EXHIBIT TWENTY-TWO

_.docx

Main document changes and comments

Page 3: Commented [1] **Veronica Kolstad** **11/29/2023 5:09:00 AM**

is this covered in 2.1 lineage?

Page 3: Commented [2] **Henry Aspegren** **11/29/2023 5:54:00 PM**

Lineage is our first step towards that goal - since we can view and audit what data went into training.

Ads personalization is the first-use case for technical enforcement

Page 3: Commented [3] **Zhenyu Yang** **11/30/2023 5:47:00 PM**

qq: how do we define the boundary of e2e lineage? for instance, will it be the immediate training datasets(first-step away from the real training job), or x-steps all the way up to some boundary predefined by e.g., lama reviews?

Page 3: Commented [4] **Mike Clark** **11/30/2023 10:22:00 PM**

That is 1.1 in the understand work

Page 3: Commented [5] **Veronica Kolstad** **11/29/2023 4:47:00 AM**

please add more specificity around what "automated coverage" means -- does that mean the GenAI systems send near real-time lineage data to ULS/AIM? there are more details below, but if folks are only seeing KRs in rollup reporting, it's helpful to be clear on expectations

Page 3: Commented [6] **Veronica Kolstad** **11/29/2023 4:56:00 AM**

do we want to set a bar for priority/criticality of the assets?

Page 3: Commented [7] **James Beldock** **11/29/2023 5:21:00 PM**

Having a framework for priority/criticality (in addition to a bar) would be good too.

Page 3: Commented [8] **Ming Qiao** **12/14/2023 9:05:00 PM**

Is there any particular reason that we didn't use more specific PAI language here? e.g. does "protection" equal to "enforcement"?

Page 3: Commented [9] **Veronica Kolstad** **11/29/2023 4:59:00 AM**

do we need to limit this to a specific type of launch for it to be achievable? i see the project for 3.3 is ads<>messenger, perhaps we can limit it to that?

Page 3: Commented [10] **David Levin** **11/29/2023 9:40:00 PM**

For extreme clarity purposes, in places where "Data Infra" is indicated, we might want to mention the exact teams to partner with. If not sure we can help figure it out.
1 total reaction
Veronica Kolstad reacted with 💯 at 2023-11-30 19:56 PM

Page 4: Commented [11] **Rituraj Kirti** **11/30/2023 3:56:00 PM**

replace with "assets" ?

Page 4: Commented [12] **Chao Zhou** **12/1/2023 5:51:00 PM**

Entitlement here specifically means the api key called used to call GenAI APIs: https://www.internalfb.com/metagen/api/entitlements

Page 4: Commented [13] **David Levin** **11/29/2023 9:16:00 PM**

Is this a P0 (by color) or P1?

Page 4: Commented [14] **Veronica Kolstad** **11/29/2023 5:03:00 AM**

i was working on similar goals in AI Privacy for ML workflow tools and PZM; I would love to chat with the UXR covering this :)

Page 4: Commented [15] **James Beldock** **11/29/2023 5:22:00 PM**

...and wouldn't you know it, the UXR is none other than @[redacted]@meta.com herself! :-)
1 total reaction
Veronica Kolstad reacted with ♥ at 2023-11-30 19:58 PM

Page 4: Commented [16] **Rituraj Kirti** **11/30/2023 3:57:00 PM**

entities?

Page 4: Commented [17] **Chao Zhou** **12/1/2023 6:02:00 PM**

It's the name of GenAI api key :)
https://www.internalfb.com/metagen/api/entitlements

Page 4: Commented [18] **Veronica Kolstad** **11/29/2023 5:04:00 AM**

@[redacted]@meta.com any risks for integrating this into the tooling vs. using PZM's existing experience? are we enabling this elsewhere?

Page 4: Commented [19] **Rituraj Kirti** **11/30/2023 7:20:00 AM**

Yes I would recommend extending existing tools vs building new

Page 4: Commented [20] **Rituraj Kirti** **11/30/2023 3:57:00 PM**

Also second building this into PZM

Page 4: Commented [21] **Chao Zhou** **12/1/2023 6:04:00 PM**

Yea, those metadata supported in PZ we should use PZ
Otherwise AIDC had the tool which design is ready to add metadata.

Page 4: Commented [22] **David Levin** **12/3/2023 9:05:00 AM**

@[redacted]@meta.com, we can look into which entry points we want to enable for settings these labels, driven by ease of use.

Page 4: Commented [23] **Veronica Kolstad** **11/29/2023 5:05:00 AM**

NB: the org FKA AI Privacy created AI Privacy Settings (AIPS) to help reuse some privacy context, which could be helpful here.

Page 4: Commented [24] **Chao Zhou** **12/1/2023 6:04:00 PM**

Agreed. We should reuse if exists

Page 5: Commented [25] **Rochit Sapra** **11/29/2023 12:19:00 AM**

@[redacted]@meta.com FYI - You should cover how Genie will support the purpose limitation needs for Ads.

Assigned to jackzh@meta.com

Page 6: Commented [26]	Sonia Kim	12/14/2023 2:45:00 AM

Would love to collaborate here to work on joint KRs. (I'm a SWE in GenAI Trust). We have an e2e platform (GenStorm) for red teaming that enables testing for text->text MetaGen models.

Page 6: Commented [27]	Arun Rao	11/28/2023 7:36:00 PM

Red teaming is pretty limited - it's hard to cover all the edge cases but it can help us find relevant categories. Ultimately, RLHF is the most scalable and some annotation can take insights from both to make a more robust set of classifiers. We need to talk about RLAIF too - how to use that well for safety.

Page 7: Commented [28]	Sonia Kim	12/14/2023 2:46:00 AM

This half, GenAI Trust has already started on this process https://docs.google.com/document/d[redacted]

Page 7: Commented [29]	Sonia Kim	12/14/2023 2:49:00 AM

Have there been any discussion in recommending / enforcing an internal model card? This would provide information on how models were fine-tuned / input / output classifiers, etc. This would help us 1. keep track of all the model information in one place, and 2. if a model is unsafe, provide us with a starting point for debugging why it may be so (ie. maybe the fine-tuning dataset was unsafe)

Page 9: Commented [30]	Chao Zhou	12/4/2023 7:01:00 AM

I felt hard to stab out PL projects from other Privacy projects when I composed the list. Meanwhile, I have many projects about data I can't find right section to fit in.
I think I'd consider to merge Privacy & PL and add a section for data

I understand the 4 sections were figured out after many-rounds discussions among the planning leads. Just want to put a frontline eng's thought on your radar :) @[redacted]@meta.com [redacted]@meta.com @[redacted]@meta.com

Page 10: Commented [31]	Arun Rao	11/28/2023 7:38:00 PM

What about 1P training data? Can we unblock public posts and such, and the reviews for that UGC?

Page 10: Commented [32]	Mike Clark	11/28/2023 8:02:00 PM

Those get solved in KR4.1, as those all require PXFN approval. The 3P data is a separate governance problem today.

Page 10: Commented [33]	David Levin	11/29/2023 9:31:00 PM

@[redacted]@meta.com - we plan to invest in documenting datasets mitigations and evidence in AIDC. Might contribute to this one.
Assigned to [redacted]@meta.com

Page 11: Commented [34]	Arya Talebzadeh	1/9/2024 10:08:00 PM

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 11: Commented [35]	Arya Talebzadeh	1/9/2024 10:08:00 PM

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 11: Commented [36] **Prasad Buddhavarapu** **1/23/2024 5:53:00 AM**

Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Page 11: Commented [37] **Arya Talebzadeh** **1/9/2024 10:08:00 PM**

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 11: Commented [38] **Prasad Buddhavarapu** **1/23/2024 5:53:00 AM**

Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Page 11: Commented [39] **Arya Talebzadeh** **1/9/2024 10:08:00 PM**

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 14: Commented [40] **Praveen Gollakota** **12/8/2023 9:03:00 AM**

There are no projects that refer to this OKR. Is this out of scope for H1 @[redacted]@meta.com _Assigned to [redacted]@meta.com_

Page 14: Commented [41] **Arya Talebzadeh** **11/27/2023 11:04:00 PM**

There's a missing pre-req here to **Redacted - Privilege** **Redacted - Privilege**

Page 14: Commented [42] **Mike Clark** **11/28/2023 8:04:00 PM**

Can you add them?

Page 14: Commented [43] **Praveen Gollakota** **12/8/2023 9:02:00 AM**

How is this measurement different from existing violation rate and refusal rate metrics that are owned by safety team? @[redacted]@meta.com cc: @[redacted]@meta.com _Assigned to [redacted]@meta.com_

Page 14: Commented [44] **Arya Talebzadeh** **1/9/2024 10:08:00 PM**

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 14: Commented [45] **Arya Talebzadeh** **1/9/2024 10:08:00 PM**

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 14: Commented [46] **Arya Talebzadeh** **1/9/2024 10:08:00 PM**

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 14: Commented [47] **Prasad Buddhavarapu** **1/23/2024 5:53:00 AM**

Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Page 14: Commented [48]	Arya Talebzadeh	1/9/2024 10:08:00 PM

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 14: Commented [49]	Arya Talebzadeh	1/9/2024 10:08:00 PM

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 15: Commented [50]	Arya Talebzadeh	1/9/2024 10:08:00 PM

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 15: Commented [51]	Prasad Buddhavarapu	1/23/2024 5:53:00 AM

Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Page 15: Commented [52]	Arya Talebzadeh	1/9/2024 10:08:00 PM

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Page 15: Commented [53]	Arya Talebzadeh	1/9/2024 10:08:00 PM

I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Header and footer changes

Text Box changes

Header and footer text box changes

Footnote changes

Endnote changes

[STATUS: Work in Progress | STO: Mike Clark]

Gen AI Platform 2024H1: Privacy, Safety, and Data Strategy

Part of 2024 H1 planning for the GenAI Platform

Main Roadmap: GenAI Platform 2024.H1 (permalink fb/genai_platform_2024h1)
Sub-Roadmaps: MetaGen | Genie | Model Trains | **Privacy, Safety & Data Strategy**
Privacy, Safety & Data Roadmaps: 1. GenAI Purpose Limitation | 2. Safety | 3. Privacy | 4. IP

Rolled Up OKRs

OKRs

Objective 1: Ensure Compliance with Regulations and Commitments
KR1.1: Establish GDA labels and default purpose limitation policies
KR1.2: Define principles for safety mitigation effectiveness
KR1.3: Define obligations, commitments, and prioritization framework for privacy
Objective 2: Implement Robust Safety and Privacy Measures
KR2.1: Establish process for developing, deploying, and updating auto-eval for CRS and CRSM
KR2.2: Make IXFN reviews faster for teams using MetaGen
KR2.3: Ship youth required safety mitigations by default for MetaGen models
KR2.4: Ship regular recurring auto-eval to 50% models powering generative AI products on the platform
KR2.5: Meet data flywheel purpose limitation requirements for new product launches
Objective 3: Enable Efficient and Secure Data Use
KR3.1: Establish end-to-end lineage coverage for fine-tuning and pre-training
KR3.2: Enable training of "clean" GenAI models with verifiable lineage
KR3.3: Develop a pipeline of potential third-party non-user data for review
KR3.4: Develop tooling to automate or accelerate the analysis of privacy and scraping impact of datasets
KR3.5: Prevent domains that represent outsized privacy risk from being used across GenAI applications

Planning STO: Mike Clark

1. GenAI Purpose Limitation

Planning STO: Vishy Poosala, Henry Aspegren, Mike Clark, Chao Zhou
- MetaGen POC: Chao Zhou
- Genie POC: Jack Zhang

V-team: AI Infra, DI, Central Privacy and GenAI

Background

A key tenet of privacy commitments and obligations is on using data for intended purposes. In order to launch generative AI products across Meta in 2023, Access Control Lists (ACLs) were used to meet those commitments and obligations across different product groups. While ACLs helped us meet the launch criteria, they can be fragile and are difficult to evidence their effectiveness over time. Because of the complexity, they are also expensive for teams to implement across many products.

Broadly in the company there has been an investment in Privacy Aware Infra (PAI) to create more durable solutions to solve for purpose limitation. GenAI Purpose Limitation leverages some of the same components of PAI to solve the same problems for generative AI products across Meta.

One note, purpose limitation is a prerequisite to enabling opt-out. Currently there is no explicit pressure on providing an opt-out for interaction data from any regulators but we anticipate at some point in the future that may be a requirement. This will be captured as a risk and not a KR for H1 2024.

Mission

- Enable product teams deploying generative AI products at Meta to efficiently and effectively meet data use privacy commitments (statements we have made externally representing our practices) made to regulators and end-users.
- Decreasing the cost and effort for Meta to do purpose limitation on data for generative AI products to meet our obligations (regulatory requirements or requirements from litigation or judicial order).

OKRs

Objective 1: Understand the purpose limitation requirements for generative AI
KR1.1: Establish GDA (*Governable Data Annotation*) labels for prioritized data boundaries needed for GenAI Purpose Limitation
KR1.2: Define a set of default purpose limitation policies for generative AI products across Meta
KR1.3: Ship a prioritization framework for which combination of labels and policies need to be solved first based on obligations and commitments
Objective 2: Generative AI models are trained on DXA-approved corpora and have verifiable lineage
KR2.1: GenAI models have known end-to-end lineage coverage for fine-tuning and pre-training, with [25%] done via automated coverage
KR2.2: Unblock ads personalization use-cases by enabling them to train "clean" GenAI models with verifiable lineage
Objective 3: MetaGen meets purpose limitation requirements for flywheel data
KR3.1: Ship the generative AI recipe as the durable solution to ensure that AI Agent flywheel data can only be used for approved purposes (L1189414PRV fast

Commented [1]: is this covered in 2.1 lineage?

Commented [2]: Lineage is our first step towards that goal - since we can view and audit what data went into training.

Ads personalization is the first-use case for technical enforcement

Commented [3]: qq: how do we define the boundary of e2e lineage? for instance, will it be the immediate training datasets(first-step away from the real training job), or x-steps all the way up to some boundary predefined by e.g., lama reviews?

Commented [4]: That is 1.1 in the understand work

Commented [5]: please add more specificity around what "automated coverage" means -- does that mean the GenAI systems send near real-time lineage data to ULS/AIM? there are more details below, but if folks are only seeing KRs in rollup reporting, it's helpful to be clear on expectations

follow work)
KR3.2: [25%] of GenAI flywheel assets have automated PAI protection
KR3.3: Meet data flywheel purpose limitation requirements for new product launches

Projects

Detailed version: GenAI PL Roadmap Planning
Priorities **P0**: Must-have • **P1**: Want-to-have • **P2**: Nice-to-have.
T-Shirt Size: **S**: < 2 weeks • **M**: 2~4 weeks • **L**: 4~8 weeks • **XL** : > 8 weeks

Priority	OKR	Project	Description	Owning Team	T-Shirt Size	Partner Teams
P0	1.1, 1.2 and 1.3	Purpose limitation understanding	Collaborate with Central Privacy to define the purpose labels, policies and prioritization framework.	GenAI		Privacy
P0	2.2 and 3.1	Privacy aware components	Many current GenAI components were built recently and are privacy unaware, including MetaGen models, foundation models, training data, lineage, API, prompt eng and entitlement etc. We should integrate these components with the privacy infra (PAI). Examples: onboarding MetaGen entitlement to OneCatalog, integrating GenAI models with AIM, datasets to AIDC etc.	GenAI	XL	AIM, Data Infra and Privacy
P1	2.1	Privacy visualization tools	Provide a consistent top-level surface to intuitively present the privacy elements on GenAI toolings. Examples: presenting purposes on MetaGen entitlement, privacy labels on MetaGen/MetaBot model cards and Genie foundation model tool, privacy powersearch for models, one-stop-shop for setting and presenting GDA annotations etc	GenAI	XL	Privacy, AIM, AIDC, MetaGen tooling, Genie tooling
P0	3.1	Privacy metadata enrichment	Following the privacy aware GenAI components, privacy metadata s should be backfilled, including labels, descriptions, lamas etc. Examples: metadata enrichment on training dataset with labels, descriptions, oncalls etc, define training policies for foundation models	GenAI	XL	Data Infra, Privacy, AIM
P0	2.1	End to end GenAI lineage	Build model lineage in GenAI and data lineage with Data infra. Connect model and data lineage to build a comprehensive lineage graph. Examples: GDA propagation and reclassification for GenAI models. All type training data lineage support (hive-partition, NFS, manifold and AWS)	GenAI	XL	AIM, Data Infra
P0	3.1 and 3.2	Flywheel data protection	Implement a policy zone to protect AI Agent flywheel data and limit its use to approved purposes under L1189414PRV (LaMa fast follow). Build a data router to ensure that flywheel data goes into the appropriate table with the correct GDA labels.	GenAI	L	Flywheel
P0	2.2	Ads personalization using approved GenAI models	Clearly mark the models that can be used for Ads personalization and empower Ads customers to use.	GenAI	M	Monetization
P1	2.2 and 3.1	Runtime enforcement for chokepoints	Enforce in runtime for the choke points before model serving. It'd further safeguard the data usage.	GenAI	XL	AIDC

Commented [6]: do we want to set a bar for priority/criticality of the assets?

Commented [7]: Having a framework for priority/criticality (in addition to a bar) would be good too.

Commented [8]: Is there any particular reason that we didn't use more specific PAI language here? e.g. does "protection" equal to "enforcement"?

Commented [9]: do we need to limit this to a specific type of launch for it to be achievable? i see the project for 3.3 is ads<>messenger, perhaps we can limit it to that?

Commented [10]: For extreme clarity purposes, in places where "Data Infra" is indicated, we might want to mention the exact teams to partner with. If not sure we can help figure it out.
1 total reaction
Veronica Kolstad reacted with 👀 at 2023-11-30 19:56 PM

Commented [11]: replace with "assets" ?

Commented [12]: Entitlement here specifically means the api key called used to call GenAI APIs: https://www.internalfb.com/metagen/api/entitlements

Commented [13]: Is this a P0 (by color) or P1?

Commented [14]: i was working on similar goals in AI Privacy for ML workflow tools and PZM; I would love to chat with the UXR covering this :)

Commented [15]: ...and wouldn't you know it, the UXR is none other than @[redacted]@meta.com herself! :-)
1 total reaction
Veronica Kolstad reacted with ❤️ at 2023-11-30 19:58 PM

Commented [16]: entities?

Commented [17]: It's the name of GenAI api key :) https://www.internalfb.com/metagen/api/entitlements

Commented [18]: @rituraj@meta.com any risks for integrating this into the tooling vs. using PZM's existing experience? are we enabling this elsewhere?

Commented [19]: Yes I would recommend extending existing tools vs building new

Commented [20]: Also second building this into PZM

Commented [21]: Yea, those metadata supported in PZ we should use PZ
Otherwise AIDC had the tool which design is ready to add metadata.

Commented [22]: [redacted]@meta.com, we can look into which entry points we want to enable for settings these labels, driven by ease of use.

Commented [23]: NB: the org FKA AI Privacy created AI Privacy Settings (AIPS) to help reuse some privacy context, which could be helpful here.

Commented [24]: Agreed. We should reuse if exists

			Examples: Enforce that only approved dataset are used for training in Genie. 100% training data registration.			
P1	2.1	Non-LLM Models Lineage	Following the projects of LLMs, build the same thing for non-LLM models (LDM, music and multi-modal)	GenAI	XL	AI LDM and AIM

Risks

Risk	Level	Description	Mitigation
1. Ads personalization use-case scope	High	Technical design for how GenAI will support this use-case is not yet clear. Impact: Delays to the launch if it takes additional work we haven't planned for.	Working with POCs to determine design and timeline: - Metagen: Ryan Fox - Genie: Jack Zhang - Monetization: Nimish Shah
2. Opt-Out Timeline	High	Opt-out could become a requirement in the future, but we have not prioritized this work in H1. Impact: If we have to support it we will have to drop other work.	We are doing the foundational work now, which is a prerequisite to supporting opt-outs in the future. We will closely monitor this requirement and re-prioritize our work if needed.
3. EU Launch	High	EU launch could have additional purpose limitation requirements that we have not accounted for here. Impact: Potential delays to EU launch if we cannot support them.	We are taking KR 2.1 in the Privacy roadmap to understand EU requirements, including for purpose limitation. We will use the framework from KR 1.3 to re-prioritize our work as needed.
4. Developer Friction	Medium	Enforcing purpose limitation requirements could increase friction for developers when training and deploying new models. Impact: Slower time to market.	Our design aims to minimize friction by relying heavily on automation to do policy checks quickly. We will roll out enforcement slowly so that we can identify and fix any issues that arise.
5. Messaging Data Use Purpose Policies	Medium	Purpose policies for messaging data are not clear. We aren't planning on doing any additional purpose limitation work for CM46010. Impact: If this changes, we will have to drop/re-prioritize our work to support it.	Shaun Lindsay from Privacy Area is POC to deliver a clear policy for messaging data. We are adding KR to define data use and purpose policies for GenAI with central privacy. Our approach will be to build a general purpose limitation solution that is flexible enough to support changing policies.

Dependencies We Have on Others

Dependencies	Level	Description	Mitigation
1. AI Infra, DI and Central Privacy	High	We have a number of dependencies on infra teams to provide GDA integrations with Genie, Bento, Diaquery and others, support for registering datasets in AIDC, model registration metadata and lineage coverage as well as mapping model types to GenAI model trains with AIM.	We've stood up a v-team working group to plan the work across these orgs. Plan is under development here.
2. Monetization	Medium	We are dependent on monetization to collaborate around the ads personalization use-case and, potentially, implement portions of the design.	We reached an agreement with monetization around a co-development model and are actively planning out what work is required for this use-case.

Commented [25]: [redacted]@meta.com FYI - You should cover how Genie will support the purpose limitation needs for Ads.
Assigned to [redacted]meta.com

Related/Reference Documents

- GenAI Purpose Limitation Design and Plan

---

2. Safety

Planning STO: Mike Clark
- MetaGen POC: Chao Zhou
- Genie POC:
- AI Integrity: Yi Zhong
- IXFN: Michael Feldergraf

V-team: GenAI Trust, XI, CI, IXFN

Background

With the increasing volume of 1st Party GenAI launches, there is a need to develop standardized best practices and solutions for products to integrate that enable safe launches through the IXFN Review Process. Meta's Content Risk Standards (CRS) and Content Risk Standards for Media (CRSM) are the policies that govern what our AI models are encouraged to produce. These policies, which apply to all our generative AI products, were written primarily to address legal risks and youth protection issues. We use these policies every day to perform red teaming and reinforcement learning from human feedback. Teams have already used these policies to label more than 7 million pieces of content created by our models. We also use these policies to train classifiers that detect when content created by our models is likely to violate our content standards.

Though these policies are in many ways similar to pre-existing policy sets, like our Community Standards, there are also many differences. The most obvious reason for these differences is that Meta's pre-existing policies were written to govern interactions between human beings. Policies for generative AI, on the other hand, are used to govern non-human behavior. These differences may - in select instances - result in confusing user experiences. Someone could, for example, use our image generation products to create an image that would be removed for violating our policies if posted on Facebook.

All of Meta's generative AI models - no matter what product they're incorporated into - must abide by these policies. Teams are, however, also free to add on additional safeguards above and beyond these baseline policies. When a given product creates content intended to be shared with others beyond the content creator or when the team has a unique product vision, product teams may also consider additional safeguards. Because we generally want to avoid punishing people for sharing content that Meta's AIs generate for them, in certain instances product teams should consider also scanning model outputs using other, pre-existing sets of classifiers (ex: those intended to implement the Community Standards or Community Guidelines).

Mission

- Enable product teams deploying generative AI products on the GenAI Platform at Meta to efficiently and effectively meet the safety requirements for generated content established by the Content Risk Standards and Content Risk Standards for Media.
- Decreasing the cost and effort for products on the GenAI Platform at Meta to meet the Content Risk Standards.
- Make it easier for the IXFN to evaluate product teams deploying models on the MetaGen platform.

OKRs

Objective 1: Understand what it takes for product teams to meet generative AI safety requirements
KR1.1: Establish a safety runbook for product teams deploying generative AI products on the platform
KR1.2: Define and publish a set of principles for the effectiveness of safety mitigations on the platform
Objective 2: Solve for most of the safety work needed for generative AI products with MetaGen

Commented [26]: Would love to collaborate here to work on joint KRs. (I'm a SWE in GenAI Trust). We have an e2e platform (GenStorm) for red teaming that enables testing for text->text MetaGen models.

Commented [27]: Red teaming is pretty limited - it's hard to cover all the edge cases but it can help us find relevant categories. Ultimately, RLHF is the most scalable and some annotation can take insights from both to make a more robust set of classifiers. We need to talk about RLAIF too - how to use that well for safety.

Commented [28]: This half, GenAI Trust has already started on this process https://docs.google.com/document/d[redacted]edit?usp=sharing

Commented [29]: Have there been any discussion in recommending / enforcing an internal model card? This would provide information on how models were fine-tuned / input / output classifiers, etc. This would help us 1. keep track of all the model information in one place, and 2. if a model is unsafe, provide us with a starting point for debugging why it may be so (ie. maybe the fine-tuning dataset was unsafe)

KR2.1: Establish the process to develop, deploy and regularly update an auto-eval for CRS and CRSM for 50% of models deployed on MetaGen
KR2.2: Make IXFN reviews for generative AI products on the platform faster for teams using MetaGen
KR2.3: Ship youth required safety mitigations by default for MetaGen models
KR2.4: Ship regular recurring auto-eval to 50% models powering generative AI products on the platform
Objective 3: Provide the tools to help teams measure and monitor the effectiveness of their safety mitigations
KR3.1: Ship required default input and output filters for MetaGen hosted models
KR3.2: Ship baseline input and output filters based on a variety of standards, including Community Standards, for teams not using MetaGen
KR3.3: Ship IXFN required Measurement Tools by default for MetaGen models

Projects

Priorities **P0**: Must-have • **P1**: Want-to-have • **P2**: Nice-to-have.
T-Shirt Size: **S**: < 2 weeks • **M**: 2~4 weeks • **L**: 4~8 weeks • **XL** : > 8 weeks

Priority	OKR	Project	Description	T-Shirt Size	Owning Team	Partner Teams	Alignment Status
P0	KR 1.1						Not Started
P0	KR 1.1						Not Started
P0	KR 1.2						Not Started
P0	KR 1.2						Not Started
P0	KR 2.1						Not Started
P0							Not Started
P0							Not Started
P0							Not Started
P0							Not Started
P0							Not Started
P0							Not Started
P0							Not Started
P1	KR 2.1						Not Started
P1	KR 2.1						Not Started
P1	KR 1.2						Not Started
P1	KR 1.2						Not Started
P1	KR 1.3						Not Started

P1	KR 2.2.						Not Started
P2	KR 1.3						Not Started
P2	KR 1.3						Not Started
P2	KR 2.1						Not Started
P3	KR 2.2						Not Started
P3	KR 2.2						Not Started
P3	KR 2.1						Not Started
P3	KR 2.2						Not Started
							Not Started
							Not Started
							Not Started
							Not Started

Related/Reference Documents

- GenAI: IXFN Review/Strategic Risk Review - Sign Up
- [WIP: v0 only] - GenAI x IXFN/MIS
- [WIP] GenAI x IXFN/MIS
- GenAI Meta Integrity Standards
- https://www.internalfb.com/intern/wiki/XIAI/GenAI_Enable/onboarding_new_products/New_Rule_Onboarding/
- MIS = https://www.internalfb.com/intern/wiki/Integrity/(WIP)_MIS_2.0_Standards/(WIP)_Ensure_functioning_SRT_Review_for_Content_Conduct/

---

3. Privacy

Planning STO: Mike Clark
- Eng - (unfunded)
- MetaGen POC: Chao Zhou
- Genie POC: Jack Zhang
- Central Applied Science POC: Prasad Buddhavarapu, Will Bullock(EM)
- Privacy PM/TL: Jonathan McPhie/Shaun Lindsay

V-team: AI Infra, DI, Central Privacy and GenAI

Background

The Privacy landscape for generative AI is dynamic and changing rapidly. Regulations are in development and there are no standards to base future assumptions.

Commented [30]: I felt hard to stab out PL projects from other Privacy projects when I composed the list. Meanwhile, I have many projects about data I can't find right section to fit in.
I think I'd consider to merge Privacy & PL and add a section for data

I understand the 4 sections were figured out after many-rounds discussions among the planning leads. Just want to put a frontline eng's thought on your radar :) [redacted]@meta.com @[redacted]@meta.com @[redacted]@meta.com

Generative AI creates a unique set of potential risks which had been mitigated manually for the sake of launching Llama 2 and products at Connect. Scaling generative AI across the company requires automating the mitigations which solve for key privacy risks like; memorization, use of human characteristics data, and the key privacy principles including data collection, usage, deletion, retention and transfer. We have established a "privacy bar" with our disclosures to regulators and our transparency to users on our data practices.

Mission

- Enable product teams deploying generative AI products at Meta to efficiently and effectively meet privacy commitments made to regulators and end-users.
- Decreasing the cost and effort for Meta to meet privacy requirements in PXFN for generative AI products.

OKRs

Objective 1: Understand generative AI Privacy requirements
KR1.1: Define generative AI obligations and commitments and a prioritization framework for execution planning.
KR1.2: For each privacy risk area, establish an aligned set of guidelines/requirements
KR1.3: Establish a Privacy runbook (bringing together requirements across all privacy risk areas) for product teams deploying generative AI products on the platform
Objective 2: Unblock the launch of generative AI products in the EU
KR2.1: Land the requirements for launching existing generative AI features in the EU
KR2.2: Update/expand existing strategic frameworks to include current and future EU considerations
KR2.3: Update Privacy runbook with EU-specific requirements to enable teams to efficiently launch features globally
Objective 3: Unblock dataset reviews for 1P training data
KR3.1: Develop a pipeline of potential third party non-user data for review
KR3.2: Develop tooling to automate or accelerate the analysis of privacy and scraping impact of datasets
KR3.3: Ad hoc dataset reviews take less than one week and approved datasets are registered and labeled in AIDC
KR3.4: Prevent domains that represent outsized Privacy risk from being used across GenAI applications.
Objective 4: Solve for most of the privacy work for generative AI products using the platform
KR4.1: Make PXFN reviews for generative AI products on the platform more efficient cutting review time by 50%
KR4.2: Define framework for PII memorization measurement
KR4.3: Ship measuring response content risk to track re-identifiability of PII risk associated with the actual responses generated by existing Meta generative AI products.
KR4.4: Measure model extractability to track whether reidentifiability of PII can potentially be extracted from generative AI models through normal user interactions or adversarial prompts.

Commented [31]: What about 1P training data? Can we unblock public posts and such, and the reviews for that UGC?

Commented [32]: Those get solved in KR4.1, as those all require PXFN approval. The 3P data is a separate governance problem today.

Commented [33]: @[redacted]@meta.com - we plan to invest in documenting datasets mitigations and evidence in AIDC. Might contribute to this one. _Assigned to [redacted]@meta.com_

Projects

Priorities **P0**: Must-have • **P1**: Want-to-have • **P2**: Nice-to-have.
T-Shirt Size: **S**: < 2 weeks • **M**: 2~4 weeks • **L**: 4~8 weeks • **XL** : > 8 weeks

Priority	OKR	Project	Description	T-Shirt Size	Owning Team	Partner Teams	Alignment Status
P0	KR 1.1	Track the deployment of generative ai models.	Enumerate the inference infrastructure and instrument to allow the centralized tracking of deployed generative AI models.				Not Started
P0	KR 1.1	Connect Dataset Reviews to lineage chains					Not Started
P0	KR 1.2						Not Started
P0	KR 1.2						Not Started
P0	KR 2.1						Not Started
P0	KR 1.2	Develop the framework and plan for the potential need for creator opt-out solutions in future versions of models					Not Started
P0	KR 5.2	Land measuring response content risk for existing models on the platform.	Response Content Risk measures the reidentifiability risk associated with the actual responses generated by existing Meta GenAI products and shared with Meta users. This category of risk has the highest severity and needs immediate mitigations if detected.	L	AI Privacy Area		Not Started
P0	KR 5.3	Measure model extractability to track whether PII related information can potentially be extracted from generative AI models through normal user interactions or adversarial prompts.	Model Extractability Risk measures the reidentiable information that can potentially be extracted from GenAI models through normal user interactions or adversarial prompts (e.g., jailbreak prompts). This category of risk is a superset of the response content risk, but might be less severe if the vulnerability has not been exploited.	XL	AI Privacy Area,		Not Started
P0	KR 5.4	Measure Model Memorization Score.	Model Memorization Score measures GenAI model's ability to memorize reidentifiable content in its training data. Having a score that's independent of the extraction method would provide a more holistic understanding of reidentifiability risk of the GenAI model, especially for open source models. The risk reflected by this score is expected to be a superset of the model extractability risk. Specifically, the score should be able to cover the majority of known extractability risk and also be able to flag new extractability risks. The northstar goal is to establish and adopt an industry standard on measuring model memorization risk.	XL	AI Privacy Area, Central Applied Science		Not Started
P0	KR 5.1	Define framework for using Privacy risk measurement	Identify the various thresholds of risk associated with Privacy risk measurement and how/where they should be applied.	L	AI Privacy Area Central Applied Science		Not Started
P2	KR 1.3						Not Started
P2	KR 1.3						Not Started
P2	KR 2.1						Not Started
P3	KR 2.2						Not Started

Commented [34]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [35]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [36]: Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Commented [37]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [38]: Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Commented [39]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

P3	KR 2.2						Not Started
P3	KR 2.1						Not Started
P3	KR 2.2						Not Started
							Not Started
							Not Started
							Not Started
							Not Started
							Not Started
							Not Started
							Not Started
							Not Started
							Not Started
							Not Started

Related/Reference Documents

- Privacy/Copyright Risk Management Strategic Position 1-Pager Technical Strategy
- a/c priv - AI Privacy Area Canonical Plan

---

4. Intellectual Property (IP)

Planning STO: Mike Clark
- Eng - (Unfunded)
- MetaGen POC:
- Genie POC:
- Central Applied Science POC (not active yet): Prasad Buddhavarapu, Will Bullock(EM)
- IP Team: James Whymark, Olena Ripnick-O'Farrell and Lucas Dahlin

V-team: AI Infra, DI, Central Privacy and GenAI

Background

Redacted - Privilege

Today, we rely on fair use as a defense to copyright claims where we are making copies/training models in the US. **Redacted - Privilege**

Redacted - Privilege

The industry direction of travel is changing for the industry where OpenAI has announced plans for creators to opt-out of their visual content being used for training DALL-E.

The proposed EU AI Act input transparency obligation (if passed) could provide visibility into the content used in our datasets. Redacted - Privilege Redacted - Privilege As currently proposed, this transparency requirement would apply to any models released for use in the EU, even if trained in the US. Redacted - Privilege Redacted - Privilege

Mission

- Establish the plan to help product teams be ready to respond to the dynamic IP regulatory and litigation environment.
- Every model, system or application relying on generative AI goes through Redacted - Privilege
- Each IP mitigation is documented with the associate models Redacted - Privilege

OKRs

Objective 1: Understand generative AI IP requirements
KR1.1: Land a plan for potential future EU Transparency obligations.
KR1.2: Land a framework and plan for Redacted - Privilege
KR1.3: Establish an IP runbook for product teams deploying generative AI products on the platform
Objective 2: Solve for most of the IP work needed for generative AI products with the platform
KR2.1: Define framework for using memorization measurement
KR2.2: Ship Redacted - Privilege
KR2.3: Measure Redacted - Privilege Redacted - Privilege
Objective 3: Unblock dataset reviews for 3P training data
KR3.1: Develop a pipeline of potential third party non-user data for review
KR3.2: Develop Redacted - Privilege
KR3.3: Ad hoc dataset reviews take less than one week and approved datasets are available in AIDC

Commented [40]: There are no projects that refer to this OKR. Is this out of scope for H1
[redacted]@meta.com
Assigned to [redacted]@meta.com

Projects

Priorities **P0**: Must-have • **P1**: Want-to-have • **P2**: Nice-to-have.
T-Shirt Size: **S**: < 2 weeks • **M**: 2~4 weeks • **L**: 4~8 weeks • **XL** : > 8 weeks

Priority	OKR	Project	Description	T-Shirt Size	Owning Team	Partner Teams	Alignment Status
P0	KR 1.1	Develop the plan for potential future EU Transparency obligations		M	Platform	IP Legal, Policy, Trust, AI Privacy Area	Not Started

Commented [41]: There's a missing pre-req here to Redacted - Privilege

Commented [42]: Can you add them?

P0	KR 1.3	Establish an IP runbook for product teams deploying generative AI products on the platform		S	Platform	IP Legal, Policy, Trust, AI Privacy Area	Not Started
P0	KR 2.1	Define framework for using memorization measurement	Redacted - Privilege		Platform	IP Legal, Policy, Trust, AI Privacy Area	Not Started
P0	KR 2.2	Land Redacted - Privilege Redacted - Privilege	Redacted - Privilege	L	AI Privacy Area	AI Privacy Area	Not Started
P1	KR 1.2	Develop the framework and plan for t Redacted - Privilege		M	Platform	IP Legal, Policy, Trust, AI Privacy Area	In Progress
P1	KR 2.2	Prototype t Redacted - Privilege Redacted - Privilege	Redacted - Privilege	XL	AI Privacy Area	AI Privacy Area, Central Applied Science	In Progress
P1	KR 2.3	Prototype Redacted - Privilege	Redacted - Privilege	XL	AI Privacy Area	AI Privacy Area, Central Applied Science	In Progress

Related/Reference Documents

- a/c priv - Key Points of AI Impact Assessment
- Privacy/Copyright Risk Management Strategic Position 1-Pager Technical Strategy
- Eng Estimate for GenAI Privacy & Copyright Measurement
- AC PRIV - 11/15/23 - Options to mitigate IP risk in the EU for GAI training

Commented [44]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [43]: How is this measurement different from existing violation rate and refusal rate metrics that are owned by safety team? @[redacted]@meta.com cc: @[redacted]@meta.com _Assigned to [redacted]@meta.com_

Commented [45]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [46]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [47]: Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Commented [48]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [49]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [50]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [51]: Given Central Applied Science have been working with Central Privacy (last half), and Genie (in H1 2024) on these execution areas.

Commented [52]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.

Commented [53]: I added these when I was part of the Privacy org last half. IIUC the AI Privacy area is no longer driving execution on these projects.