EXHIBIT TEN

a/c priv - reflects advice of EEA-qualified external counsel and attorney-client communications | privileged and confidential | do not share

DEPRECATED: Please use [AIR531] Evolving Pressure on Potentially Pirated Content (a/c priv)

[AIR531] EU AI Act Counterfeit and Piracy Watch List Doc (a/c priv)

Decision Status: Decision doc WIP
Control Center Decision ID: AIR531
Date: Dec 17, 2024
POC: Zef RosnBrick, Anna Grosul, James Whymark, Elizabeth L. Kendall, Ahuva Goldstand, Ania Helseth, Malikat Rufai, Rob Arcamona, *please @ your name when you contribute*

TLDR

The EU AI Act "put in place a policy to comply with Union law on copyright and related rights" requiring Meta to "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model".

Additionally, EU AI provided sub-measure 4.5 in their (first) draft Code of Practice, which is a voluntary measure we can choose to sign, which provides a strongly suggested manner by which we can comply with their introduced policy. Sub-measure 4.5 requires "Signatories [to] take reasonable measures to exclude pirated sources from their crawling activities, such as by excluding websites listed in the Commission Counterfeit and Piracy Watch List" (more context). On this watchlist we have Libgen and it implies Anna's Archive which are two of the most valuable data sources for text and image modalities that are used to train Llama foundational models. We must now decide how to respond to these new policies.

Key Risks:

- **EU AI Act Compliance**: Meta must establish a policy to comply with EU copyright laws. The draft Code of Practice, particularly Sub-measure 4.5, suggests excluding pirated sources like Libgen and Anna's Archive from crawling activities. Some of these sources are crucial for training Llama foundational models.
- **Voluntary AI Act Code of Practice:** Although the Code is voluntary and still in draft form, there is a strong expectation for AI providers to sign it. It will also serve as a compliance benchmark by the AI Office.
- **Transparency Obligations:** The EU AI Act is likely to require Meta to disclose the use of these sources, with details to be confirmed in January.
- **Policy risks**: major reputational risk that will impact how LLAMA model is perceived overall by policy makers and likely to undermine the open source narrative and we should expect it to affect the willingness of policymakers and influential companies to ally with us in public, a core part of our policy strategy. We can expect legislation to come into place that will remove opt-out as insufficient protection for rightholders.
- **Litigation** Redacted Redacted
- **External scrutiny:** Public figures, such as authors, actors, and musicians, view the use of this data as theft and are increasingly vocal in their opposition. We can expect a public backlash driven by influential figures.

Please note: the code of practice is still in draft format and is not yet finalized and the EU AI act will require a series of decisions that we will litigate in future L1s (see appendix).

L1 Inform Decision

Inform: We are not making any changes to Llama-4 Flagship in response to sub-measure 4.5.

After carefully analyzing all of our short term options, we have decided that the only viable path forward is to keep the status quo in terms of crawling and using Anna's Archive and Libgen for training the Llama-4 Flagship. Our training timelines are too tight, with training starting on January 20, to implement any significant changes based on our longer term outlined options. Additionally, the product impact of removing this data is too significant, especially since the provisions in the code are not yet finalized and may change.

While our view is that we have arguments to demonstrate compliance with the current AI Act regulatory requirements, there will remain [redacted] legal and policy [redacted] [redacted]

Commented [1]: I thought we were going to say something like "we recommend exploring if there are alternatives for the future" with the understanding that it would take a lot of time/not pinning anyone down to a particular timeframe or even outcome. Or can that be policy's recommendation? I don't want to have this recommendation and just say all the risks are high, it seems like we're not working on any way forward that's useful (again, consistent with the point that no change is happening right now, etc.)

Definitions

- **Synthetic Generation:** Synthetic generation refers to the process of creating new, artificial data that resembles existing data but is not derived from it. This involves generating entirely new data points, samples, or instances that are similar in structure and characteristics to the original data.
- **Synthetic Transformation:** Transformation refers to the process of modifying or altering existing data into a new form while preserving its essential characteristics.
- **Crawling:** the act of collecting 3P publicly available data
- **Data use**: the act of using data in training production models.

Context

Under the EU AI Act, Meta is required to comply with the following requirements:

> *"Article 53 - Providers of General Purpose AI-Models shall:*
>
> *(c) put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790;*
>
> *(d) draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office."*

These requirements apply to any General Purpose AI Model put on the market in the EU from August 2, 2025 (models already on the market before this point would only require a policy/transparency by August 2027).

The AI Office is developing a voluntary "Code of Practice" to guide compliance with the AI Act. While not mandatory, there is a strong expectation for AI providers to adopt it, as it outlines compliance expectations for general-purpose AI models. The draft Code of Practice includes guidelines for a compliant 'copyright policy' and addresses content crawling from 'pirated sources.' If Meta signs it, these measures will apply to any model released from August 2, 2025.

> *Sub-Measure 4.5. No crawling of piracy websites*
>
> *Signatories will take reasonable measures to exclude pirated sources from their crawling activities, such as by excluding websites listed in the Commission Counterfeit and Piracy Watch List. Signatories are also encouraged to comply with analogous exclusion lists published by relevant public authorities in the jurisdictions where they are established.*

Current State

The draft "Code of Practice" is not yet finalized and is expected to be completed in May 2025. Although the terms may change, it's likely that key provisions will remain largely unchanged.

Meta must provide a transparency summary based on a template from the AI Office, which is expected to be received on **January 13, 2025**. We currently do not have specific details of what will be expected to be contained in a transparency summary although the language of the Act (and recitals) provide some limited detail:

- The disclosure must be "comprehensive in scope instead of technically detailed".
- It is intended to facilitate parties with "legitimate interests" including copyright holders, to exercise their EU rights.
- The disclosures can take into account trade secrets and confidential business information.
- An example of how this may be achieved is through "listing main data collections, such as large private or public databases or data archives".
- Alternatively, the law provides that a "narrative explanation about other data sources used" could be used.

Impact/Risk

We have historically used and crawled data from the Commission Counterfeit and Piracy Watch List, such as Library Genesis, for training. This data makes up a large portion of our high quality token count for text and multimodal and significantly enhances model performance (see here). However, due to transparency requirements, Meta may need to disclose the use of these sources for training GPAI models in the EU after August 2, 2025. This could affect the Llama4 Flagship, set for release in September 2025. Since training will begin within a few weeks, changes to the use of Library Genesis or similar datasets are not feasible.

Key Risks:

- **EU AI Act Compliance**: Meta must establish a policy to comply with EU copyright laws. The draft Code of Practice, particularly Sub-measure 4.5, suggests excluding pirated sources like Libgen and Anna's Archive from crawling activities. Some of these sources are crucial for training Llama foundational models.
- **Voluntary Code of Practice:** Although the Code is voluntary and still in draft form, there is a strong expectation for providers to sign it. It may also serve as a compliance benchmark by the AI Office.
- **Transparency Obligations:** The EU AI Act may require Meta to disclose the use of these sources, with details to be confirmed in January.
- **Policy risks**: major reputational risk that will impact how LLAMA model is perceived overall by policy makers and likely to undermine the open source narrative and we should expect it to affect the willingness of policymakers and influential companies to ally with us in public, a core part of our policy strategy. We can expect legislation to come into place that will remove opt-out as insufficient protection for rightholders. We can also expect legislation to be passed prohibiting training on piracy sites, as they may be defined.
- **Litigation** Redacted Redacted
- **External scrutiny:** Public figures, such as authors, actors, and musicians, view the use of this data as theft and are increasingly vocal in their opposition. We can expect a public backlash driven by influential figures. We have strategic partnerships with content providers.

This document **informs** on a decision to make **no immediate changes** to the use or crawling of datasets from the Watch List for Llama4 Flagship training, while outlining the significantly increased legal and policy risks. Additionally in the Appendix we outline a series of outstanding decisions we will need to make as a result of or related to these increased risks this Act.

Please note: Anna's Archive is not explicitly mentioned in the Piracy Watch List but we will consider it as part of product impact as Libgen is a part of Anna's Archive and it will likely be associated with this list.

Product Context

TLDR: Libgen and Anna's Archive (listed as a similar pirated source in the US privacy watchlist) make up a significant proportion of our **Llama 4** datamix and if removed will significantly damage our model performance.

For **Text Llama 4** datamix:

- 100% of Fiction Books
- 100% Science Books
- 100% Multilingual Books
- 82% of Papers

And for **multimodal Llama 4 Flash 17B data** mix, nearly 21% of total tokens are contributed by PDF data out of which 69% of PDF tokens are contributed by Libgen and Anna's archive.

Removing these sources from our datamix would have a huge impact on our performance, specifically for reasoning benchmarks like GPQA, MMLU, etc, where we are already behind competition like o1 and are actively trying to catch up. Additionally, obtaining all the same data from other sources like licensing or crawling is extremely challenging if not impossible.

Performance regression

Ref Doc: Value of Anna's Archive & Libgen

Experiment Set up: 2024-12-02: 60T datamix 1013 baseline vs. 60T without Libgen & Anna's Archive (N6278652) and remove Libgen & Anna's archive related datasets and give weights back to knowledge-web.

Result summary:

- [redacted]



Possible options to fill the gap

Below is a list of **possible options** that we need to explore **in parallel** in order to try and fill the data gap in the event that Anna's Archive and Libgen are no longer available to us.

⚠️ **Important note**: none of these options are viable for our Llama-4 Flagship timeframes, we will need significant investment and lead time to determine if any of these options are even possible.

Option	Description	Tech Risk Profile	Est. % to close gap	AI Risk Decisions Required
1. Synthetic Data Transformation	Leverage these data sources to **transform** them into net new data sources that we use for model training	high *depending on the solution needed*	**100%** This % can change based on what is required	decide: can use this mechanism and what is transformed enough to be considered net new?
2.Synthetic Data Generation	Leverage these data sources to **generate** net new data sources that we use for model training	very high	**0-100%** *Nobody knows how to do this at this scale and this will be very GPU*	decide: can we use this mechanism and if generated content is considered net new ?

			costly	
3.Licensing as many publishers as possible	License publishers that have content that is available in Anna's Archive or libgen.	medium	**<5% (non starter)** • *The data does not exist at this scale through licensing* • *need to have unlimited funds to license every publisher (ea. around $20-$50M)* • *have 1-2 years of lead time to close deals*	n/a
4.Directed Crawling	Build mechanisms to specifically search for sites that contain similar content that is contained in Anna's Archive and Libgen	medium	**<5% (non starter)** The content in Anna's Archive and Libgen is around 98% unique to what we already crawled on the web. See here	Decide: how we should interpret the definition of "pirated content".

Llama-4 Key Dates

Llama 4 Model	Training Start	Release
Flash 17B	10/8/2024	Feb/March 2025
Flash 288B	10/14/2024	Internal
Flash 288B Distilled	12/20/2024	April/May 2025
Flagship	1/20/2025	July - September 2025

Privacy / Risk Considerations

HOW TO: 1-2 sentences describing key privacy / risk axes to consider.

Commented [2] [redacted]@meta.com
Assigned to [redacted]@meta.com

Competitor Approaches

We will likely not have confirmed signals on how our competitors are approaching use of datasets named on the EU Commission Counterfeit and Piracy Watch List or similar sources. However, based on statements companies have made as well as their responses thus far to the draft code of practice, we can make educated guesses about their positions.

OpenAI

As of OpenAI from June 4, 2024 publicly announced that they

> "...affirmatively **avoid crawling** websites that we know employ paywalls for access to their content, websites that primarily aggregate personally identifiable information, and websites that have content that violates our policies. We **exclude sites that are known to engage in IP infringement**, such as the "notorious markets" identified annually by the Office of the U.S. Trade Representative. We also exclude domains that have been flagged to us by copyright owners and their representatives as hosting infringing content."

OpenAI specifically calls out that they avoid crawling but seemingly intentionally, make no mention on how they "use" (train on) the data that they have previously collected from these sites before they made this committment.

Additionally, many competitors publicly listed books3 in their training data and which subsequent lawsuits are in process. However, OpenAI mentions a non-standard "books2" which is rumored to be libgen.

Google

Google shares Meta concerns on the draft provision of the code of practice and voiced that in their own submissions. We have partnered also via the AI coalition to voice these concerns more broadly to the policy makers (regarding unclear process of creating of the watchlist).

MSFT

Microsoft is implementing new methodologies and options for webmasters to control the usage of their content in Bing Chat.

"We are pleased to share that we have built on standard controls that webmasters use today to control indexing and snippet length on Bing in order to empower publishers to make choices about use of their content in Bing Chat and for training Microsoft's generative AI foundation models."

Inform: We are not making any changes to Llama-4 Flagship in response to sub-measure 4.5.

After carefully analyzing all of our short term options, we have decided that the only viable path forward is to keep the status quo in terms of crawling and using Anna's Archive and Libgen for training the Llama-4 Flagship. Our training timelines are too tight, with training starting on January 20, to implement any significant changes based on our longer term outlined options. Additionally, the potential product impact of removing this data is too significant, especially since the provisions in the code are not yet finalized and may change.

While our view is that we have arguments to demonstrate compliance with the current AI Act regulatory requirements, there will remain l Redacted legal and policy **Redacted**

Argument for Position in relation to the use of these sources in Llama4 flagship training

We will continue to crawl and use datasets mentioned on the EU Commission Counterfeit and Piracy Watch List for Llama4 Flagship training, based on the following considerations:

- The Code is not finalized, and its specific wording may change further.
- We have reasonable legal arguments that:
  - The Code references crawling these sources but does not specify requirements for usage.
  - The Code is prospective in application, so prior crawling should not be impacted.
- The Code is voluntary, and Meta is not required to sign up. However, there is a strong expectation for AI providers to adopt it, as it represents the only guidance from the AI Office on compliance with general-purpose AI model requirements in the Act.
- Meta uses this content for training in the US and so takes the position that such use is permitted on the basis of 'fair use'. Redacted

Redacted

Decision Risks <WIP>

Key Considerations:

- **Code of Practice:** The Code is still in draft form, with two more iterations expected. Ou expectation is that this type of measure will remain in the Code although there is material risk that the specific terms could become more restrictive before the Code is finalised in May.
- **Transparency Template:** We will receive the first draft in mid-January, which will guide how we communicate about datasets used in training.
- **EU Commission Counterfeit and Piracy Watch List Update:** The list is updated approximately every two years, with the next update planned for **2025 Q2.**
- **Reputation risk:** Many of the outlined risks, including regulatory investigations, regulatory changes, policy maker pressure, partner backlash, and litigation, pose significant reputation and comms risks to the company as they will be publicly scrutinized in the press and public sphere.

Type	Risk	Risk Severity (WIP)
Regulatory POC: Zef RosnBrick	**Transparency risks** - The AI Act requires GPAI providers to provide the AI Office with a transparency report on all models released onto the EU Market after 2 August 2025. We do not currently know the scope or detail required of such a report but will have a draft template and Code from the AI office in mid-Jan. There are material risks that Meta would be required to disclose that it has trained on content from these sources which would result in Redacted	Very High
	Prospective vs retrospective application - Our view is that we can take the position that the AI Act applies prospectively and that activities undertaken before the AI Act or the Code is effective are not impacted. The Act however applies in relation to GPAI models when they are put on the market in the EU. We will therefore likely get RFIs and greater scrutiny from the AI Office in relation to Llama4 if it becomes clear that Meta has trained on these sources.	Very High

Formatted: Space Before: 0 pt

Commented [3]: [redacted]@meta.com

Commented [4]: @[redacted]@meta.com [redacted]@meta.com [redacted]@meta.com @[redacted]@meta.com can you guys help finish this table please and add risk severity please

	Redacted legal Redacted **Redacted**	Redacted
	Regulatory investigation - As we do not have a clear definition of a compliant 'copyright policy' we can expect significant regulatory scrutiny where we are continuing to train a GPAI model on these sources that will be released onto the EU market. The AI Office has significant powers to investigate and can require models to be removed from the market or issue fines for non-compliance (up to 3% of global turnover).	Very High
Policy [WIP] POC: Elizabeth L. Kendall	**Messaging**: We will likely be asked by policymakers if we are crawling listed sites, in contexts in which we must answer. OpenAI will be stating that it is not crawling listed sites, as it has in the past. Our opening message is "we don't break paywalls and we respect robots.txt, and we crawl the open internet, consistent with U.S. copyright law," but then we would have to fall back to confirm that we crawl listed sites or be silent, permitting conclusions to be drawn. • [We can work with comms on an alternative narrative, but would need something concrete related to training.]	High
	[WIP Policymakers/Law Change If it becomes clear that we are training on the Listed sites, even if we are not legally precluded, we can expect policymakers and rightsholders to take negative actions toward us in the EU and possibly in the US, as there is close communication between rightsholders in these jurisdictions. Potential actions could include: • Longer term: a. Introduction of specific regulation to intervene on behalf of rightholders	Very High

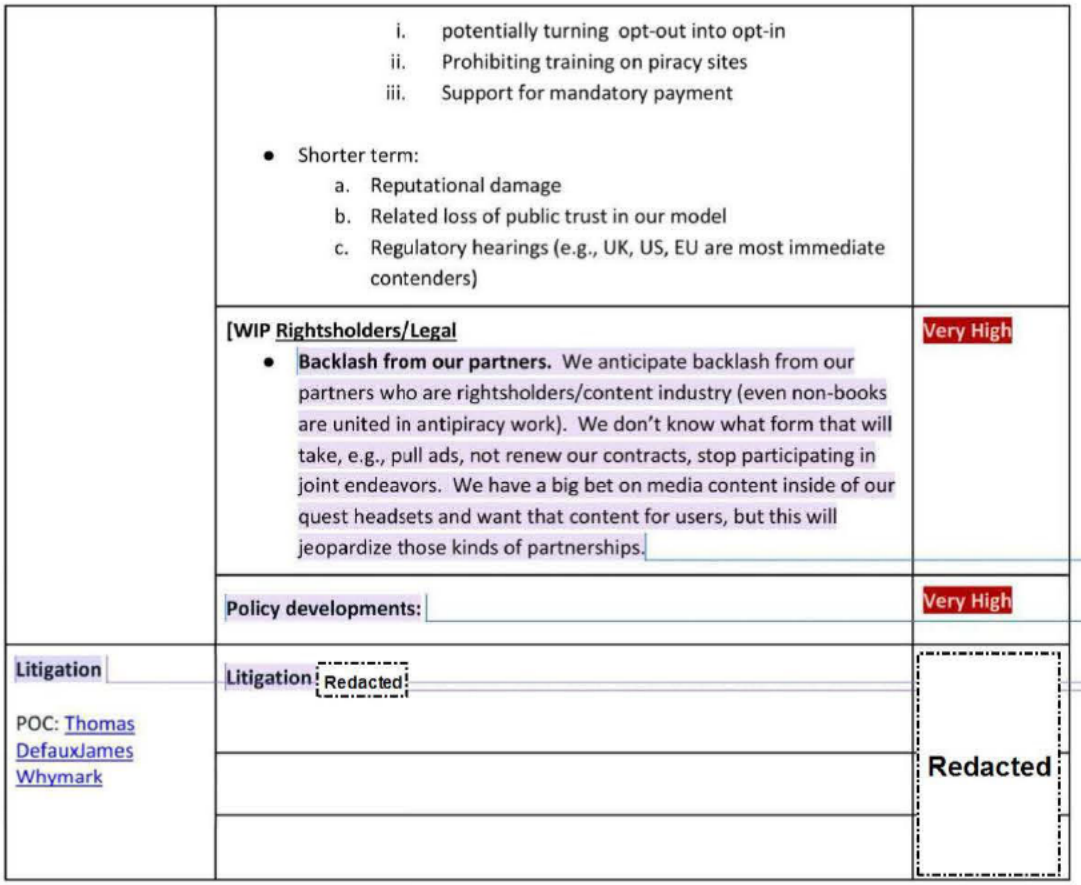
Commented [5]: [redacted]@meta.com @[redacted]@meta.com [redacted]@meta.com to fix

Commented [6]: Some of this has been converted from a previous doc and no longer fits the format

Commented [7]: Should we loop in Comms now?

Commented [8]: We can only use our general narrative, not sure if Open AI here is of relevance. Please also note the announcement of Youtube.

Commented [9]: I'm not certain what you mean, can you add?



	i. potentially turning opt-out into opt-in ii. Prohibiting training on piracy sites iii. Support for mandatory payment • Shorter term: a. Reputational damage b. Related loss of public trust in our model c. Regulatory hearings (e.g., UK, US, EU are most immediate contenders)	
	[WIP Rightsholders/Legal • **Backlash from our partners.** We anticipate backlash from our partners who are rightsholders/content industry (even non-books are united in antipiracy work). We don't know what form that will take, e.g., pull ads, not renew our contracts, stop participating in joint endeavors. We have a big bet on media content inside of our quest headsets and want that content for users, but this will jeopardize those kinds of partnerships.	Very High
	Policy developments:	Very High
Litigation POC: Thomas DefauxJames Whymark	Litigation Redacted	Redacted

Commented [10]: How does this relate to book data?

Commented [11]: Use of piracy sites will be perceived as condoning piracy, which all sectors are sensitive about.

Commented [12]: @[redacted]@meta.com _Assigned to [redacted]all@meta.com_

Commented [13]: isn't that tab repetitive to the one above?

Commented [14]: I don't know what different ground this is supposed to cover

Commented [15]: [redacted]@fb.com @[redacted]meta.com to add

Commented [16]: [redacted]@fb.com _Assigned to [redacted]@fb.com_

Appendix

Future Decisions

#	Decision	Context	Deadline
1	What data should we exclude from crawling going forward?	Focused on whether we can crawl and/or train on datasets named in the EU Commission Counterfeit and Piracy Watch List going forward - Llama 5+.	
2	Which other watch lists should we use to exclude from crawling?	Considers whether to expand the scope of Decision to other, similar watch lists to hedge for future risks.	
3	How can we use data we are excluding from crawling in the future?	Focuses on how else we may be able to use data, and to what extent, from excluded datasets for other purposes, such as synthetic generation or transformation.	
4	How should we interpret the definition of "pirated content"?	Watch Lists generally refer to two concepts when describing content that should be excluded from use: • **Websites:** Specific URLs that serve a dataset, e.g., Libgen.onl, and • **Datasets:** The specific collection of data which is considered pirated or counterfeit, e.g., "libgen." Focus on determining if we can use content from "pirated sites" found elsewhere on the web.	

Commented [17]: @[redacted]@meta.com [redacted]@meta.com @[redacted]@meta.com [redacted]@meta.com [redacted]@meta.com what else is missing?

Commented [18]: Optimizing opt-out via rights manager..? We need sth to offer rightholders in the back pocket.

Commented [19]: @[redacted]@meta.com can we discuss target dates here?

5	How are we interpreting what a general purpose model is?	Focus on defining a definition of "general purpose model" - i.e. is it limited to OSS?	

Counterfeit and Piracy Watch List Context

The 'Counterfeit and Piracy Watch List' is not an exhaustive list of the reported services and marketplaces and does not purport to make findings of legal violations. Neither does it provide the European Commission services' analysis of the state of protection and enforcement of IPR in the country or countries concerned. The European Commission aims to identify online service providers and physical marketplaces located outside the European Union that are reported to engage in, facilitate or benefit from substantial IPR infringements, and in particular piracy and counterfeiting.

The latest EC Watch List is dated from December 2022. We are expecting an updated list in March 2025. The analysis attached to the watchlist finds that "*Around 35% of digital related public conversations on social media could possibly relate to piracy, with film and music piracy being the areas most discussed, especially on Reddit and Twitter.*" **Potential copyright infringements in the context of the metaverse** is also mentioned, and "may require further monitoring in the future." Social media platforms were also reported as facilitating the sale of counterfeit goods. However, these platforms also demonstrate better efforts to comply with the 2018 Recommendations on measures to effectively tackle illegal content online.

List of online service providers offering or facilitating access to copyright-protected content:

- Mega.nz/.io
- Uptobox.com/Uptostream.com
- Rapidgator
- Uploaded.net
- Dbree.org
- Y2mate
- Savefrom
- Flvto and 2conv
- Snappea.com

- Fmovies/Bmovies/Bflix
- Seasonvar
- Rlsbb
- Rezka
- The Pirate Bay
- Rarbg
- Rutracker
- 1337x
- Music Bazaar
- Sci-hub.io
- Library Genesis
- IPTV Smarters
- EVPAD
- Shabakaty
- DDoS-Guard.net
- Private Layer
- Amarutu/Koddos
- Istq
- Host Palace
- BestBuyIPTV
- King365tv.com
- VolkaIPTV.com
- V Kontakte (*Note – the only social media platform included in the list*)
- 2embed.ru
- Fembed.com
- Shopee
- Dhgate
- Tiu.ru, Deal.by, Satu.kz
- Tokopedia

Privacy Considerations

HOW TO: *This section should objectively frame up the privacy risks and other adjacent privacy considerations to consider as they relate to this proposal. It is not the place for the privacy, legal or policy teams to make their case for why they want one option over another. Options are not introduced until the stoplight chart, where the team has an opportunity to build their case.*

Commitments

<add commitments> (e.g., "We don't use your app and website data if you opt out of this control". In this context, we've defined app and website data as x,y,z.)

Precedent: Key Historical CPE Decisions and Outcomes

- <add past relevant summaries>
  - <add Name (Date)>
  - Analysis/Rule: <add>
  - Risks: <add>
  - Mitigations: <add>

Privacy Risks

HOW TO: *Outline the residual privacy risks - those NOT already addressed with agreed upon mitigations (the full list of which should be in the Appendix). Risks should be measured by potential severity and potential likelihood of something bad happening with consequences, i.e., enforcement, damages or fines.*

- <add unmitigated risk(s) here>
  - Example:
    - "High likelihood of regulatory inquiry:
      - a) Moderate possibility that regulators will view this bulk connect as an unfair or deceptive practice and investigate whether users were given sufficient disclosure / transparency / control, even with clarified/deprecated commitments.
      - b) FTC may view this as a violation of FTC Act and Section 1 of consent order prohibiting misrepresentations.
      - c) This could also feed into Consumer Financial Protection Bureau (CFPB) arguments that Meta "poses risk to consumers," increasing chances of supervisory designation. If a federal enforcement investigation is opened,

regulatory investigation risk becomes high as regulators could expand into other data use issues, where we could expect fines at 2% of annual revenue or worse."

Note: We will not be discussing all mitigations in this escalation. The focus of this escalation is on the unmitigated risk highlighted above. Re-opening any agreed upon mitigations may trigger a new escalation.

Additional Legal narrative

1. Litigation Redacted

Redacted

Policy Considerations

The policy environment is characterized by a delicate balance between embracing the potential of Artificial Intelligence (AI) and addressing concerns from rightholders, particularly news publishers and book publishers. On one hand, the EU institutions are eager to harness the benefits of AI, as reflected in the Draghi report and the mission letters of the incoming EU Commission. On the other hand, rightholders are pushing for stricter scrutiny over AI training, citing concerns that:

1. Opt-outs are not respected: AI model providers often disregard the opt-outs of EU rightholders.

2. Pirated sites are used: AI models are frequently trained on pirated websites to circumvent opt-outs.
3. Inadequate robot.txt usage: AI model providers rely solely on robot.txt, which is not suitable for all types of content under opt-out.

To alleviate some of these concerns, the latest draft of the code of practice on the AI Act includes provisions that aim to prohibit crawling pirated websites listed on the IP Watchlist: This measure aims to prevent AI model providers from training on pirated content.

Policy recommendation would be to undermine the process of building the list, the unclear criteria for inclusion and lack of transparency around the final list of nominated websites, resulting in legitimate companies such as Cloudflare, being nominated next to companies engaged only in the pirated activity.

- We would argue that the process is not transparent: The process of building the list should be subject to relevant checks and balances to guarantee that nominations are:
    - Relevant to piracy and counterfeits: Nominations should focus on issues related to intellectual property (IP) infringement, rather than other matters - we see multiple nominations focusing on misinformation, hate speech and other types of content that have little to do with IP
    - Reflect actual IP concerns: nominations often serve as a method to attack your competitors and there are insufficient checks and balances to address it

2. Consider alternative solutions for addressing infringing content: Instead of relying solely on the IP Watchlist, it may be more effective to:
    - Engage search engines in taking down infringing content: Search engines could play a more active role in removing content that infringes on third-party rights, helping to reduce the spread of pirated or counterfeit materials online.

Graveyard

To Dos:

- **L1 Review Date:** Friday, Dec 20, 2024
- Context on the differences between the EU Watch List and other watch lists we may include
- Additional technical context on Synthetic Generation and Transformation
- Input on Decisions 1, 2, 3

Decision 1: What data should we exclude from crawling going forward?

Decision 1 is focused on whether we can crawl or train on datasets named in the EU Commission Counterfeit and Piracy Watch List. Decision 2 considers whether to expand the scope of Decision to other, similar watch lists. Decision 3 focuses on how else we may be able to use data from excluded datasets for other purposes, such as synthetic generation or transformation.

Watch Lists generally refer to two concepts when describing content that should be excluded from use:

- **Websites:** Specific URLs that serve a dataset, e.g., Libgen.onl, and
- **Datasets:** The specific collection of data which is considered pirated or counterfeit, e.g., "libgen."

More broadly, we can also consider whether an individual piece of content is contained in a dataset, e.g., whether or not the first Harry Potter book is contained in LibGen. This is not mentioned explicitly in Sub-measure 4.5 or in the EU Commission Counterfeit and Piracy Watch List, though it is important when considering intellectual property risk more broadly.

Our options for meeting the requirements of Sub-measure 4.5 revolve around the methodology we use for determining if a piece of data or dataset should be excluded from crawling.

Stoplight/Functional Alignment Chart

Key: green = acceptable levels of risk or impact, yellow = carries significant risk or product impact, red = high risk or impact, would continue to escalate for higher decision

	Public Narrative	Product Anna Grosul	Privacy Zef RosnBrick	Legal Ahuva Goldstand, James Whymark, Becca Felsenthal	Policy Ania Helseth, Elizabeth L. Kendall	Comms Toby Partlett,Ben Walters
Option 1. Status Quo. Continue to crawl websites named in the EU Commissions Counterfeit and Piracy Watch List. *Note: This would likely imply that we do not sign the EU AI Act Code of Practice, and may have other implications which we would need to explore in a separate decision.* Crawl (going forward): ✅ Yes						
Option 2. Exclude websites named in the EU Commission Counterfeit and Piracy Watch List. Use the specific URLs mentioned in the watch list to exclude content from crawling: • Disallowed: LibGen downloaded from Libgen.onl	Meta avoids crawling websites named in the Commission's Counterfeit and Piracy Watch List.					

	Public Narrative	Product Anna Grosul	Privacy Zef RosnBrick	Legal Ahuva Goldstand, James Whymark, Becca Felsenthal	Policy Ania Helseth, Elizabeth L. Kendall	Comms Toby Partlett,Ben Walters
• Allowed: LibGen downloaded from another url not included on the watch list. Crawl (going forward): ❌ No						
Option 3. Exclude websites and datasets named in the EU Commission Counterfeit and Piracy Watch List. Use both specific URLs mentioned in the watch list and datasets to exclude content from crawling: • Disallowed: LibGen downloaded from Libgen.onl • Disallowed: LibGen downloaded from another url not included on the watch list. • Allowed: A specific book included in LibGen, but downloaded not as part of the dataset, and from a website not named in the watch list. Crawl (going forward): ❌ No	Meta avoids crawling websites and datasets named in the Commission's Counterfeit and Piracy Watch List.					
Option 4. Exclude content found in websites and datasets named in the EU Commission Counterfeit and Piracy Watch List. Use both specific URLs mentioned in the watch	Meta avoids crawling content contained in datasets named in the Commission's Counterfeit and Piracy Watch List.					

	Public Narrative	Product Anna Grosul	Privacy Zef RosnBrick	Legal Ahuva Goldstand, James Whymark, Becca Felsenthal	Policy Ania Helseth, Elizabeth L. Kendall	Comms Toby Partlett,Ben Walters
list and datasets to exclude content from crawling. Also exclude content from other sources which is known to be contained in a dataset named in the watch list. • Disallowed: LibGen downloaded from Libgen.onl • Disallowed: LibGen downloaded from another url not included on the watch list. • Disallowed: A specific book included in LibGen, but downloaded not as part of the dataset, and from a website not named in the watch list. Crawl (going forward): ❌ No						

Decision 2: Which other watch lists should we use to exclude from crawling?

Sub-measure 4.5 encourages signatories "to comply with analogous exclusion lists published by relevant public authorities in the jurisdictions where they are established," and OpenAI specifically mentions the list of "notorious markets" identified annually by the Office of the U.S. Trade Representative. This decision considers whether we should include watch lists beyond the EU Commission Counterfeit and Piracy Watch List in Decision 1.

Todo: Differences between EU Watch List of USTR

Commented [20] [redacted]@meta.com
Assigned to [redacted]@meta.com

Note: If we choose to include additional datasets, we will expand the Public Narrative as needed.

	Product Anna Grosul	Privacy Zef RosnBrick	Legal Ahuva Goldstand, James Whymark, Becca Felsenthal	Policy Ania Helseth, Elizabeth L. Kendall	Comms Toby Partlett,Ben Walters
Option 1. Exclude based on the EU Commission Counterfeit and Piracy Watch List only. Use websites and datasets (depending on the option chosen for Decision 1) solely from the • EU Commission Counterfeit and Piracy Watch List to determine content to be excluded.					
Option 2. Exclude based on the EU Commission Counterfeit and Piracy Watch List and the Notorious Markets list identified by the USTR. Use websites and datasets (depending on the option chosen for Decision 1) from both the • EU Commission Counterfeit and Piracy Watch List • USTR Notorious Markets list to determine content to be excluded.					

Decision 3: How can we use data we are excluding from crawling in the future?

Data found in LibGen is both highly valuable to model training (todo: datapoint) and unique to other data sources (approximately 98% of data found in LibGen is not found elsewhere in crawled data), and while the current draft of the code of practice specifies "crawling" as opposed to use for training, the transparency template may require us to disclose the use of potentially pirated content in training.

Moreover, indirectly utilizing these already crawled datasets may require novel techniques, having potentially long lead times, and yield different estimated likelihoods of success. To that end, this decision considers the extent to which we can use datasets which we've already crawled, but will not in the future based on Decisions 1 and 2.

Synthetic Generation

Synthetic generation refers to the process of creating new, artificial data that resembles existing data but is not derived from it. This involves generating entirely new data points, samples, or instances that are similar in structure and characteristics to the original data.

Synthetic Transformation

Transformation refers to the process of modifying or altering existing data into a new form while preserving its essential characteristics.

Commented [21]: @[redacted]@meta.com can you help us add more details on generation and transformation? I think in particular some context on how different the synthetic content would or could be in each case will be useful to informing the risk. _Assigned to [redacted]@meta.com_

Note: The options in Decision 3 are not mutually exclusive. As long as we do not decide that we cannot use data excluded in Decisions 1 and 2 for any purpose going forward, we can choose any subset of Options 2-4.

	Product Anna Grosul	Privacy Zef RosnBrick	Legal Ahuva Goldstand, James Whymark, Becca Felsenthal	Policy Ania Helseth, Elizabeth L. Kendall	Comms Toby Partlett ,Ben Walters
Option 1. Datasets which will be excluded in Decisions 1 and 2 cannot be used for any purpose going forward.					
Option 2. Datasets which will be excluded in Decisions 1 and 2 can be used for Synthetic Generation. Likelihood of successfully replacing LibGen data: 20%					
Option 3. Datasets which will be excluded in Decisions 1 and 2 can be used for Synthetic Transformation. Likelihood of successfully replacing LibGen data: 50%					
Option 4. Status Quo. Datasets which will be excluded in Decisions 1 and 2 can continue to be used directly in model training.					

Meeting Notes

HOW TO: Be sure the QB or another representative from the team is prepared to take notes. Summarize all critical information from the notes after the meeting. Be sure to post completed decisions in Central AI Risk XFN after reaching alignment.

- <add during meeting>

Decision Summary

Decision:

<add decision summary>

Action Items:

1. <add action items and POCs>

Strong example (remember to remove):

Post CPE Summary (updated Dec 16th):

Decision: teams to proceed with option #2 - IAB Link History can be classified as 1PD and used for ads (opt-in in EU/similar; opt-out in ROW)

Action items:

1. Enhance the "user value", i.e transparency, sense of place (via more direct messaging, native app entry points, "app feature as opposed to IAB feature") to increase the overall defensibility and probability of approval by regulators. (POC: Jamie Wells)
2. PXFN will continue to review iterations of UX. Fred Leach to reach out to Tom Hobbs to align on a Privacy POC to partner with Ads on the iterations required. (POC: Fred Leach)
3. Provide an update to MZ on the decision (communicate that we have a legally Redacted and the near term ~0.3% iRev opportunity. (POC: Michel Protti, John Hegeman)
4. Not discussed live in CPE, but PXFN to continue work to solve potential tangential issues relating to deletion/retention, AYI/DYI, deduplication, sensitive data filtering, and privacy policy update/regulatory briefing plan. (POC: Danielle Wisneiwski)

Graveyard

Overall Options

When thinking about potentially pirated data(sets) we will exclude from training, there are a number of factors to consider:

1. Does that data(set) show up on a Piracy Watch List?
2. For data(sets) that show up on a Piracy Watch List, there are two cases:
   a. We have explicitly agreed to not train on data on the watch lists, e.g., in the case we agree to sign the EU AI Act Code of Practice.
   b. We are explicitly required by legislation or regulation to not train on data from the watch list
3. If the data(set) is not on a watch list, is there another reason why it may be considered potentially pirated, e.g., are we facing legal pressure about a particular piece of content?

In addition, for data we decide to not train on directly, there is a question about whether we can use such data to create synthetic data for use in training as a replacement.

Commented [22]: Redacted
The legal position we need to convey is:

Redacted

Commented [23]: we also have the transparency template requirement

	Can be crawled	Can be included in training	Can be used from alternate sources	Can be used to create synthetic data
1. Data(sets) which show up explicitly on a disallowed watch list AND cannot be found elsewhere	Yes	No	No	Yes
2. Data(sets) which show up explicitly on a disallowed watch list, but are included in other freely usable sources	Yes	Yes	Yes	Yes
3. Data(sets) which are "problematic," e.g., show up on a watch list but not one which is explicitly disallowed	Yes	Yes	Yes	Yes
4. Other data	Yes	Yes	Yes	Yes

Commented [24]: Redacted
Redacted

Option	Legal Risk	Policy Risk	Product Risk
(1) Cease to crawl and train on only data sources on the EU Piracy Watch List (in particular **Libgen**)	**Redacted**		

Commented [25]: Redacted
Redacted

	Redacted		
(2) Cease to crawl and train on data in (1) + data sources on other piracy watch lists including the USCO watch list (in particular **Anna's Archive**)			
(3) Cease to crawl and train on data in (1) and (2) + other data sources that are subject to litigation / may be considered 'pirated sources' (in particular **Books3**)			
(4) Cease to crawl Contentious Sources and only use pre-existing crawled content from Contentious Sources			
(5) Cease to crawl Contentious Sources and replace existing sourced content with content from alternative websites before training (i.e. identify equivalent collated datasources/websites)			

(6) Cease to crawl Contentious Sources and replace existing sourced content with content from alternative sources before training (i.e. identify equivalent pieces of content)			
(7) Cease to crawl Contentious Sources and replace existing sourced content with synthetic content before training			

Decisions:

1. When should we exclude potentially pirated content from use in training?
2. For content which we have excluded from training for being potentially pirated, when can we use that data for the creation of synthetic replacement data?
3. Based on Decisions 1 and 2, should we not sign the Code of Practice due to the Piracy Watch List provision

Open Questions

Missing Definitions

1. **Pirated Content:**
    a. Who is defining pirated content?
    b. How is it defined?
    c. Is the way it is defined "fair"?
2. **General Purpose Models:** What does "The model is placed on the market" mean, can we limit this definition to "OSS" models only?

General Purpose AI models is defined in the AI Act itself -

'general-purpose AI model' means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market;

Please see product context on prior decisions we need to make in order to lay out product options more effectively.

1. **Option 1:** Synthetic generate or transform this data sufficiently enough to be considered "net new, non pirated data" to replace Anna's Archive and Libgen data. *Please note:* this comes at significant investment and presents a research risk, we have not yet solved how to generate data at a pre-training scale.
2. **Option 2:** Maintain current operations without releasing open-source software (OSS) in the EU.
3. **Option 3:** Discontinue use of Anna's Archive and Libgen data, accepting roughly [redacted] performance loss for STEM text, and [redacted] performance loss for text overall and an XX% performance loss for multimodal, falling even further behind on the competitive leaderboard.
    a. **Option 3.1**: In the case of option 3, we could find alternative copies from other parts of the web. Use only minimal metadata from Libgen/Anna's Archive to find the same or similar copies of the content elsewhere and replace the actual content from Libgen/Anna's Archive completely. There is a huge risk we will not be able to recover over 90% of this data

Commented [26]: @[redacted]@meta.com

Decisions Needed to Inform Product Options <WILL GO IN APPENDIX>

In order to outline appropriate option to make up the gap, if this data becomes unavailable, **product is dependent on a series of decisions:**

#	Decisions Needed	Trigger Point / Date	Product Context	Policy Context	Legal Context

1	What is our interpretation of the EU AI Counterfeit and Piracy Watch List for our product team, and what concrete actions do we need to take in response?	Flagship Datamix for Pre-training is due mid January but we need to know ASAP if this decision is no. **ASAP** Dec 11, 2024	Retroactive Removal of Usage (non-starter) - this will be detrimental to all of our products as this data makes up a significant proportion of our current and historically released models. Additionally this is **technically infeasible** without re-training every single model from scratch. Go Forward No Usage- We will cease crawling and using this data moving forward. This approach will result in a significant data gap for future models like Llama 5+, and we currently have low confidence in our ability to fill this gap. Go Forward No Crawl - We will stop crawling data from these sites in the future but will continue to use the data already collected. This approach aligns with OpenAI's current claims and helps avoid regressions or	• Policymakers expect that we will not crawl sites listed in the Commission list or the USTR Notorious Markets List at a minimum, but they likely also think this means there will be no training on these materials as well. I am not sure that the crawl/usage distinction would be respected by policymakers. Pirate or notorious sites are very politically stigmatized as illegal, foreign bad actors. • Even assuming the most blunt reading, I don't think they could reasonably expect retroactive removal of usage, I am not aware of other things in the Code that work that way. • Continuing to crawl and use the sites would be inevitably discovered (template, other declarations, etc.) and would be considered to be Meta's declaration that it cannot be trusted to adhere to voluntary	**Redacted**

Commented [27]: There are two pieces here:

1/ For August 2025, only models which are put on the market after that date are in scope.

2/ There is a grace period of 2 years to address models that were already on the market.

Commented [28]: [redacted]@meta.com what do you mean by "grace period?" - as in after two years we will have to take the models off of the market completely?

Commented [29]: This is a good question. [redacted]meta.com, Redacted

Redacted

Commented [30]: This also affects LLama 4 Flagship, no? We would release it after Aug 1, 2025

Commented [31]: Policy - Welcome others' thoughts to confirm this, and legal Redacted Redacted

			significant data gaps that could occur if we removed the data entirely. Continue To Crawl & Use - The code of conduct is optional and we do not have to agree to it until it is specifically mandated by a government for a specific date.	norms/standards. In addition to potential legal consequences, we could expect more restrictive legislation or rulemaking (opt-in v. opt-out for training material use, or preferencing rightsholders in determining standards to replace robots.txt, or heavy downstream liability penalties, etc.	Redacted
1.1	How are we interpreting what a general purpose model is? Specifically, the AI Act defines it as a "model that is placed on the market", can we limit this definition to "OSS" models only?	**ASAP** Dec 11, 2024	limit definition to OSS- we can add a 5th option to the list above to continue to crawl and use this data for all OSS models outside of the EU and we can decide to not release the model in OSS in EU specifically. Definition applies to all Meta GenAI products built on foundational models - the product/ business tradeoffs will		Redacted

			become significantly larger and may not make business sense to exclude the EU from meta products built on top of GenAI foundational models.		Redacted
1.2	Library Genesis is listed as pirated content, does this include both Anna's Archive and Libgen or is it more targeted towards Libgen?	**ASAP** Dec 11, 2024	The "notorious markets" identified annually by the Office of the U.S. Trade Representative " calls out Libgen specifically but Anna's Archive is only referenced as " Nominated as libgen.rs. related sites include libgen.is, libgen.me, libgen.st, library.lol, annas-archive.org, annasarchive.gs, and other mirror sites. Reportedly operated from Russia." making it ambiguous on how we should deal with Anna's Archive.	Policymakers will consider anything containing LibGen to be as bad as LibGen. If we were to use Anna's Archive and document that we never used the LibGen part, maybe we could make a counterargument, but if we just use the same site under a different name, we will be in more trouble than just using the original (for trying to be sneaky, especially when Anna's Archive is also a pirate site, etc.).	Redacted

1.3	What is our interpretation of the definition of pirated content?	Jan 31, 2025	Implementing a clear definition of pirated content would be essential for the data team to anticipate future developments, build upon previous work, and avoid starting from scratch every time an arbitrary list changes.		
2	Can we use this data for synthetic transformation or generation?	Jan 1, 2025 Should be a quick follow up to the decision above. Synthetic Data efforts are incredibly risky and costly, we need to invest in them ASAP if we are to have any chance of meeting any of the EU deadlines.	If Yes (costly & complicated research problem) -Anna's Archive and Libgen contain approximately 98% unique data not found in general webcrawls. Licensing this data would be costly, with only a small fraction available at a high price. In the event of this data unavailability, our team proposes synthetic data transformation and generation as a potential solution to close the data gap. However, this approach requires extensive research, as we have yet to solve the		Redacted

Commented [32]: @[redacted]@meta.com set some arbitrary dates here but feel free to change these - just want to have a deadline to work towards

Commented [33]: I wonder if we should shift this from defining "what is pirated content" to defining what are the requirements for excluding content from training for being potentially pirated. I've added some options at the top of the doc I'd like to discuss live that I think set some razors for this.

			synthetic data generation problem for pre-training scale. If No - There is currently no other obvious way of how we could close this data gap otherwise.		
2.1	If **Yes (above)**, By how much do we need to transform the data in order for it to account as non pirated data?	Jan 1, 2025 Same as above.	Product needs clear guidelines on what we believe is good enough to be considered data that is "not pirated" when generating data or synthetically transforming data.		

Decision 3: How should we handle historically collected data?

	Product Anna Grosul	**Privacy** Zef RosnBrick	**Legal** Ahuva Goldstand, James Whymark, Becca Felsenthal	**Policy** Ania Helseth, Elizabeth L. Kendall	**Comms** Toby Partlett,Ben Walters
Option 1: **Remove the data from future and previous models** Crawl (going forward): ❌ No Use (going forward): ❌ No Use (historical models): ❌ No	NOT POSSIBLE / MAJOR BLOCKER				

	Product Anna Grosul	Privacy Zef RosnBrick	Legal Ahuva Goldstand, James Whymark, Becca Felsenthal	Policy Ania Helseth, Elizabeth L. Kendall	Comms Toby Partlett,Ben Walters
Option 2: **Remove the data from future models** Crawl (going forward): ❌ No Use (going forward): ❌ No					
Option 2: **No longer crawl this data** Crawl (going forward): ❌ No Use (going forward): ✅ No					

	Public Narrative	Product Anna Grosul	Privacy Zef RosnBrick
Option 1. Status Quo. Include datasets named in the EU Commission Counterfeit and Piracy Watch List in training for Llama4 Flagship. • Does not explicitly go against the Code of Practice • We will likely need to be transparent about the inclusion of these datasets, but we don't know specifically given we do not yet have a draft of the transparency template.			
Option 2. Perform trivial transformation on datasets named in the EU Commission Counterfeit and Piracy Watch List prior to use in training for Llama4 Flagship. • Transformation would likely only consist of minor changes, such as inserting additional spaces.			

	Public Narrative	Product Anna Grosul	Privacy Zef RosnBrick
• However, this may allow us to be less specific in any transparency statement.			
Option 3. Perform a non-trivial transformation on datasets named in the EU Commission Counterfeit and Piracy Watch List prior to use in training for Llama4 Flagship. • A non-trivial transformation could ensure the data we train on is "more" different than the datasets in question. • There is significant technical risk in this approach.			

Risk vectors:

- AI conduct code (or copyright policy)
- Transparency

Timing:

- Flagship: one-way door <to Transparency>
- AI code of conduct signing: anytime before August

Strategic tenor:

Anna's notes:

Most immediate decision: how should we proceed with Llama 4 Flagship given the Transparency act

Options:

1. Remove AA and Libgen from training - huge product impact
2. Proceed with training as is and figure out our response
3. Try and synthetically modify data before training - almost impossible to do this in a fruitful way in 1 month

(An inform on 2)

Optional Decision: given we choose option two, what is our strategic response to the transparency act

_.docx

Main document changes and comments

Page 3: Commented [1] **Elizabeth L. Kendall** **1/17/2025 4:45:00 PM**

I thought we were going to say something like "we recommend exploring if there are alternatives for the future" with the understanding that it would take a lot of time/not pinning anyone down to a particular timeframe or even outcome. Or can that be policy's recommendation? I don't want to have this recommendation and just say all the risks are high, it seems like we're not working on any way forward that's useful (again, consistent with the point that no change is happening right now, etc.)

Page 5: Added **Elizabeth L. Kendall** **1/17/2025 4:47:00 PM**

We can also expect legislation to be passed prohibiting training on piracy sites, as they may be defined.

Page 5: Added **Elizabeth L. Kendall** **1/17/2025 4:48:00 PM**

We have strategic partnerships with content providers.

Page 9: Commented [2] **Zef RosnBrick** **12/17/2024 3:26:00 PM**

@[redacted]@meta.com
Assigned to [redacted]@meta.com

Page 11: Formatted **James Whymark** **12/18/2024 11:00:00 AM**

Space Before: 0 pt

Page 11: Added **Ania Helseth** **12/18/2024 4:59:00 PM**

regulatory changes,

Page 11: Commented [3] **James Whymark** **12/18/2024 11:00:00 AM**

@[redacted]@meta.com

Page 11: Commented [4] **Anna Grosul** **12/18/2024 2:16:00 AM**

[redacted]@meta.com [redacted]@meta.com @[redacted]@meta.com
@[redacted]@meta.com can you guys help finish this table please and add risk severity please

Page 12: Commented [5] **Anna Grosul** **12/18/2024 2:14:00 AM**

@[redacted]@meta.com @[redacted]@meta.com @[redacted]@meta.com to fix

Page 12: Commented [6] **Elizabeth L. Kendall** **1/6/2025 3:41:00 PM**

Some of this has been converted from a previous doc and no longer fits the format

Page 12: Commented [7] **Elizabeth L. Kendall** **12/16/2024 7:19:00 PM**

Should we loop in Comms now?

Page 12: Commented [8] **Ania Helseth** **12/18/2024 4:40:00 PM**

We can only use our general narrative, not sure if Open AI here is of relevance. Please also note the announcement of Youtube.

Page 12: Commented [9] **Elizabeth L. Kendall** **1/17/2025 4:33:00 PM**

I'm not certain what you mean, can you add?

Page 13: Commented [10] **James Whymark** **12/17/2024 10:00:00 PM**

How does this relate to book data?

Page 13: Commented [11] **Elizabeth L. Kendall** **1/6/2025 3:44:00 PM**

Use of piracy sites will be perceived as condoning piracy, which all sectors are sensitive about.

Page 13: Commented [12] **James Whymark** **12/18/2024 11:00:00 AM**

[redacted]@meta.com
Assigned to [redacted]@meta.com

Page 13: Commented [13] **Ania Helseth** **12/18/2024 4:43:00 PM**

isn't that tab repetitive to the one above?

Page 13: Commented [14] **Elizabeth L. Kendall** **1/17/2025 4:40:00 PM**

I don't know what different ground this is supposed to cover

Page 13: Commented [15] **Anna Grosul** **12/18/2024 2:12:00 AM**

@[redacted]@fb.com [redacted]@meta.com to add

Page 13: Commented [16] **James Whymark** **12/18/2024 11:01:00 AM**

[redacted]@fb.com
Assigned to [redacted]@fb.com

Page 14: Commented [17] **Anna Grosul** **12/18/2024 2:32:00 AM**

@[redacted]@meta.com [redacted]@meta.com @[redacted]@meta.com @[redacted]@meta.com @[redacted]@meta.com what else is missing?

Page 14: Commented [18] **Ania Helseth** **12/18/2024 4:44:00 PM**

Optimizing opt-out via rights manager..? We need sth to offer rightholders in the back pocket.

Page 14: Commented [19] **Anna Grosul** **12/18/2024 2:32:00 AM**

[redacted]@meta.com can we discuss target dates here?

Page 23: Commented [20] **Zef RosnBrick** **12/10/2024 3:57:00 PM**

[redacted]@meta.com
Assigned to [redacted]@meta.com

Page 25: Commented [21] **Zef RosnBrick** **12/10/2024 5:17:00 PM**

@[redacted]@meta.com can you help us add more details on generation and transformation? I think in particular some context on how different the synthetic content would or could be in each case will be useful to informing the risk.
Assigned to [redacted]@meta.com

Page 27: Commented [22] **James Whymark** **12/9/2024 11:41:00 AM**

I think these are combined issues. The legal position we need to convey is:
- There is an overall legal obligation to have a copyright policy on how we comply with EU copyright law.
- The code of conduct has been released which is the expected (though not the only) way that

providers will develop a compliant policy.
- In that code, there is a requirement to use reasonable measures not to crawl pirated sources including those on specific watch lists.

Page 27: Commented [23] **Ania Helseth** **12/9/2024 3:58:00 PM**

we also have the transparency template requirement

Page 28: Commented [24] **James Whymark** **12/9/2024 11:38:00 AM**

Which watch lists are we expecting this category to cover?

Page 28: Commented [25] **James Whymark** **12/9/2024 1:24:00 PM**

have taken a stab at reformulating the decision options. I think this is broadly what we need to consider based on the below. Does this assist with putting together the decision matrix for the L1 at the end of the week?

Page 31: Commented [26] **Anna Grosul** **12/4/2024 9:20:00 PM**

@[redacted]@meta.com

Page 32: Commented [27] **Zef RosnBrick** **12/4/2024 5:26:00 PM**

There are two pieces here:

1/ For August 2025, only models which are put on the market after that date are in scope.

2/ There is a grace period of 2 years to address models that were already on the market.

Page 32: Commented [28] **Anna Grosul** **12/4/2024 6:25:00 PM**

@[redacted]@meta.com what do you mean by "grace period?" - as in after two years we will have to take the models off of the market completely?

Page 32: Commented [29] **Zef RosnBrick** **12/5/2024 5:35:00 PM**

This is a good question. @[redacted]@meta.com, for models that were created before the code of practice went into effect but violate something agreed to in the code, would we need to take such models off the market by 2027?

Page 32: Commented [30] **Sean Bell** **12/6/2024 1:19:00 AM**

This also affects LLama 4 Flagship, no? We would release it after Aug 1, 2025

Page 32: Commented [31] **Elizabeth L. Kendall** **12/5/2024 1:31:00 AM**

Policy - Welcome others' thoughts to confirm this, and legal if you're aware of other parts of the code that would work retroactively.

Page 35: Commented [32] **Anna Grosul** **12/4/2024 6:23:00 PM**

@[redacted]@meta.com set some arbitrary dates here but feel free to change these - just want to have a deadline to work towards

Page 35: Commented [33] **Zef RosnBrick** **12/5/2024 5:39:00 PM**

I wonder if we should shift this from defining "what is pirated content" to defining what are the requirements for excluding content from training for being potentially pirated. I've added some options at the top of the doc I'd like to discuss live that I think set some razors for this.

Header and footer changes

Text Box changes

Header and footer text box changes

Footnote changes

Endnote changes