Weights are read from the repository. The probabilities in the examples are illustrative, not measurements from X.

What raises and lowers your reach on X's For You feed: the tables

The reference tables from the report, on one page. Every value traces to a file in the repository. The calculator and simulation are on the main page.

Version 2026.11.03 · commit b220448 (2026-11-03) · Created by @SidneyCur

ACTION WEIGHTS
— PENALTIES
ENFORCEMENT OUTCOMES
REPLY-RELATED TRIGGERS
EXEMPTIONS
AUTHOR DIVERSITY MULTIPLIER
FINAL REPORT: DO
FINAL REPORT: DON'T
FINAL REPORT: WORKED EXAMPLES

A Action weights

The weight the ranker multiplies by a predicted probability.

TABLE 01 — ACTION WEIGHTS — COMMIT B220448

#	ACTION	WEIGHT	PARAMETER	NOTE
01	**Reply**	5.0	ReplyWeight	Every reply action
02	**Reply on an original post from a mutual follow**	20.0	5.0 + 15.0 boost	Original posts only, not replies or reposts
03	**Quote**	5.0	QuoteWeight	—
04	**Share via DM**	5.0	ShareViaDmWeight	—
05	**Share via copy link**	20.0	ShareViaCopyLinkWeight	Largest non-reply gain

#	ACTION	WEIGHT	PARAMETER	NOTE
06	**Share**	2.0	ShareWeight	—
07	**Follow author**	4.0	FollowAuthorWeight	Predicted follow after seeing the post
08	**Repost**	1.0	RetweetWeight	In-network only
09	**Favorite**	0.66	FavoriteWeight	—
10	**Click**	0.3	ClickWeight	—
11	**Original post (no action)**	none	no weight	Posting itself is not scored

B Penalties (malus)

Negative weights. Each one lowers a post's score in proportion to the predicted chance of that action.

TABLE 02 — PENALTIES (MALUS) COMMIT B220448

#	ACTION	WEIGHT	PARAMETER	EXAMPLE
01	**Not interested**	−52.272	NotInterestedWeight	A viewer taps not interested. P = 0.03 → −52.272 × 0.03 = −1.57
02	**Block**	−31.2	BlockAuthorWeight	A viewer blocks you. P = 0.005 → −31.2 × 0.005 = −0.16
03	**Mute**	−58.8	MuteAuthorWeight	A viewer mutes you. P = 0.02 → −58.8 × 0.02 = −1.18
04	**Report**	−234	ReportWeight	A viewer reports you. P = 0.01 → −234 × 0.01 = −2.34
05	**Not dwelled (scrolled past)**	−0.02	NotDwelledWeight	A viewer scrolls past. P = 0.50 → −0.02 × 0.50 = −0.01

C Enforcement outcomes

What happens to an account or post after a flag, with an example of what the author sees.

TABLE 03 — ENFORCEMENT OUTCOMES — COMMIT B220448

#	OUTCOME	EFFECT	DURATION	EXAMPLE
01	**SpamHighRecall (account)**	Drops your posts for people who don't follow you	30 days	Your account is labelled. Strangers' For You feeds skip your new posts for 30 days. Your followers still see them. Source: enforcement_user.yaml, author_rules.rs
02	**SpamHighRecall (post)**	Drops that post as spam	30 days	One reply is labelled as spam and disappears from feeds for 30 days. Your other posts are not affected. Source: enforcement_post.yaml, golden_corpus/oon_tweet_label.rs
03	**RiskyHighVizReply (reply)**	Applied to a flagged reply	30 days	A reply under a popular post is labelled risky. The repo does not trace its visibility effect, so treat it as a warning flag. Source: enforcement_post.yaml
04	**Liveness check / CAPTCHA / Arkose**	Challenge before continuing	Until passed	The account must complete a check before it can post again. Source: enforcement_user.yaml

#	OUTCOME	EFFECT	DURATION	EXAMPLE
05	**Temporary suspension**	perm: false, policy PlatformManipulation	Temporary	The account is locked under the platform-manipulation policy. The repo does not set the length. Source: enforcement_user.yaml
06	**Permanent suspension**	CSE and campaign cases	Permanent	Reserved for the most serious cases, such as child-safety reports and campaign abuse. The account does not return. Source: enforcement_user.yaml
07	**Notice**	Keeps the post in the feed with a notice attached to it	While the rule matches	A fourth outcome alongside show, filter and drop. The post is not removed or downranked; a notice is attached instead. Nothing in this repository draws the notice itself. Source: visibility-filtering/rules/engine.rs, README

Reply-related triggers

Reply behaviours the classifiers and rules flag, with an example of each.

TABLE 04 — REPLY-RELATED TRIGGERS COMMIT B220448

#	TRIGGER	WHAT IS FLAGGED	LABEL	EXAMPLE
01	**Fast reply spam**	Replies in quick bursts	`SpamHighRecall`	A burst of replies in a few minutes, under many different posts. The repo does not publish the speed threshold. Source: enforcement_post.yaml

#	TRIGGER	WHAT IS FLAGGED	LABEL	EXAMPLE
02	**LLM-written text (llm_slop_post)**	Machine-generated look	`RiskyHighVizReply + account label`	The same polished, generic line, "Great point, totally agree!", posted under many posts. Source: enforcement_post.yaml, enforcement_user.yaml
03	**Gibberish**	Nonsense text	`SpamHighRecall`	A reply like "asdf asdf 🔥🔥 " that carries no meaning. Source: enforcement_post.yaml
04	**Multi-step reply spam**	Patterns across a thread	`Classifier flag`	A chain of replies under one post, each adding a little more promotion. The newest reply in the chain must be flagged for the label to stick. Source: grox/flows/reply_spam/
05	**Coordinated spam**	Patterns across accounts	`Classifier flag`	A group of accounts that reply to each other's posts within minutes. Source: grox/flows/reply_spam/

E Exemptions

Accounts that skip some reply-spam labels or enforcement.

TABLE 05 — EXEMPTIONS COMMIT B220448

#	EXEMPTION	EFFECT	TYPE	SOURCE
01	**High PageRank (score ≥ 66)**	Skips RiskyHighVizReply	PageRank	task_filter.py, constants.py
02	**Grey badge**	Skips RiskyHighVizReply	Badge	task_filter.py

#	EXEMPTION	EFFECT	TYPE	SOURCE
03	**Very high follower count**	Skips some enforcement (threshold in YAML is a mock value)	Follower count	enforcement_user.yaml

F Author diversity multiplier

Applied to your 2nd, 3rd, and later posts in one feed. Decay 0.5, floor 0.25.

TABLE 06 — AUTHOR DIVERSITY MULTIPLIER (VM-RANKER) COMMIT B220448

#	POST POSITION	MULTIPLIER	CALCULATION	NOTE
01	**1st**	1.000	`k = 0: 0.75 × 1 + 0.25`	No penalty
02	**2nd**	0.625	`k = 1: 0.75 × 0.5 + 0.25`	Score is 62.5% of base
03	**3rd**	0.438	`k = 2: 0.75 × 0.25 + 0.25`	—
04	**4th**	0.344	`k = 3: 0.75 × 0.125 + 0.25`	Same as the worked example
05	**5th**	0.297	`k = 4: 0.75 × 0.0625 + 0.25`	—
06	**6th**	0.273	`k = 5: 0.75 × 0.03125 + 0.25`	—
07	**7th and after**	→ 0.250	`Floor: decay 0.5, floor 0.25`	Never reaches zero

G Final report: DO

Actions that raise the score. Numbers pair with the DON'T table.

TABLE 07 — FINAL REPORT: DO COMMIT B220448

#	ACTION	SCORE	HOW THE CODE USES IT	EXAMPLE
01	**Post original posts to mutual follows**	● Gain 5 → 20 (boost +15)	`bidirectional_boost_eligible()` needs `!is_reply && !is_retweet && is_mutual_follow_author`. `reply_weight_for` adds 15 (xai-value-model/weights.rs).	Your original post "Shipping v2 today" reaches a mutual follow with predicted reply P = 0.10. Contribution: 0.10 × 20 = **2.00** (instead of 0.10 × 5 = 0.50).
02	**Earn copy-link shares**	● Gain 20	`ShareViaCopyLinkWeight` = 20.0 (home-mixer/params/param.rs).	A viewer copies the link to your post. P = 0.05 → 0.05 × 20 = **+1.00**.
03	**Earn DM shares**	● Gain 5	`ShareViaDmWeight` = 5.0.	A viewer sends your thread to a friend by DM. P = 0.04 → **+0.20**.
04	**Earn quotes**	● Gain 5	`QuoteWeight` = 5.0.	A viewer quote-posts your take with their own comment. P = 0.02 → **+0.10**.
05	**Earn replies on your posts**	● Gain 5	`ReplyWeight` = 5.0.	"Which tool did you use?" invites replies. P = 0.03 → **+0.15**.
06	**Earn follows**	● Gain 4.0	`FollowAuthorWeight` = 4.0.	After a post that shows expertise and a clear bio, a viewer follows. P = 0.01 → **+0.04**.
07	**Earn shares (share sheet)**	● Gain 2.0	`ShareWeight` = 2.0.	A viewer uses the share sheet. P = 0.02 → **+0.04**.
08	**Earn reposts (in-network)**	● Gain 1.0	`RetweetWeight` = 1.0. Only in-network, because out-of-network reposts are filtered.	A follower reposts you. P = 0.05 → **+0.05**.
09	**Earn favorites**	● Gain 0.66	`FavoriteWeight` = 0.66.	A viewer likes the post. P = 0.10 → **+0.066**.

#	ACTION	SCORE	HOW THE CODE USES IT	EXAMPLE
10	**Earn clicks**	● Gain 0.3	`ClickWeight` = 0.3.	A viewer taps your link. P = 0.08 → **+0.024**.
11	**Earn link opens**	● Gain 0.2	`OpenLinkWeight` = 0.2.	A viewer opens the link card. P = 0.05 → **+0.01**.
12	**Earn video opens**	● Gain 0.02	`VideoOpenWeight` = 0.02.	A viewer taps to open your video. P = 0.10 → **+0.002**.
13	**Earn photo expands**	● Gain 0.05	`PhotoExpandWeight` = 0.05.	A viewer expands your image. P = 0.20 → **+0.01**.
14	**Earn dwell and click-dwell time**	● Gain 0.05 dwell · 0.4 click-dwell	`DwellWeight` = 0.05, `ContClickDwellTimeWeight` = 0.4, `ContDwellTimeWeight` = 0.006 (multiplied by predicted time).	A viewer clicks in and reads. Predicted click-dwell time 2 (model units) × 0.4 = **+0.80**.
15	**Earn quoted-post clicks**	● Gain 0.05	`QuotedClickWeight` = 0.05.	A viewer clicks the post you quoted. P = 0.10 → **+0.005**.
16	**Earn in-network posts the viewer hasn't explored**	● Gain 0.015	`PostUnexploredWeight` = 0.015, applied only when `in_network = true`, unless `PostUnexploredIncludeOutOfNetwork` is on (default false).	A follower who has never engaged with your topic sees your post. P = 0.30 → **+0.0045**.
17	**Post original posts for new audiences**	● No weight No fixed weight	Phoenix and SimClusters retrieve out-of-network original posts, and `popular_posts_source.rs` adds a shared list of popular recent posts (`EnablePopularPostsSource` is now true by default). `OONRetweetReplyFilter` removes out-of-network replies and reposts.	A stranger in France sees your original thread through retrieval. Your reply to someone else's post would be filtered for that stranger.

#	ACTION	SCORE	HOW THE CODE USES IT	EXAMPLE
18	**Reply to followers and mutuals**	● No weight No fixed weight	In-network replies pass `OONRetweetReplyFilter`. Out-of-network replies don't, unless `EnablePhoenixOonReplies` is on (default false).	Your reply to a follower's post is visible to your followers. Your reply to a stranger's post is dropped for strangers.
19	**Reply within 48 hours of the post**	● Filter	`MAX_POST_AGE` = 48 × 60 × 60 seconds, enforced by `AgeFilter`.	Reply at 3 hours: eligible. Reply to a post from 3 days ago: removed before scoring.
20	**Post regularly**	● No weight No fixed weight	Thunder keeps recent posts from followed accounts in memory (thunder/, README).	Four posts a week keeps your last 48 hours of posts ready for your followers' in-network feed.

Final report: DON'T

Actions that lower the score or trigger enforcement. Paired by number with the DO table.

TABLE 08 — FINAL REPORT: DON'T — COMMIT B220448

#	ACTION	OUTCOME	HOW THE CODE USES IT	EXAMPLE
01	**Automate or schedule replies**	● Label Post label `SpamHighRecall`, 30 days	enforcement_post.yaml: `fast_reply_spam_post` → `act_add_fast_reply_spam_post_label`, TTL 2,592,000,000 ms.	A script posts 200 replies in an hour. Flagged replies get the post label, which drops them as spam for 30 days.
02	**Post LLM-written or templated replies**	● Label Post label `RiskyHighVizReply`, 30 days. Account label `SpamHighRecall` if flagged `llm_slop_user`	enforcement_post.yaml: `llm_slop_post` → `RiskyHighVizReply`. enforcement_user.yaml: `llm_slop_user` → `SpamHighRecall`.	"Great point, totally agree!" under many posts is flagged `llm_slop_post`, so the reply is labeled `RiskyHighVizReply`.

#	ACTION	OUTCOME	HOW THE CODE USES IT	EXAMPLE
03	**Post gibberish or filler**	● Label Post label `SpamHighRecall`, 30 days	enforcement_post.yaml: `gibberish_post` → `SpamHighRecall`.	A reply of "asdf asdf 🔥🔥" is flagged `gibberish_post` and gets the post label.
04	**Repeat near-identical replies across threads**	● Classifier Classifier flag (threshold not published)	grox/flows/reply_spam/classifier_multi_step_reply_spam.py. It drops flags unless the newest reply is flagged.	The same "Totally agree!" posted under 30 threads is flagged by the multi-step classifier.
05	**Coordinate engagement with groups or swaps**	● Classifier Classifier flag (threshold not published)	grox/flows/reply_spam/classifier_coordinated_spam.py and plan_coordinated_spam.py.	Ten accounts that reply to each other within minutes match the coordinated-spam pattern.
06	**Burst many replies in minutes**	● Label Post label `SpamHighRecall`, 30 days	enforcement_post.yaml: `fast_reply_spam_post` (same rule as the automation row).	Fifteen replies in two minutes, by hand, can still trigger the fast-reply rule.
07	**Trigger a block**	● Loss −31.2	`BlockAuthorWeight` = −31.2 (home-mixer/params/param.rs).	A viewer blocks you. P = 0.005 → 0.005 × −31.2 = **−0.16**.
08	**Trigger a mute**	● Loss −58.8	`MuteAuthorWeight` = −58.8.	A viewer mutes you. P = 0.02 → **−1.18**.
09	**Trigger a report**	● Loss −234	`ReportWeight` = −234.0.	A viewer reports you. P = 0.01 → **−2.34**.
10	**Trigger "not interested"**	● Loss −52.272	`NotInterestedWeight` = −52.272.	A viewer taps not interested. P = 0.03 → **−1.57**.
11	**Lose dwell time (not dwelled)**	● Loss −0.02	`NotDwelledWeight` = −0.02.	A viewer scrolls past in under a second. P = 0.50 → **−0.01**.

#	ACTION	OUTCOME	HOW THE CODE USES IT	EXAMPLE
12	**Post spam or malicious links**	● Drop Dropped	visibility-filtering/rules/golden_corpus/oon_tweet_label.rs: `malicious_url/drop/undesirable` and `spam_high_recall/drop/undesirable`.	A link to a flagged domain removes the post from feeds.
13	**Run multiple or inauthentic accounts**	● Drop Temporary suspension (`PlatformManipulation`)	enforcement_user.yaml: `act_inauthentic_detection_v45_suspend` → `act_suspend_user`, `perm: false`.	Two accounts that only like and reply to each other trip inauthentic detection and get suspended.
14	**Keep the same pattern after a reach drop**	● Label Labels last 30 days regardless	All label rules use `ttl_msec: 2592000000` (30 days).	Changing your pattern now doesn't shorten a label that's already set. Wait out the 30 days.
15	**Expect replies to reach strangers**	● Filter Filtered for out-of-network viewers	`OONRetweetReplyFilter` with `EnablePhoenixOonReplies` = false.	Your reply to a follower's post is removed from a stranger's For You feed.
16	**Expect reposts to reach strangers**	● Filter Filtered for out-of-network viewers	`OONRetweetReplyFilter: is_retweet && in_network == false` is removed.	Your repost of someone you follow is removed from a stranger's feed.
17	**Count on video views or watch seconds**	● No weight Adds 0 through VQV and watch seconds	`VqvWeight`, `QuotedVqvWeight`, `VideoContinuationWeight`, `UserVideoContinuationWeight` and `PostVideoWeight` are all 0.0.	A 40-second video with 10,000 plays adds nothing through VQV.
18	**Count on profile clicks**	● No weight Adds 0 directly	`ProfileClickWeight` = 0.0 and `ProfileVisitSecsWeight` = 0.0. Only the follow action (4.0) counts.	500 profile visits add 0. One follow adds 4.0 × P(follow).

#	ACTION	OUTCOME	HOW THE CODE USES IT	EXAMPLE
19	**Post into a viewer's muted keywords**	● Filter Removed for that viewer	`ViewerMutedKeywordFilter`.	A viewer muted "crypto". Your crypto post is removed for that viewer.
20	**Post a 4th or later item from you in one feed**	● Loss Score × 0.344 at 4th, floor × 0.25	vm-ranker/params.rs: `EnableAuthorDiversity` = true, decay 0.5, floor 0.25.	Four of your posts rank for one viewer. The 4th is multiplied by 0.344 before the final sort.

Final report: worked examples

Illustrative probabilities, not measured values.

TABLE 09 — FINAL REPORT: WORKED EXAMPLES — COMMIT B220448

#	SCENARIO	CALCULATION	RESULT	WHAT IT MEANS
01	**Largest single gain: boosted reply on your original post**	`20 × P(reply) = 20 × 0.10`	+2.00	The biggest positive single term.
02	**Large gain: copy-link share**	`20 × P(share) = 20 × 0.05`	+1.00	Copy-link shares are the largest non-reply gain.
03	**Largest single loss: report**	`−234 × P(report) = −234 × 0.01`	−2.34	The biggest negative single term.
04	**Break-even report probability**	`2.00 ÷ 234`	P ≈ 0.85%	About one report-likely viewer in 117 cancels one boosted reply.
05	**Mute vs. favorite**	`−58.8 × 0.02 vs +0.66 × 0.30`	−1.18 vs +0.20	One mute costs about 6 favorites.
06	**Combined post**	`2.00 (boosted reply) + 0.20 (favorite) + 0.04 (follow) − 1.18 (mute)`	+1.06	Positive and negative terms add up before the final sort.
07	**Out-of-network penalty**	`2.00 × 0.75 (OonWeightFactor)`	1.50	Out-of-network posts keep 75% of their score.

#	SCENARIO	CALCULATION	RESULT	WHAT IT MEANS
08	**4th post from you in one feed**	`2.00 × 0.344`	`0.69`	Multiplier from `decay^k`, with k = 3.
09	**10th post from you in one feed**	`0.25 + 0.75 × 0.5^9`	`0.2515`	Approaches the 0.25 floor and never reaches zero.
10	**New-account out-of-network penalty**	`0.00001 (NewUserOonWeightFactor)`	`Off in repo`	Only applies when NewUserAgeThresholdSecs > 0. The repo default is 0.

Created by **@SidneyCur**. Built from the x-algorithm-source repository, commit b220448 (2026-11-03). Not affiliated with X.