, titles and descriptions are considered to be ‘content’ within this definition, as are livestreaming videos or audio, and hyperlinks”</p>

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
Content tagging - Eg hashtags	Suicide and self-harm content Eating disorder content Violent content Harmful substances content Dangerous stunts and challenge content	None recommended.	None recommended.
Screen capturing or recording	Bullying content Violent content	None recommended	None recommended
USER FUNCTIONALITIES			
User tagging	Pornography Violent content	None recommended.	None recommended.
User profiles	Eating disorder content Abuse and hate content Violent content	Limited US2 (user support) sets out that all U2U services that have user profiles, <u>and</u> certain user interaction functionalities (user connections, posting content and are medium to	None recommended

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
		<p>high-risk of one or more of bullying content, abuse and hate content and violent content allow blocking or muting of other users' accounts.</p> <p>The Government produced its own "best practice" guide for safety by design for platforms that enabled private or public interaction in 2021: https://www.gov.uk/guidance/private-and-public-channels-improve-the-safety-of-your-online-platform</p>	
User connections	Pornography Suicide and self-harm content Eating disorder content Abuse and hate content Bullying content Violent content Dangerous stunts and challenge content	None NB this is a lower level of protection than in the illegal harms codes, where recommendation A7 (only for services at high-risk of grooming, or a large service at medium-risk of grooming) requires that default settings do not include children in network expansion prompts and connection lists	None recommended
User groups	Pornography	Limited	None recommended

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
	Suicide and self-harm content Eating disorder content Violent content	Recommendation US1 requires that all services that have group chats and are medium or high risk of one or more of: pornographic content, eating disorder content, bullying content, abuse and hate content, and violent content; “provide children with an option to accept or decline an invite to a group chat”.	
User base profile	Pornography Suicide and self-harm content Eating disorder content Abuse and hate content Bullying content Violent content Harmful substances content	Extensive Measures AA1-AA5 which set out the age assurance requirements on companies, depending on whether they host PPC and/or PC are the tools by which the concerns identified in the risk profiles re user base (eg the presence of children). However, there are no measures to address the evidence that Ofcom has collected relating to the differential impact of different types of harm on children of different ages. The age assurance recommendations are a blunt tool to	Limited Recommendation CM5 re content moderation says the services needs to take into account “the particular needs of its child user base as identified in its risk assessment, <u>in relation to languages.</u> ” Recommendation UR2 mirrors the illegal harms measure re complaints and says “In designing its complaints processes for relevant complaints, including its reporting tool or function, the provider should have regard to the particular needs of its United Kingdom user base as identified in its children’s risk

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
		<p>either a) prevent children accessing the entire service if its “principal purpose is the hosting or dissemination of one or more kinds of PPC”</p> <p>And, while this is an “ex-ante” measure - in the sense that it comes before the access to content - it does not address the fundamental design choices relating to specific functionalities, identified elsewhere in this table, that can cause harm to children</p>	<p>assessment. This should include the particular needs of: a) children (considering the likely age of the children using that service); and b) disabled people”</p> <p>Neither of these address the way in which the service design might ensure that users identified in the risk assessment might be protected in the first instance from harm.</p>
Stranger pairing (Children’s risk profiles only)	Abuse and hate content	None recommended	None recommended
RECOMMENDER SYSTEMS			
Recommender systems	Pornography Suicide and self-harm content Eating disorder content Abuse and hate content Bullying content Violent content Harmful substances content Dangerous stunts and	Limited There are three recommendations re recommender systems (a functionality that was not considered in the illegal harms codes of practice). These include: RS1 for U2U services that operate a	Limited RS3 for U2U services meeting the two sets of combined conditions for RS1 and RS2 (LH column) <u>and are large</u> , they must “enable children to provide negative feedback on content that is recommended to them”.

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
	challenge content	<p>content recommender system; <u>and</u> are medium or high-risk for at least one kind of PPC, they must “ensure that content likely to be PPC is not recommended to children”</p> <p>RS2 for U2U services that operate a content recommender system; <u>and</u> are medium or high-risk for at least one kind of PC (excluding bullying), they must “ensure that content likely to be PC is reduced in prominence on children’s recommender feeds”.</p> <p>There is no upstream requirement in the code to ensure that services consider the overall design of their recommender systems in the first place, just measures that are related to the content that flow over them.</p>	<p>However, unlike the illegal harms recommendation (A8) that required some services to analyse the safety metrics from tests of its recommender system to “understand if changes to the recommender system would increase the risk of users encountering illegal content”, there are no requirements on services to act on the negative feedback that they might receive as a result of implementing RS3. The recommended measure in the code simply says that “where a child user has signalled negative sentiment towards a specific piece of regulated user-generated content present on any of their recommender feeds that piece of regulated user-generated content, and any other piece of regulated user-generated content that shares significant characteristics with it, is given a low degree of prominence in that child user’s recommender feeds” (Annex 7, p46)</p>
MESSAGING FUNCTIONALITIES			
Group messaging	Pornography Suicide and self-harm content	Limited	None recommended

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
	Eating disorder content Abuse and hate content Bullying content Violent content	US1 puts a requirement on “services likely to be accessed by children where it has a group messaging functionality and there is a medium or high risk of one or more abusive content, bullying content, content inciting hatred, eating disorder content, pornography content and violent content” to “provide children with an option to accept or decline an invite to a group chat.	
Encrypted messaging	Suicide and self-harm content Eating disorder content Violent content	None recommended	None recommended
Direct messaging	Pornography Suicide and self-harm content Eating disorder content Abuse and hate content Bullying content Violent content	Limited US2 (user support) sets out that all U2U services that have user profiles, <u>and</u> certain user interaction functionalities (user connections, posting content and are medium to high-risk of one or more of bullying content, abuse and hate content and violent content allow blocking or muting of other users’ accounts.	None recommended

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
Ephemeral messaging	Bullying content Violent content	None recommended	None recommended
ANONYMOUS/FAKE ACCOUNTS			
Anonymous user profiles and bot accounts	Pornography Eating disorder content Abuse and hate content Bullying content Violent content	None recommended NB the Government produced its own “best practice” guide to “safety by design” for anonymous or multiple account creation in 2021; https://www.gov.uk/guidance/anonymous-or-multiple-account-creation-improve-the-safety-of-your-online-platform	None recommended
Fake Profiles	Pornography Suicide and self-harm content Bullying content	None recommended	None recommended
MISCELLANEOUS			
Business model: • Ad revenue	Pornography Suicide and self-harm content Eating disorder content	None recommended	None recommended
User location	Bullying content	None recommended	None recommended

Functionality	Types of content (PPC, PC or NDC) *content in bold is where the functionality is highlighted in the introductory summaries in the risk register	Code of practice: systemic or ex-ante mitigation?	Code of Practice: recommended ex-post mitigation?
Editing visual media	Suicide and self-harm content Eating disorder content Bullying content	None recommended	None recommended
UGC content searching or filtering	Pornography Suicide and self-harm content Eating disorder content Violent content	None recommended	None recommended

COMPARISON OF FUNCTIONALITIES WITH CODE OF PRACTICE MITIGATIONS ([ANNEX 8](#)) - SEARCH SERVICES - FULL TABLE

The analysis on the functionalities related to user access to illegal content via search services is presented in a different way by Ofcom in volume 3 (as it was in the illegal harms consultation volume 2): a high-level summary narrative that talks about functionality. It is even more high-level in the children's version in that it makes no mention of specific types of PPC, PC or NDC. The number of search-related functionalities are also more limited. This table is included for completeness / comparative purposes with our work on the illegal harms consultation but please see our full response for more detail. NB the Government produced its own "best practice" guide for "safety by design" for search functionality in 2021: <https://www.gov.uk/guidance/search-functionality-improve-the-safety-of-your-online-platform> (It is not referenced by Ofcom.)

Functionality	Types of content (PPC, PC or NDC)	Code of practice: systemic or ex-ante measures?	Code of practice: ex-post measures?
Typing in searches	Yes - but no specific content mentioned	None recommended .	Extensive Content is primarily dealt with in the codes via the search moderation duties Eg: SM1: The provider should have systems or processes designed to downrank or blur search content in relation to PPC, PC and NDC SM2: when a user is believed to be a child, filter identified PPC out of their search results through a safe search setting. Users believed to be a child should not be able to turn this setting off. (large and general search services) SM3: internal content policies (large and multi-risk) SM4: performance targets (ditto)

Functionality	Types of content (PPC, PC or NDC)	Code of practice: systemic or ex-ante measures?	Code of practice: ex-post measures?
			SM5: prioritization for review (ditto) SM6: resourcing (ditto) SM7: training (ditto)
Search prediction or personalisation 6T.37 “It is reasonable to assume that these functionalities can increase the risk of accessing illegal content amounting to a range of offences, unless effective mitigations are in place to prevent this, or indexed content is blocked.”	Yes - but no specific content mentioned	None recommended	Limited All large general search services need to “offer users a means to easily report predictive search suggestions relation to PPC and PC” (SD1) They must also provide “crisis prevention information in response to known PPC search request regarding suicide, self-harm and eating disorderse
Revenue models - ad-based models .	Yes - but no specific content mentioned.	None recommended	None recommended
Commercial profile/size	Yes - but no specific content mentioned	None recommended	Limited Search services that are multi-risk for content provided to children must “provide age-appropriate user support materials”.