[SantosDesign] Santos Catalog Moderation

Notes for Santos Design Review

Thank you for reviewing the doc! Key Topics that we are asking for feedback from Santos PE Group:

1. Review ⚓ Moderation Risk Evaluation Function Strategy Discussion Section to make sure we are making right decision on platform and model choice based on current use case.
2. Review Catalog Moderation Model Performance Evaluation Section to see any more evaluation or research we should do to gain confidence.
3. Review Moderation Risk Evaluation Function System Design Section and raise any concerns or improvement we should do.

Overview

FAM team's goal is to protect Buy with Prime (BwP) and Amazon against reputational, financial and compliance related risks while providing a safe and secure experience to BwP merchants and shoppers. To achieve it, we envisioned to build flexible monitoring and moderation (M&M) capabilities that can be tailored to support use-cases ranging from policy violation detection to site content/structure optimization. Our product team presented the PR/FAQ during Coco Q1 2023 onsite and we decided M&M characters are our project mascots after that, shown on the right side. :)

With the PRFAQ, FAM engineer team dived deep and separated the overall system into two major parts which could operate independently - Monitoring system and Moderation system.

- Monitoring **System** focus on continuously finding and understanding the data shown on sites we wish to monitor (BwP-enabled sites). Its main responsibility is to find all the page/content (crawl function), and transform them into the generic format (scrape function) that could be used by downstream system in a certain cadence. For example, to check BwP AUP, we need monitoring system crawl and scrape merchant DTC website weekly which moderation system will check compliance.
- **Moderation System** focus on moderating the data based on the configured policy and raises cases to risk managers to review and take actions if detecting violation.

In this document, we focus on the moderation system, which is our first M&M capability launch, and describe the system we build to support Santos catalog moderation.

WHY WE CHOSE SANTOS CATALOG MODERATION AS FIRST FEATURE LAUNCH

We selected Santos Catalog Moderation as our first moderation use case due to the following reasons (in order of importance): **(i) Lima launch:** Lima prioritization (ranked #5 in KPR; beta launch ETA 1/30/24) expedited the need to moderate Santos Catalog as Santos Catalog data fields (product title, product description, and all product images related to that product) will show up as content on Lima PDP pages, **(ii) Ballooning scan cost:** if we were to use our existing EverC solution and daily scan cadence to monitor Lima PDP URLs, the cost would reach **~$32MM by end of 2024** (see calculation here). Lima aside, the cost to monitor BwP-enabled DTC sites through EverC is expected to reach $1.5MM in 2025, **(iii) Increase in # of shopper facing avenues with unmoderated Santos catalog data:** When we initially accepted the risk of unmoderated Santos Catalog, the unmoderated data was just showing up on Hosted Widget post order emails *(we have not received any escalation for BwP Acceptable Use Policy (AUP) non-compliant content on post order emails).* However, it was time to re-visit our decision with the launch of new shopper facing avenues - Lima, COM & Shopper Hub - which also pull data from Santos Catalog (see Appendix 1 for screenshots), and **(iv) Availability of data for moderation:** Santos Catalog data is already present within catalog systems and thus there was no need for a monitoring solution to crawl and scrape data separately.

Business Requirement

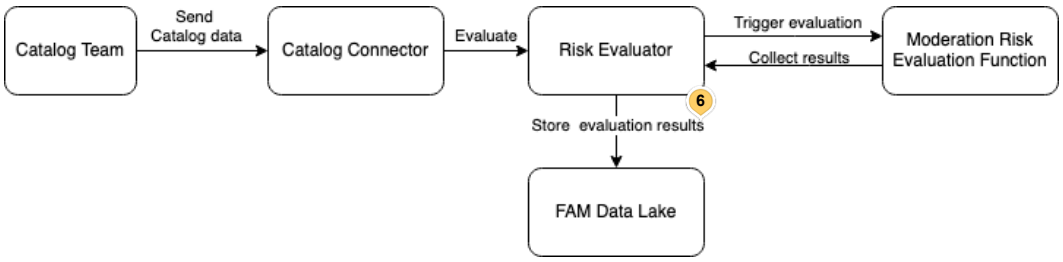
FAM Team's end-state goal is to continually moderate every updated/newly created Santos Catalog to ensure that Santos

Catalog data is in adherence with the BwP Acceptable Use Policy (AUP), and raise alarms to Risk Investigation Team (RIT) to manual moderate and take actions if needed. Detailed policies are listed below:

	A	B	C
1	**Fraud and Abuse Scenario**	**V1 Focus Moderation Use-Case**	**Theme**
2	Merchant lists products or publishes content featuring nudity	Detecting Child Sexual Abuse Materials (CSAM) in images on merchants' sites	BwP AUP
3	Merchant lists products or publishes content that is hateful toward a protected class of people or is violation of the rights of others	Detecting profane language/slurs/hate speech directed toward protected classes of people in text and images on merchants' sites	BwP AUP
4	Merchant lists products or publishes content that depicts or promotes graphic violence	Detecting products, images, and text depicting or promoting self-harm on merchants' sites	BwP AUP
5	Merchant lists counterfeit products on their site	Detecting presence of a high risk brand name or trademark in text and images on merchants' sites	FOLEX P0 requirement (ETA: April 2024)

Note - BRD link which has more details

Catalog Moderation E2E Workflow



To moderate catalog, *FAM Catalog Connector* retrieves records from Santos catalog after receiving create/update event notifications, and calls FAM DomainService, *FAM Risk Evaluator (FRE)*, a service to accept sync/async risk evaluation and route them to different evaluation function based on TLR, to trigger the moderation request. Then FRE passes along the data to *Moderation Risk Evaluation Function* to evaluate the risk and store the results in the FAM data lake asynchronously once moderation is done. And if there is a violaten detected, Risk Investigation Team will review and take enforcement action if needed, (such as warning email, business account suspension)

In these E2E workflow, three main components are needed:

1. Catalog Connector to connect Santos Catalog team to FAM Risk Evaluator. For this component, we are reusing our existing SNS/SQS-Lambda service, FAMEventListener, to connect FRE with Santos Catalogs Item Notification Service (INS) and Item Administration Service (IAS). Detailed design
2. FAM Risk Evaluator (FRE) to trigger the moderation request and store the results. This is our domain service, which we are launching in Q4 and planning to onboard catalog moderation as first use case. Detailed design
3. Moderation Risk Evaluation Function, the core function to run moderation, which we need to build. And we will focus on this component in the following sections.

Moderation Risk Evaluation Function Strategy Discussion

We need to build a generic moderation function to evaluate the compliance of extracted objects/entities against specific predefined policies, and give the moderation results for real-time streaming data. To build this function, we evaluated different platforms and models to see whether we could reuse instead of re-invent the wheels.

Based on our research, we made a two-way door decision to build an FAM moderation function, underlying integrated with Santos Oberon Service which is designed for review moderation (detailed intro in later part), but also providing the flexibility to leverage different tech providers if they fit better for different use case (such as brand guideline violation detection).

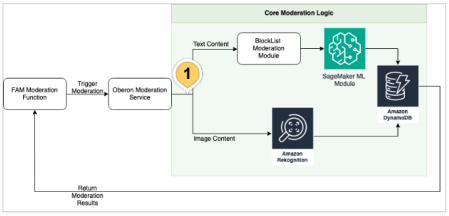
We think Oberon Service works best for us for catalog moderation due to these major benefits:

1. Oberon Service already supports all our current needed policies.

2. Oberon Service, acting as a platform, is flexible to onboard new models if we find better performance ones for certain use case.
3. Oberon Service is owned by Santos which is easier to work together for new feature requests.

Plus, building a well-trained moderation function from scratch would require significant effort from engineers and data scientists. And Leveraging Amazon existing platform also has limitation. Amazon Classification and Policy Platform (CPP), the leading platform used by Amazon Restricted Product team, are tied to Amazon ecosystem and designed for product moderation, which will be a limitation when we want to expand our moderation system to moderate non-product info (e.g. Homepage, FAQ, blog) scraped from merchant DTC website, which are required when doing BwP AUP compliance check. In addition, data handling is a concern too if we use CPP as the data access permission is shared. So using Oberon to handle moderation evaluation logic helps FAM accelerate the moderation process in its initial stages. Detailed comparison is listed Appendix 3: FAM Moderation Possible Solutions Comparison

Quick intro - Moderation in Oberon
Oberon is the service that BwP Review team (Soapbox) built to serve review moderation requirements. This service has already provided a set of comprehensive moderation policies used by Amazon review team. Currently it supports the moderation for Text and Image content types. Detailed of the workflow are explained in Appendix with quick intro of Amazon Rekognition.



Why not directly using Oberon Service as moderation risk evaluation function?
Although Oberon Service is designed as a generic moderation service and supports all our current needed policies, we don't think it's right decision to directly use it as the moderation risk evaluation function. First, FAM has feature requirement which is not what Oberon plans to support as a generic function, such as paragon case creation, multiple moderation task aggregation (stateful moderation). Secondly, as our moderation requirement grows, we may find different solutions that fit us for certain use case (such as brand guideline violation detection). So we want to build FAM moderation function flexible and extensible to provide the possibility to integrate with different moderation technical providers if needed for future moderation request. The approach enables FAM to protect Santos products efficiently while providing fast moderation response with minimal efforts.

How to gain confidence that the model built for reviews works for catalog?
For image moderation, Oberon leverages Amazon Rekognition, which is built for generic use case. For text moderation, although they have models trained for review, Oberon provides functions to configure customized allowList, manualList, blockList which we could use to match our use cases.

In addition, to gain confidence on the model performance, we are separating the catalog moderation launch into three phases:

- **Phase 1** - **Solution Performance Evaluation + One-time Manual Santos Catalog Moderation**: The focus was on a one-time Santos Catalog moderation with the intent of (i) getting granular insights on the presence of any BwP AUP non-compliant content in Santos Catalog, (ii) validating model accuracy and suitability for our requirements, and (iii) give some feedback to scientists to improve the model. We have COMPLETED this phase and will be reviewing the details in the next section.
- **Phase 2 - Monthly Cadence Manual Santos Catalog Moderation**: In phase 2, we plan to do periodic check(s) of the newly created/updated Santos catalog data since phase 1 completion date to cover the business need for catalog moderation before phase 3 delivery. We plan to run 4 monthly checks (on 11/01/23, 11/30/23, 12/30/23 and 1/29/24) before phase 3 launch (on 1/30/24) which will help (i) cover the unmoderated risk; (ii) improve model performance; and (iii) identify any other potential optimization opportunities.
- **Phase 3** - **Automatic Santos Catalog Moderation**: The end state solution with automated & real-time Santos Catalog moderation during creation/update time with Paragon case creation for FAM Investigators to investigate ML-model flagged content. We will dive deep and provide details on the technical solution we are building in a later section.

Catalog Moderation Model Performance Evaluation

Oberon serves as a well-trained moderation service for reviews data. However, it remained unknown on its performance on evaluating BwP Catalog data against BwP AUP. So while we are building an automatic moderation E2E workflow, we started with one time manual execution of Santos Catalog Data (product title, product description and product images) to initiate a proof-of-concept (POC) and evaluate Oberon solution performance on catalog data. The goal was to find out the right configuration (in terms for blocklist/allowlist and ML policies to be applied on Santos catalog data), validate its accuracy and

suitability for FAM requirements, and identify the potential optimization areas before phase 3 launch.

PROOF OF CONCEPT

Before full Santos catalog run, we decided to run a POC on smaller data set to get initial performance result of Oberon system quickly. We generated two data sets - 1) 857 Santos Catalog SKU randomly selected from different merchants across all DTC site traffic buckets; 2) 149 verified BwP AUP non-compliant/egregious SKU from the Regulatory Intelligence, Safety, and Compliance (RISC) (formerly RP) team of Amazon.com (data). For 857 Santos Catalog SKU, we saw **86.12% APPROVE rate and 13.88% REJECT rate (meaning non-compliant)**; and for 149 non-compliant/egregious SKUs, we saw **APPROVE rate is 3.36% and 96.64% REJECT rate.** Key takeaway were**:**

1. **13.88% REJECT rate for Santos Catalog data:** We were initially surprised by the high reject rate of 13.88% for Santos Catalog data but after a manual inspection we realized that some of the policies we copied from Reviews policy (like TOBACCO, DRUGS etc.) do not apply to our use-case of BwP AUP. We took an AI to identify and eliminate these extraneous policies before running P1
2. **3.36% APPROVE rate for Verified Egregious data:** The 3.36% false negative rate was due to 5 cases (out of 149) that would have been flagged had we included keyword matching expressions like 'Hitler', 'Nazi' and 'Suicide'. We took an AI and added these keywords into our FAM Keyword List (data) and in additional, we are continously learning from Amazon.com to proactively update the keyword in long term.
3. **Need to resolve technical challenges:** a) Arcade was storing only one image for a product whereas MCUI allowed to add up-to 25 images. FAM worked with the Santos Analytics team to include all images in the data transformation job along with backfilling the past data, and b) Internal bugs in Oberon where the service was throwing NPE when a item is deleted and images are not available for downloading. These issues were fixed by code changes in Oberon service.

FULL SANTOS CATALOG RUN RESULTS

FAM carried out complete (as of 9/18/23) Santos Catalog data moderation of **456,371 Prime Intent SKUs** after working with Oberon to solve couple issue faced during runtime (details)**. First run of Santos Catalog Moderation shows 68,133 (14.9%) REJECT cases out of 456,371 SKUs.** We manually scanned through data samples that had been REJECTED and realized: (i) For image moderation, some of the secondary-level policies that AWS Recognition moderated (like 'male/female swimwear', 'revealing clothes') do not apply to our use-case; (ii) the Oberon TextML model was flagging non-egregious words as BAD_LANGUAGE since the model is trained for review case (e.g. series number detection) but not applicable for catalog moderation; and (iii) some of the FAM owned keywords were producing false positives

Based on the learning, we worked with Oberon team to (i) deselect non-applicable secondary policies (see Appendix 2 for full list of policies deselected) for image moderation; (ii) for text moderation, we eliminated the false positive ones from FAM keyword list and (iii) remove Oberon textML from the flow to only use FAM owned keyword list and Oberon's generic blocklist keyword library. We re-ran the Catalog moderation after making these changes and the **second run of Santos Catalog Moderation shows 13,921 (3%; decreased from 14.9% in first round) REJECT cases out of 456,371 SKUs.** Reviewed with RIT team and we think 3% is an acceptable number for manual moderation, so we used this round of results for performance evaluation.

PERFORMANCE EVALUATION

Performance Baseline

Before evaluating Oberon model performance, we did some research and tried to collect data from two resources to setup baseline. However, we found it's hard to do apple to apple comparison. And in the absence of that, we are relying on ensuring an acceptable/manageable false positive rate, while avoiding false negatives.

More context of the two resources we researched:

1. **3P Model - EverC**: We currently use EverC as our vendor to do BwP AUP check on merchant DTC websites. Although we know what websites they are scanning, they don't share how many pages in total they monitor and moderate, which made it impossible to calculate the REJECT/APPROVE rate. Only data we could collect from EverC is with all the issues they raise, how many are real issues. Among the 690 issues they raised from Aug to Oct, none of them are violating BwP AUP. (Details could be found in EverC issues tracker).
2. **Amazon Model - RISC team**: Amazon RISC team owns the Amazon Catalog moderation for restricted products. Based on our learning, they sample the catalog data to moderate and the moderate process are divided into three major processes - 1) Manual Process - For the policy they don't have confident to run ML model or text regex match, they go through manual process; 2) Text Regex Check Process - For text moderation, regex match check is the major solution they are using; 3) ML Model - They own ~100 ML models to do different GL and type of policy moderation. For image, they use OpenAI CLIP to connect image and text first, moderate text next, and decide on the results. They also use

Amazon Rekognition for image moderation and apply additional model based on the result to provide more accurate results. When we review our current policy with them, nudity is the only policy they support now through ML model, which means we won't be able to do apple to apple comparison with all the policies we want to moderate. But we are working with them to run a nudity related experiments to compare that with Oberon model.

Performance Evaluation Summary

From the performance evaluation run, we are confident that Oberon model works for Santos Catalog moderation with current policies since we didn't miss any risk when sampling the approve cases. Although false positive rate is high, it's the same situation as we see from 3P model - EverC (100% false alarm). In addition, we do see opportunities that could help continuously improve the performance.

Performance Evaluation Details on Oberon owned ML Model + Oberon's Generic Keyword Blocklist

In order to evaluate Oberon ML model's performance, we randomly spot checked a sample of **200 APPROVED and 200 REJECT** cases from full Santos Catalog run results with the help of the FAM Risk Investigations Team (RIT). FAM RIT manually annotated each of the 400 cases depending on whether a particular case should have been flagged for human moderation (see detailed results here). Our key takeaway were:

1. **0% of APPROVED cases violated BwP AUP policy (False Negative: 0%):** We were glad to see that model only APPROVED cases that should have been approved and did not allow any egregious content/BwP non-compliant content to pass through.
2. **0% of REJECTED cases violated BwP AUP policy (False Positive: 100%*) while 11% REJECTED cases were benefited from human moderation (False Positive: 89%):** At the other end of spectrum, 0% of the REJECTED cases actually violated BwP AUP. While it is tempting to state False Positive is 100%, on deeper look we believe 11% of REJECTED cases could be benefited from human moderation (even though none of them violated BwP AUP) a.k.a we are happy that the model flagged these cases. While our aim is to reduce the % of false positives, we also need to be cognizant of the fact that trying to minimize % of false positives might lead to actual violations not getting flagged (i.e. false negatives) which is a reputational concern, so we will continually evaluate model performance while we are working with Oberon team for optimization.

FAM Owned Keywords Details [BwP AUP related keywords + brand name keywords we monitor related to FOLEX]

To find performance of FAM Owned Keywords, we audited additional 200 cases that were REJECTED by the model due to FAM Keyword List. Our key takeaways were:

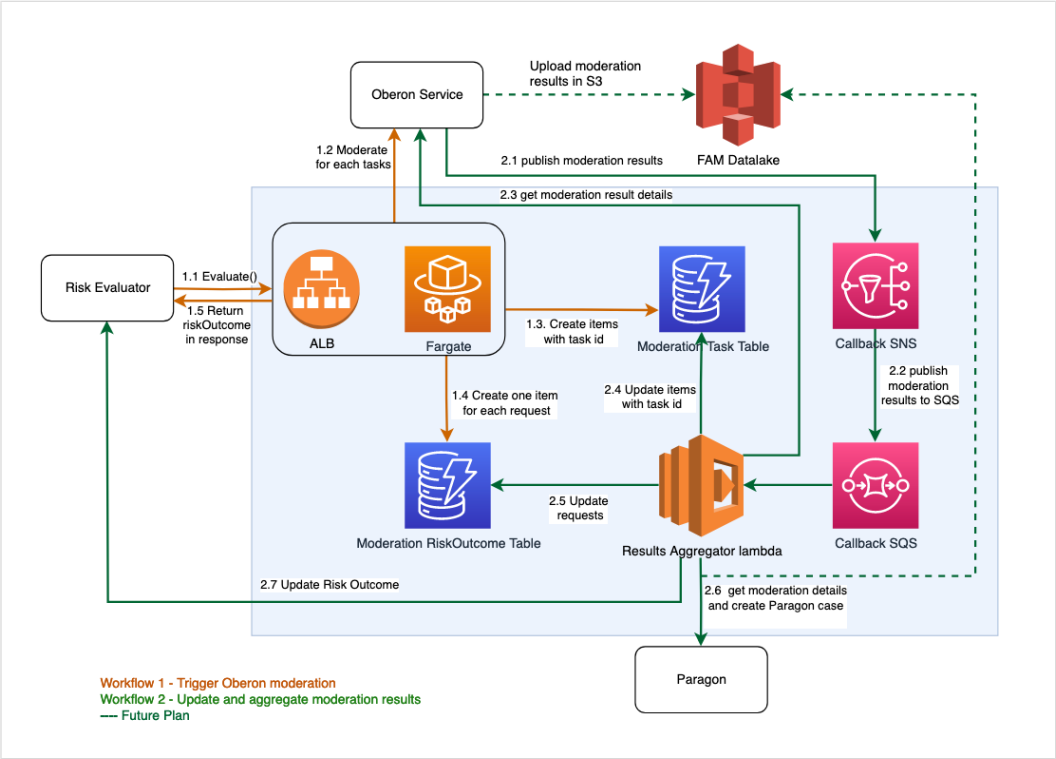
1. **0% of REJECTED cases violated BwP AUP (False Positive: 100%*) while 27% REJECTED cases were benefited from human moderation (False Positive: 73%):** The 27% (54 cases) that benefited from human moderation is made up of 11% (6 cases) of BwP AUP keywords and 89% (48 cases) of brand name keywords we monitor related to FOLEX. The 73% (146 cases) actual false positive is made up of 15% (22 cases) of BwP AUP keywords and 85% (124 cases) of FOLEX keywords. See case examples here

Optimization Opportunities

1. **Oberon's Generic Keyword Library Performance:** FAM observed some of the false positives happened due to Oberon's generic blocklist library. We plan to continuously review those REJECTED cases and work with Oberon team to either remove the keywords from the generic keyword library or add them to FAM TLR allowlist if not applicable for FAM use cases.
2. **BwP AUP BlockList Keywords:** Similar. We will continue to update the BwP AUP keyword blocklist list as we review the REJECTED cases.
3. **Image Moderation Performance:** Currently, we observed high false alarm rate (of 89%) on REJECT image moderation. To reduce the false positives FAM planned to tune the right confidence score threshold for some policies by creating the score distribution graphs using sample data from phase 1 results. In addition, we are exploring two options to see whether they could improve the performance - 1) Leveraging GenAI tech (AIGC to describe the image and run text moderation; or CLIP to connect text and image) after Amazon Rekognition moderation results; 2) Leveraging Amazon model - RISC to moderate nudity which we saw a high false alarm rate too.
4. **FOLEX Counterfeit Detection Performance:** Planning to build a quick solution, we currently use high risk brand names as keywords to detect related product and manual moderate them as first phase for counterfeit detection solution. While we only apply FOLEX keywords on product title to avoid any false alarms (eg: Apple is a common word used in product description in a non-brand context), we still see a number of false positives. We are still brainstorming how we could improve the performance. (Please reach out to us if you have any suggestion. Thank you! :D)

Moderation Risk Evaluation Function System Design

While we are running manual moderation to test model performance, we need to build a system to automate the process, which is our target goal for phase 3. Below is the system architecture.



FAM Moderation Risk Evaluation Function have two workflows: Trigger the evaluation (labeled as 1.x) and Evaluation Result Update (labeled as 2.x)

- **Trigger the evaluation (labeled as 1.x):** FAMRiskEvaluator (FRE) receives the moderation request and invokes the evaluate SPI implemented by Moderation Risk Evaluation Function. The moderation function then creates the moderation record in DynamoDB and initiates moderation using the identified moderation technical provider (such as the Oberon Service for catalog data). Depending on the moderation content, multiple calls may be triggered to create more than one moderation sub-tasks. Once all moderation sub-tasks are generated, the moderation function will respond to FRE with a status of IN_PROGRESS.
- **Evaluation Result Update (labeled as 2.x):** Once the Moderation technical provider completes the moderation process for sub-tasks, it sends the results to the FAM team. Taking Oberon as an example, when Oberon publishes the moderation completion message through SNS, the moderation function calls its API back to gather the detailed moderation results. Following this, the moderation function batch updates the results in the DynamoDB, and aggregates outcomes when we validate all sub-tasks are completed. Then the aggregated moderation outcome(REJECT/APPROVE/MANUAL) will be sent to FRE with a status of COMPLETE. If manual review is necessary, the moderation function creates a paragon case and provide enough infos for FAM RIT agents to evaluate whether these manual cases should be REJECT or APPROVE, track the results manually. In addition, we build script to do periodical sample checking on the APPROVE and REJECT cases to provide feedback and improve the model continuously.

In the future, we will automate feedback loop. When investigators close paragon cases, outcomes of manual reviews can be utilized for ML model training and moderation performance optimizing.

What's next?

As we closed out phase 1 with confidence of the Oberon model performance and identifying the optimization opportunities, engineer team is mainly working on phase 3 implementation to have an automatic and real-time catalog moderation system build-up, which we target to launch on 2024/01/30. In addition, we are manually running monthly check to continuously improve our configuration setting and model performance.

Besides catalog moderation, the moderation of merchant DTC website content is the next moderation feature. It will moderate the data crawled and scraped from monitoring system. The monitoring system is in design phase. We did tech solution dive deep and decided to integrate with Amazon Selection Monitoring Team (SMT) team for monitoring function. We did an POC with SMT and results look promising. We are working with their leadership to settle down the goal and will deliver a tech design by Q4 2023, and work on implementation on Q1 2024. Once we begin monitoring merchant websites, the moderation service is equipped to moderate a substantial volume of websites data within a single request from the monitoring system.

Also, we are working on enforcement process domain model exercise so that we could close the risk evaluation and mitigation loop by taking action based on the moderation results.

Could stop here for the review

Meeting Notes from Santos Design Review

1. [Owner: Xiaoxi Bai] Integration Pattern with Upstream - Current Catalog Connector might only works for Fire and Forgot Pattern. Need to follow up with Catalog what if we want to support sync in long run.
2. [Owner: Aniruddh Menon + Akanksha Sharma] Human Power - we should measure human FP and FN rate as well. And continuously compare them.
3. [Owner: Pavan Deshpande] Enforcement - Need more details. Could share to Catalog team once we have some details.

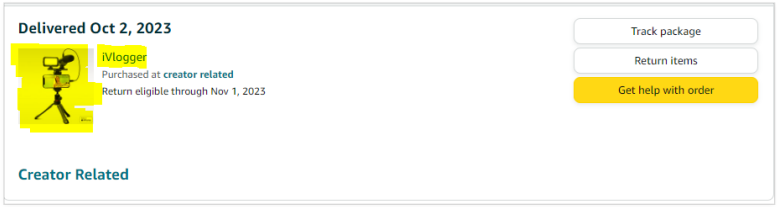
Resource

- Monitoring and Moderation PRFAQ - link
- Monitoring and Moderation BRD - link
- Monitoring and Moderation High Level Discussion - Monitoring and Moderation High Level Discussion
- Moderation Strategy Discussion [2022 version] - Moderation Solution Strategy Discussion
- Moderation Strategy Discussion [2023 version] -FAM Moderation Solution Discussion
- Catalog Moderation System Design - FAM Moderation System Design
- Catalog Connector Design - FAM Catalog Connector Design
- Catalog Moderation P2 Dive deep - Catalog Moderation Phase-2
- Monitoring Solution Research - Monitoring Solution Research

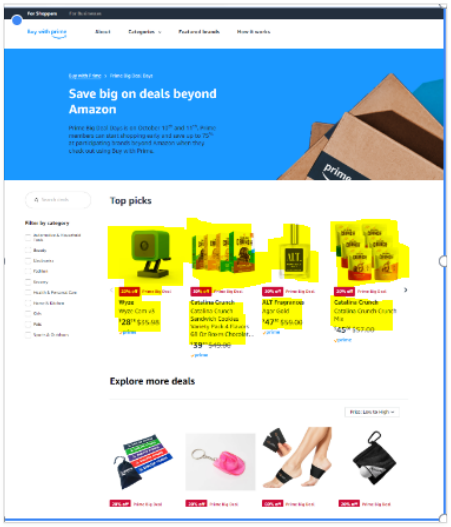
Appendix

APPENDIX 1: HOSTED CHECKOUT EMAILS, COM, SHOPPER HUB, LIMA PAGES

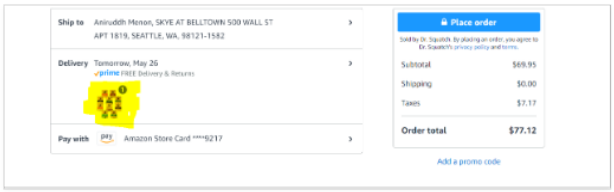
COM Screenshot

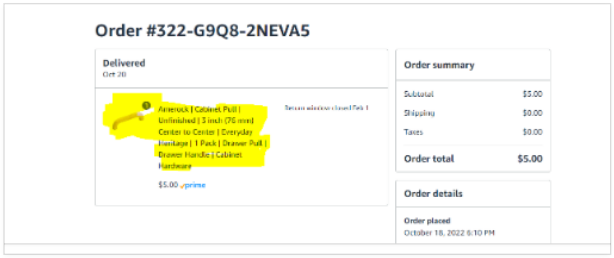


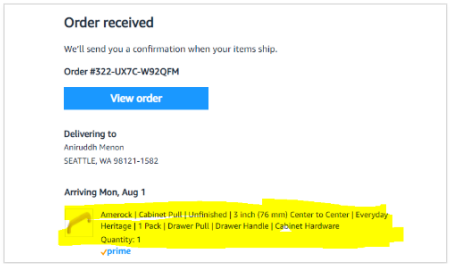
BwP Shopper Hub Screenshot



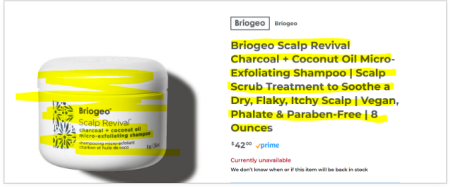
Hosted Checkout - Checkout Page, Order Confirmation and Shopper Tracking Email







LIMA



APPENDIX 2: OBERON POLICY LIST

	A	B
1	**Oberon Policies (based on Amazon Rekognition Policies)**	**Applicable to BwP AUP ? (Yes/No)**
2	Nudity	Yes
3	Graphic Male Nudity	Yes
4	Graphic Female Nudity	Yes
5	Sexual Activity	Yes
6	Illustrated Explicit Nudity*	Yes
7	Adult Toys	No
8	Female Swimwear Or Underwear	No
9	Male Swimwear Or Underwear	No
10	Partial Nudity	Yes
11	Barechested Male*	Yes
12	Revealing Clothes	No
13	Sexual Situations	Yes
14	Graphic Violence Or Gore	Yes
15	Physical Violence	Yes
16	Weapon Violence	Yes
17	Weapons	No
18	Self Injury	Yes
19	Emaciated Bodies	Yes
20	Corpses	Yes
21	Hanging	Yes
22	Air Crash	Yes
23	Explosions And Blasts	Yes
24	Middle Finger	Yes
25	Drug Products	No
26	Drug Use	No
27	Pills	No
28	Drug Paraphernalia	No
29	Tobacco Products	No
30	Smoking	No
31	Drinking	No
32	Alcoholic Beverages	No

33	Gambling	No
34	Nazi Party	Yes
35	White Supremacy	Yes
36	Extremist	Yes

APPENDIX 3: FAM MODERATION POSSIBLE SOLUTIONS COMPARISON

	Current Solution	Approach 1	Approach 2	Approach 3	Approach 4
Name	Apay/EverC	Reuse Amazon Classification and Policy Platform(CPP)	Create FAM moderation platform	Create FAM platform with Oberon ML Model	Reuse Oberon Platform
Description	FAM is using EverC and Incopro to do AUP monitoring checking	Reuse CPP platform which provides self-service classification of products in the catalog and management of policies.	Build own platform using existing fundamental methods and create FAM owned ML models.	Build FAM-owned platform with Oberon models from Review team.	Oberon moderation system is the service Review team build to serve moderation requirements.
Moderation methods	Unknown	FAM can reuse existed classifications If it's covered in our current use case already	FAM will build ML models. Needs collaboration with ML scientists to develop models based on our use cases from scratch	Oberon ML model is already onboarded 5 machine learning models from CS team and we can reuse their sagemaker endpoint to use their ML models	Platform provides Oberon ML model as a set of comprehensive moderation policies proven in Amazon's decades of e-commerce operation
Manual Moderate	Once EverC send the violations back, FAM team will review to determine if the violation is true	CPP has the UI portal for investigators. Need extra efforts to connect CPP platform with paragon	Needs to connect paragon for manual moderation during implementation	Needs to connect paragon for manual moderation during implementation	Oberon Platform has the UI for investigators. Extra efforts needed to connect paragon case
Flexibility	**N/A** No visibility and flexibility on what pages and the content they scanned	**+** New feature requests need to fill Project Intake Process template	**++++** FAM can respond rapidly and take action items with new features internally	**+++** FAM can respond rapidly and take action items with new features internally. But ML models maintenance needs more help from Review team.	**++** FAM needs Review team to support new features in Oberon service.
Development Time estimation	N/A	**+++** CPP has a detailed onboarding plan to support new project. On-demand classification needs extra SDE efforts.	**+++++** 1. Code implementation 2. Full Security review 3. Large size efforts from Scientists	**+++** 1. Code implementation 2. Full Security review 3. Medium size efforts from Scientists in Review team	**++** 1. SDEs efforts from Review teams 2. Delta Security review 3. Medium size efforts from Scientists in Review team
Ownership and Maintenance	N/A	FAM will be a user in CPP. New feature addition/enhancement request in CPP platform needs to follow intake process	FAM owns and maintains platform and ML models	FAM owns and maintains the Platform; Review team owns ML model and takes over new model requests from FAM	Review team owns platform and ML model. FAM call Oberon as a dependency
Long term plan	Cannot fit for [Feature 2]	It's not a long term solution	Can be extended to a santos wide moderation system	Can be extended to a santos wide moderation system	Can be extended to a santos wide moderation system
Pros	1. + EverC is the third-party which strengthen and streamline risk management 2. + Friendly UI with moderation results.	1. + Well maintained platform with clear UI which is widely used in Amazon. 2. + Move fast: FAM can reuse existed classification from offensive team to get moderation results without too much SDE/ML efforts	1. + FAM will have complete ownership for the moderation platform 2. + FAM has high flexibility to customize ML models and add new features in the future	1. + FAM will have complete ownership for the moderation system architecture 2. + Move fast: Leveraging existing ML models from review teams 3. + Accurate: Beta version implementation duplicates Amazon CS ML model with 100% decision match rate.	1. + Move fast: Oberon has already used Rekognition and ML models to moderate which mentioned in approach 2 2. + Accurate: Beta version implementation duplicates Amazon CS ML model with 100% decision match rate. 3. + Avoid duplicate work to build moderation system.
Cons	1. - Lack of data transparency 2. - Limited of functionality 3. - Ballooning Cost	1. - CPP supports data in Catalog with ASIN based classifications only. Catalog data cannot cover header/footer and non-product pages in [Feature 3] 2. - Uncertainty of the timeline within FAM's new feature requests 3. - Extra costs and efforts associated when integrating with Amazon CDO platform with Santos data.	1. - Substantial efforts from SDEs due to the complex nature of integrating various methods for text, image, and video moderation together 2. - Substantial efforts from machine learning scientists to create ML models from scratch	1. - Substantial efforts from FAM SDEs to build platform 2. - Scientists from Review team need to provide supports in maintaining moderation accuracy for FAM.	1. - Extra away team efforts to extend inputs limited, and support paragon cases in domain level 2. - The Review team owns the core moderation logic. For any new features, external team support will be required.
Decision	Goal is to replace by FAM moderation service	Not recommended for long term	Not recommended	Considering	Accept

APPENDIX 4: FAM MODERATION FUNCTION COMPONENTS

Here is the details of the two workflow:

- **Trigger the evaluation (labeled as 1.x):** FAMRiskEvaluator (FRE) receives the moderation request and invokes the evaluate SPI implemented by Moderation Risk Evaluation Function. The moderation function then creates the moderation record in DynamoDB and initiates moderation using the identified moderation technical provider (such as the Oberon Service for catalog data). Depending on the moderation content, multiple calls may be triggered to create more than one moderation tasks. Once all moderation tasks are generated, the moderation function will respond to FRE with a status of IN_PROGRESS.
- **Evaluation Result Update (labeled as 2.x):** Once the Moderation technical provider completes the moderation

process for tasks, it sends the results to the FAM team. Taking Oberon as an example, when Oberon publishes the moderation completion message through SNS, the moderation function calls their API back to gather the detailed moderation results. Following this, the moderation function saves the results in DynamoDB and aggregates task outcomes after all tasks are completed. The moderation function then sends this aggregated moderation outcome to FRE with a status of COMPLETE. If manual review is necessary, the moderation function creates a paragon case and forwards it to FAM RIT agents for further investigation.

Here are components details in moderation function:

- Core FAM Moderation API service (ALB + Fargate)
  - Implement FAM Domain Service Evaluate SPI to trigger FAM moderation function, which underlying separates them to single record if needed and call Oberon moderation service for each one and create records in dynamo DB tables to track the status.
- DynamoDB tables
  - Moderation Tasks Table to store task level data with moderation result
  - Moderation Assessments Table to save assessment level data with an aggregated moderation result
- SNS-SQS for Oberon team to publish moderation results back.
- Results Aggregator Lambda to update records and aggregate task results and create paragon case if manual review is required.

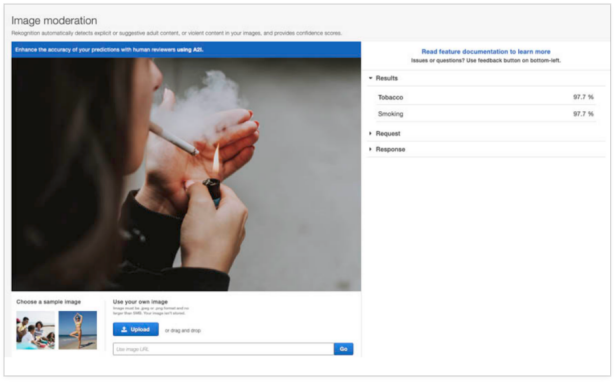
APPENDIX 5: OBERON DETAILED WORKFLOW

Text content workflow:

1. **Blocklist Moderation Module:** RegEx based library to detect inappropriate keywords present in text content. This library is build upon Amazon Community Shopping team's blocklist used for Amazon review moderation (code link).
2. **ML Moderation Module:** ML moderation module used to detect inappropriate and off-topic text content, build by Soapbox data science team. Oberon onboarded 5 ML models from CS team which has been widely used for content moderation. For text moderation workflow, there are ReviewsTextModel which is used to predict APPROVE/REJECT decision for given text content, and ReviewsRejectReasonModel which used to predict the reason code for rejected review. Here is the supported policies summary: Supported Policies in summary: Oberon Moderation Service Micro-Policy Support

Image content workflow:

1. **Amazon Rekognition:** Amazon Rekognition offers deep learning-based image analysis. This AWS service covers more general categories with great precision and has great capability to detect inappropriate, unwanted, or offensive image content. It provides a detailed taxonomy of moderation categories, such as Explicit Nudity, Suggestive, Violence, Rude Gestures, and Hate Symbols.



> Rekognition provides a hierarchical list of labels with confidence scores to enable fine-grained control over what images you want to allow. Besides, The Rekognition custom labels can help us identify the customized objects in images that are specific to FAM moderation requirements (e.g. brand logos for counterfeit detection).

APPENDIX 6: DATA SAMPLES

Cautions - please review with caution as some contents are NSFW

- 149 verified BwP AUP non-compliant/egregious SKU - Privileged & Confidential - Product List + Keywords

- FAM Keyword List - Catalog Moderation FAM Keywords
- Example to define "***human moderation is beneficial to a case"*** *(please note that all 4 examples below were REJECTED by the model)*

Example #	Did the case benefit from human moderation? (a.k.a are we glad that the case was flagged by the ML model ?)	Did the case violate the BwP AUP?
Example 1	YES	NO
Example 2	YES	NO
Example 3	NO	NO
Example 4	NO	NO

Sample on the cases needed manual moderation

Cases benefited from human moderation	SKU Title/Description
BwP AUP	*Honey Dew Gifts, A Wise Woman Once Said **F@@@** This Shit and She Lived Happily Ever After, 2.5 inch by 3.5 inch, Made in USA, Refrigerator Magnets, Decorative Funny Magnets, Fridge Magnets Adult*
FOLEX	***Sony** XB13 Extra BASS Portable IP67 Waterproof/Dustproof Wireless Speaker Bundle*
Cases NOT benefited from human moderation (i.e. actual false positive)	**SKU Title/Description**
BwP AUP	*Vance & Hines VO2 **Naked** Air Cleaner: 91-20 Harley-Davidson Sportster Models)), Con*
FOLEX *(keyword is TIFFANY but the brand is ESCENA)*	*ESCENA **Tiffany** Style Table Lamp, Antique Vintage Light Decor Swing Arm Desk Light, Handmade Lampshade, LED G45 Bulb Included, Decoration for Living Rooms, Bedrooms, Coffee Bar*

APPENDIX 7: ISSUE FACED DURING P1 RUNNING

While running the moderation, we faced many setup issues due to huge data size such as (i) Oberon client timing out before all the items are submitted for moderation, (ii) special characters in item title and description breaking the CSV parsing, and (iii) throttling from Oberon downstream dependency Santos media service. These issues were fixed one by one along with script improvements like optimizations to process the data faster and retrying the failed tasks and replacing the results for failed tasks in place (for complete list of issues and learnings see here).

APPENDIX 8 - LEARNINGS FROM PHASE 1

S.no	Issue	Action Item	Status	Notes
1	Lot of false positives were observed in image recognition as oberon allowed to configure only top level ML policies and not secondary level policies	Oberon made a code change to let clients configure ML policies at a granular level.	COMPLETE	E.g Adult Toys label as secondary and Explicit Nudity as primary [ModerationHistory(ModerationStatus=REJECT, ModerationDescription=RekognitionLabel: Adult Toys Confidence: 64.79124%. RekognitionLabel: Explicit Nudity Confidence: 64.79124%. , ModerationTime=2023-09-20T16:53:30Z, Policies=[EXPLICIT_NUDITY], Workflow=IMAGE, ContentId=https://amazon-omni-cdn.com/hgize3tf2ba/ffi02pp0xuz763/51INcybKjAL.jpeg)
2	Bad language in text recognized due to ML models was mostly false positive	Oberon modified the policy to IN_APPROPRIATE and FAM configured to remove this ML policy for text workflows	COMPLETE	E.g If content has text like series number, IDs, units etc. It is likely to trigger this policy. e.g. ModerationContentSet(ContentSet=[ContentObject(Name=text_title, Value=ContentObjectValue(StringValue=Professional Equine Horse Stainless Steel 4.5 Twisted Wire Copper Mouth D-Ring Snaffle Bit **35319A**))) ModerationContentSet(ContentSet=[ContentObject(Name=text_title, Value=ContentObjectValue(StringValue=CHALLENGER Horse **32" x 30"** Western Contoured Wool Felt Therapeutic Saddle Pad Red **39218RD**))])
3	Some of the submitted moderation requests to Oberon remained in PENDING state due to unhandled exception `ThreadPoolExecutor` exception.	Oberon fixed the issue by handing the exception and returning FAILED status which helped FAM to re-submit the moderation request and getting appropriate results.	COMPLETE	

4	FAM also observed low performance/accuracy on some image recognition by manually verifying sample MANUAL+REJECT cases	1) After analysing the results some policies are not making sense for FAM and hence planned to be removed e.g BARECHESTED MALE + ILLUSTRATED_EXPLICIT_NUDITY 2) To reduce the false positives FAM planned to tune the right confidence score for some secondary level policies by creating the score distribution graphs using sample data from PI results 3) Investigate to use AI Generated Content to reduce false positive 4) Connect with .com catalog moderation team to understand the model they use and corresponding performance and see if oberon can integrate with them	Planned for Phase 2	E.g 50% threshold might be too low for [ModerationHistory(ModerationStatus=REJECT, ModerationDescription=RekognitionLabel: Weapons Confidence: 59.415993%. RekognitionLabel: Violence Confidence: 59.415993% , ModerationTime=2023-09-20T16:10:03Z, Policies=[VIOLENCE], Workflow=IMAGE, ContentId=https://amazon-omni-cdn.com/bek4y89pgli/znsx3kcslrj296/71Tn0TQJ8KL.jpeg)
5	FAM observed some of the false positives happened due to oberon's generic blocklist library.	FAM planned to continuously review those MANUAL/REJECTED cases and remove the keywords if not applicable for FAM usecase.	RIT team to give feedback based on P1 result for keyword	E.g Oberon's generic blocklist library flags jig, Golder Shower, etc keyword which seems to create False positives
6	FAM observed False positives due to some of the blocklist keywords (Brand names) added after the POC moderation for a way to enable counterfeit products	FAM planned to moderate only products title for the brand name keywords	1) Planned to moderate only titles through oberon and manually reconcile results for Phase 1 2) Planned for Phase 2 and Phase 3	Brand names in product description should not be considered a non-compliant but in product's title can be a potential counterfeit. E.g Product WATER PILLS FOR WATER WEIGHT LOSS have keyword apple in description: *Our natural water pills produce a diuretic effect without added caffeine. Our water pills also contain beneficial herbs such as dandelion leaf, buchu leaf, juniper berries, **apple** cider vinegar, green tea extract and cranberry powder that have strong antioxidant properties, vitamins, urinary tract health support and other health benefits.*

APPENDIX 9 - PHASE 2 ROUND 1 OUTPUT

Moderation Output:

- **232,831** catalog data are moderated
- **227,704 (97.80%)** approved
- **5,090 (2.19%)** rejected/manual
- **37 (0.01%)** has unreachable image URL.