Reconstructing the Hidden Objective Functions of Modern Personalized Feeds

InTelluric / Alnitak Group

Executive Summary

The public record does not expose the exact production reward functions used by modern feeds, but it does expose enough architecture, metrics, and organizational behavior to reconstruct the class of objective functions that plausibly drove them. Across official disclosures by YouTube, Meta Platforms, and TikTok, the recurring pattern is a multi-stage system: retrieve candidates from very large corpora, score them with many signals, optimize several predicted actions, and rerank under latency, diversity, and policy constraints. What the companies disclose publicly is architecture-level. What they do not disclose publicly is the decisive layer: the production reward mix, the relative weights on long-watch versus completion versus comments versus return frequency versus ad value, the telemetry windows used to refresh those estimates, and the experiment logs that selected one mix over another.

The strongest reconstruction is that the hidden objective function was not a single scalar like click-through rate. It was a composite continuity objective: maximize the probability that a user stays in the platform-controlled behavioral loop, returns soon, stays longer, generates more predictable feedback, and remains monetizable. The YouTube record is especially clear. Its 2010 paper describes personalized recommendation built from user activity, co-visitation graphs, multi-hop expansion, ranking, and A/B tests that tracked CTR, long CTR, session length, time until first long watch, and recommendation coverage. Its 2012 post then states directly that discovery features shifted away from driving views and toward increasing time spent watching, “not only on the next view, but also successive views thereafter,” with more watching also opening more revenue opportunities. Meta’s public descriptions of Facebook and Instagram similarly center objective functions, multiple ML predictions, “long-term value,” multi-objective retrieval, and tunable source weights. TikTok’s public description says videos are ranked by weighted signals, with watch completion treated as a stronger indicator than weaker contextual cues.

That continuity-maximization reconstruction matters because it matches the observed harm cluster better than the older “screen time” framing. The clearest public-health pattern is not all internet use versus no internet use. It is high-frequency, passive or semi-passive, socially evaluative, personalized, recommendation-driven exposure at adolescent developmental stages. The Centers for Disease Control and Prevention

reported that frequent social-media use among U.S. high-school students was associated with higher prevalence of bullying victimization, persistent sadness or hopelessness, seriously considering suicide, and making a suicide plan. The U.S. Department of Health and Human Services Surgeon General’s advisory states that children and adolescents using social media more than three hours per day face double the risk of poor mental-health outcomes including depression and anxiety symptoms. A 2025 JAMA Network Open cohort study using the ABCD cohort found that more time on social media during early adolescence “may contribute to increased depressive symptoms over time.” The 2026 World Happiness Report further distinguishes between online communication/news/learning, which correlate with higher life satisfaction, and social media/gaming/browsing for fun, which correlate with lower life evaluations, especially at very high use. Exposure is large enough to treat these systems as an environmental condition, not a niche consumer choice. Pew Research Center

reported in late 2024 that nearly half of U.S. teens say they are online almost constantly, that nine in ten use YouTube, and that TikTok and Instagram are each used by roughly six in ten. Gallup

reported in 2023 that U.S. teens spend an average of 4.8 hours per day on social media and that 51% spend at least four hours daily. Over the same broad period, CDC data show severe deterioration in youth mental-health indicators: the suicide rate for ages 10–24 rose 62% from 2007 to 2021, and 2023 YRBS estimates still showed 39.7% persistent sadness or hopelessness, 20.4% seriously considering suicide, and 9.5% attempting suicide among high-school students. Among young adults, the Substance Abuse and Mental Health Services Administration

reported in its 2024 NSDUH that 12.6% of adults aged 18–25 had serious thoughts of suicide in the prior year and 4.2% made a suicide plan.

The behavioral-science lineage strengthens, rather than weakens, the reconstruction. Public material from Stanford University’s Behavior

Design Lab states that behavior occurs when motivation, ability, and prompts converge, and that the lab's "persuasive technology" work explicitly addressed behavior change and ethics. The American Psychological Association

has summarized why adolescents are especially susceptible: starting around age 10, brains become more sensitive to social rewards like attention and approval from peers, and the APA's advisory notes that brain regions linked to attention, feedback, and reinforcement from peers become increasingly sensitive in early adolescence. In other words, the platforms did not have to "control minds" in any cinematic sense. A system that reliably identifies and sequences prompts tied to curiosity, peer validation, self-comparison, uncertainty, outrage, and relief can shape time allocation, sleep timing, affective state, and repeated re-entry while remaining fully compatible with users experiencing the behavior as self-directed.

The government/influence lineage is real, but the public evidence supports a narrower claim than direct state authorship of today's recommender stacks. Official materials from DARPA describe social-media-scale programs on misinformation detection, linguistic cues, sentiment/opinion detection, information-flow analysis, and narrative effects on cognition and behavior. Official material from IARPA

describes OSI as continuous, real-time monitoring of public data including social media, web search, news feeds, internet traffic, and Wikipedia edits to anticipate significant societal events. The U.S. Supreme Court's opinion in Murthy v. Missouri records regular contact between federal officials and platforms regarding COVID-19 and election-related misinformation, including the White House, the Surgeon General, CDC, the Federal Bureau of Investigation, and the Cybersecurity and Infrastructure Security Agency; the opinion also notes that the Surgeon General's advisory encouraged redesigning recommendation algorithms to avoid amplifying misinformation. What open sources do not establish is that government actors designed, funded, or controlled the core engagement-maximizing feed objectives. That remains open. The most useful investigative conclusion is therefore narrower and stronger than generic blame. The hidden objective function can be reconstructed as a continuity-maximizing, monetization-compatible, risk-managed control system acting on individualized behavioral forecasts. The through-lines worth investigating are the lineages that would leave observable traces: candidate-graph construction, session metrics, multi-objective scorecards, long-horizon return labels, sequence-learning feature stores, monetization-weighted auctions, "non-recommendable" policy classes, and internal safety exception handling. Those traces are concrete enough to target through FOIA, discovery, procurement records, and internal repository requests.

What the public record shows and what it withholds The public record shows a consistent architecture but an incomplete provenance chain. Official publications disclose retrieval funnels, candidate generation, ranking layers, example features, and example optimization targets. YouTube's 2016 paper explicitly says it describes the system only "at a high level." Meta's Facebook explanation says the public sees "many of the details," but that "under the hood" the ranking system is "incredibly complex," composed of multiple layers of ML models and rankings. TikTok's description enumerates factor classes and relative signal strength, but not coefficients, long-horizon retention labels, or safety-loss tradeoffs. The result is a stable asymmetry: architecture is public; the production reward surface is not.

That gap matters because observed social effects are generated by the hidden layer, not by the existence of candidate generation in the abstract. Knowing that a feed has two-stage retrieval and ranking does not reveal whether the production system privileges expected watch time over completion, completion over comments, comments over shares, shares over return probability, or ad yield over all of them. It also does not reveal what negative weights exist for regret, sleep disruption, compulsive re-entry, or social-comparison spirals. The absence of these specifics in official disclosures is itself informative: the public documents inspected here explain how the systems are structured, but not which combination of behavioral outcomes won internally. The convergence timeline below summarizes the major lineages that are publicly visible.

timeline title Lineage convergence of modern personalized feed systems 1997 : Stanford Persuasive Tech Lab begins formal ethics/persuasive-tech work 2005 : YouTube launches 2010 : YouTube paper describes co-visitation graph, seed expansion, ranking, A/B metrics 2011 : IARPA OSI introduced for real-time public-data monitoring 2012 : YouTube shifts discovery toward watch time and successive views 2012 : DARPA SMISC and Narrative Networks documented 2016 : YouTube publishes deep candidate-generation and deep-ranking architecture 2021 : Meta publicly frames News Feed ranking as objective-function optimization and long-term value 2021 : Murthy opinion records intensified federal-platform contacts on misinformation 2023 : Meta describes Instagram Explore four-stage retrieval/ranking funnel 2023 : TikTok publishes weighted factor description and teen time-limit defaults 2023 : YouTube adds teen repeated-recommendation guardrails 2024 : CDC publishes YRBS analysis linking frequent social-media use to bullying, sadness, suicide-risk indicators 2024 : European Commission requests recommender-system details tied to mental well-being and addictive behavior

2024-2026 : Meta publishes people-centric sequence-learning ads infrastructure and larger ranking models The sequence above is synthesized from official platform, government, and public-health records. Objective-function map The most defensible reconstruction is

a layered objective that selects for behavioral continuity, then constrains it. Layer Publicly visible Publicly visible inputs mechanism Hidden-but-Likely inferable target institutional class

watch history, likes, favorites, playlists, co-visitation graph, populate a menu generation follows, recent retrieval sources, of likely-interactions, embeddings, Two-engaging items content metadata, Tower retrieval, fast enough for real-time and pre-candidate funnels real-time serving relevance at scale, freshness, discovery, low latency generated sources thousands of user/ content/context weighted ranking, maximize signals; video info; multiple ML immediate and predictions, separate near-horizon ranking stage, reranking behavioral yield for each item session depth, repeat usage, stability Ranking Session shaping device/account settings; event streams with timestamps autoplay, infinite supply, refresh, notifications, ranking refresh cadence session continuation optimization, “successive views thereafter,” time-to-long-watch, return nudges reduce exit habit formation, probability and inventory increase next-action growth, revenue predictability opportunities bid, estimated ad auction using total maximize Monetization action rate, quality, value, people-centric monetizable coupling cross-device and sequence-learning attention per audience signals ads models unit of user time advertiser ROI, revenue, performance marketing efficiency Layer Publicly visible Publicly visible inputs mechanism policy classes, “not demotion, “non-Safety and interested” signals, recommendable” integrity negative feedback, classes, teen constraints harm-sensitive guardrails, reranking categories penalties Hidden-but-Likely inferable target institutional class

cap the most visible or legally salient harms without abandoning the core loop regulatory defense, litigation defense, brand protection This map is grounded in the YouTube 2010 and 2012 disclosures, Meta’s News Feed and Instagram Explore explanations, TikTok’s factor description, and Meta’s ads documentation. A concise template for the hidden production score is: Score(u, i, t) = Σ_k w_k · P(action_k | u, i, t) + α · E[watch_time | u, i, t] + β · P(next_view_or_next_open | u, i, t) + γ · P(return_within_horizon | u, state_t) + δ · MonetizationValue(u, i, t) - λ · PolicyRisk(u, i, t) - μ · Redundancy(i, slate_t) + ν · Freshness(i, t) For ads, one publicly disclosed sub-case is even simpler: AdTotalValue = Bid × EstimatedActionRate + QualityAdjustment The first template is a reconstruction from multiple official disclosures; the second follows Meta’s published description of ad-auction “total value.” Technical lineage of recommender objectives The clearest early public technical lineage is YouTube’s 2010 recommendation paper. It describes a personalized top-N recommender using a user’s personal activity as seeds, then traversing a co-visitation graph built from video pairs co-watched within sessions, “usually 24 hours.” It explicitly allows multi-hop expansion to broaden recommendations beyond direct neighbors, then ranks candidates by video quality, user specificity, and diversification. It also describes batch precomputation, several updates per day, and live A/B testing with metrics including CTR, long CTR, session length, time until first long watch, and coverage. That is already a behavioral system, not merely a content-similarity engine: it uses telemetry about what users watched, how far they watched, what they favored or liked, and how those actions change session behavior.

The 2012 YouTube watch-time post is the pivotal public provenance marker because it states the objective shift, not just the architecture. The company says discovery features were previously designed to drive views, but were changed to surface videos that viewers actually watch, and that suggestions now focus on increasing time spent watching “not only on the next view, but also successive views thereafter.” It further states that more watching “opens up more opportunities to generate revenue.” That is a public statement of long-horizon session objective plus monetization relevance. It does not publish the formula. It does reveal the target behavior class. The 2016 YouTube deep-learning paper formalizes the next step: separate deep candidate generation and deep ranking. The paper is explicit that it presents the system only “at a high level,” which is important in this context. Public disclosure moves from simple co-visitation graphs toward learned retrieval and ranking, but the production objective remains only partially visible. What the disclosure does reveal is the enduring split between retrieval and ranking that later appears across the industry. Meta’s public descriptions show the same architecture class. Facebook’s News Feed documentation describes a system that must score thousands of candidate posts using thousands of signals and then define an objective function for each user. It explicitly discusses multiple predictions, aggregation into a single value, and an overarching objective of creating “the most long-term value” for the person by showing content that is meaningful and relevant. Instagram Explore, in turn, publishes a four-stage funnel of retrieval, first-stage ranking, second-stage ranking, and final reranking. It explains that retrieval can combine real-time and pre-generated sources with tunable weights and that Two-Tower models are used for multi-objective retrieval. TikTok’s public documentation is more compact but directionally consistent. The platform says its “For You” feed ranks videos using a combination of user interactions, video information, and device/account settings, and that signal weights differ, with full watch on a longer video treated as stronger than weak contextual matches like shared country. That disclosure again shows the broad shape of the scoring function without revealing the reward weights, horizon, or safety-loss mix. Candidate objectives and predicted social effects

ranking objective Public exemplar What it selects for YouTube 2010 used CTR CTR maximization and long CTR as evaluation curiosity-gap metrics; Google Ads thumbnails, novelty, defines CTR as a major immediate clicks quality signal Predicted social effects if scaled population-wide impulsive clicking, shallower quality control, lower informational trust YouTube 2012 explicit shift Expected watch to watch time and time and session successive views; 2010 also continuation tracked session length and time until first long watch

content that reduces exit probability and chains into more viewing compulsive continuation, sleep displacement, passive overconsumption, reduced self-termination

ranking objective Public exemplar What it selects for Predicted social effects if scaled population-wide Facebook's public Multi-action objective-function framing social combines likes, comments,

shares, and "long-term value" Instagram Explore uses socially provocative, outrage, social identity-relevant, comparison, peer-status emotionally activating items sensitivity, norm acceleration Retrieval for multi-objective discovery multi-stage ranking, rapid deeper personalization, tunable source weights, personalization, finer subculture and Two-Tower retrieval exploration under segmentation, faster niche learned from engagement relevance constraints escalation events Meta ads auction uses bid × estimated action rate sustained Monetization-plus quality; Meta engagement states weighted value sequence-learning ads that remain saleable models use event streams to advertisers and timestamps surveillance expansion, ad-optimized attention shaping, stronger incentives for never-idle states Safety-constrained relevance YouTube teen guardrails, TikTok teen limits, EU DSA recommender scrutiny keep the core loop partial mitigation without while suppressing objective redesign; visible the most legally harm reduction at the salient classes margins only The technical lineage therefore supports a reconstruction in which the production objective evolved from local relevance toward long-horizon continuity plus monetization, with safety layered on later as a constraint, not as the primary reward. Behavioral, ad-tech, and government lineages Behavioral-science lineage The behavioral-science lineage is not speculative. Stanford's Behavior Design Lab states that behavior change occurs when motivation, ability, and prompts converge, and its ethics page documents formal persuasive-technology work from the late 1990s onward. That matters because modern feeds operationalize exactly those levers: prompts are continuous and personalized; ability is reduced by one-tap responses, autoplay, and infinite supply; motivation is tuned through social reward, novelty, threat, or relief. Adolescent susceptibility is also well documented. APA summaries explain that children's brains become more reward-sensitive around age 10, especially to peer attention and approval, and the APA health advisory notes increased sensitivity of brain regions associated with attention, feedback, and reinforcement from peers in early adolescence. The Surgeon General's advisory further points to poor sleep, online harassment, low self-esteem, poor body image, and higher depressive symptoms in connection with heavier social-media use. In practical terms, a recommender system need not target "depression" directly to worsen mental health. Selecting for repeated exposures that heighten social comparison, fear of exclusion, novelty seeking, and bedtime continuation is sufficient to shift risk at population scale. Ad-tech lineage The ad-tech lineage makes the hidden objective more legible because advertising systems publish their value logic more openly than feed systems do. Meta states in its fairness paper that ads compete in an auction where total value is calculated from the advertiser's bid, the estimated action rate, and ad quality. The same paper defines "ad delivery inputs" as data used as features in ML models that expand ad audiences or inform delivery. Meta's newer ads-infrastructure post shows that its ads recommendation system now uses sequence learning, people-centric infrastructure, event streams, sequence lengths, timestamps, and transformer-like architectures to rank thousands of ads in a few hundred milliseconds. That is a direct admission that the company models temporally ordered behavior sequences rather than only static profile attributes.

Google's ad and analytics documentation fills in the identity and cross-device side of the lineage. Google documents cross-device reporting and modeling using signed-in users who have Ads Personalization enabled, explains how device paths are inferred across mobile, tablet, and desktop, and notes that many cross-device conversions are modeled from signed-in user behavior. Google's RTB documentation shows bid requests exposing device-level geolocation fields and supports identifiers like GOOGLE_USER_ID ; the OpenRTB specification itself treats device, user, geo, data, and segment as first-class ad-request objects. The significance is straightforward: identity-linked, cross-device, temporally ordered behavioral traces became normal inputs to ad optimization, and that same data logic is structurally compatible with feed optimization. Government and influence lineage The open-source government lineage shows interest in influence-relevant capabilities, not direct authorship of feed reward functions. DARPA's SMISC program described a "new science of social networks," with tools for misinformation/deception detection, linguistic cues, information flow, sentiment/opinion analysis, and concept tracking. DARPA's Narrative Networks program described work on how narratives influence cognition and behavior in security contexts, including radicalization, social mobilization, conflict prevention, communication, and PTSD treatment. IARPA's OSI program described continuous real-time monitoring of public data including social media, web search, news feeds, internet traffic, and Wikipedia edits to detect significant societal events. Separately, the Murthy v. Missouri record shows documented federal-platform communications about information flows. The Court's opinion recounts that White House officials, the Surgeon General, CDC, FBI, and CISA communicated with platforms about COVID-19 and election-related misinformation; it also notes that the Surgeon General's advisory encouraged platforms to redesign recommendation algorithms to avoid amplifying misinformation. The same opinion records CDC meetings and reports to platforms, FBI and CISA meetings with platforms before the 2020 and 2022 elections, CISA "switchboarding," and the existence of Facebook "Covid Insights" reports sent to White House and Surgeon General officials. That is not proof of control over feed objectives. It is proof that agencies recognized the strategic significance

of algorithmic amplification and maintained operational contact with platforms about it.

The institutional map below separates what is documented from what remains inferential. flowchart LR subgraph Platforms YT[YouTube] FB[Facebook] IG[Instagram] TT[TikTok] end subgraph TechnicalSystems CG[Candidate generation] RK[Ranking and reranking] NT[Notifications and autoplay] AD[Ads and auctions] FS[Feature stores and telemetry] end subgraph PublicHealth CDC[CDC and YRBS] HHS[HHS and Surgeon General] SAM[SAMHSA and NSDUH] WHR[World Happiness Report] end subgraph GovernmentInfluence DARPA[DARPA] IARPA[IARPA] FBI[FBI] CISA[CISA] end subgraph EvidenceChannels LIT[Litigation and AG complaints] WHB[Whistleblower disclosures] RFI[EU recommender RFIs] BLOG[Official platform blogs and papers] end YT --> CG YT --> RK YT --> NT YT --> AD YT --> FS FB --> CG FB --> RK

FB --> NT FB --> AD

FB --> FS IG --> CG

IG --> RK IG --> NT

IG --> AD IG --> FS

TT --> CG TT --> RK

TT --> NT TT --> AD

TT --> FS DARPA -. documented interest in narrative/social influence .-> GovernmentInfluence IARPA -. documented public-data monitoring .-> GovernmentInfluence FBI -. documented platform contacts .-> Platforms CISA -. documented platform contacts .-> Platforms HHS -. advisory and platform contacts .-> Platforms CDC --> PublicHealth SAM --> PublicHealth WHR --> PublicHealth BLOG --> Platforms LIT --> Platforms WHB --> Platforms RFI --> Platforms The government nodes in this diagram indicate documented research interests or documented platform contacts, not proven control of production recommender objectives. Expected observable traces by lineage Lineage If it influenced production, expect to see Technical lineage metric names such as session_length , long_watch , completion_rate , time_to_first_long_watch , next_view_rate , return_24h , source_weight , rerank_penalty , coverage ; diagrams of candidate funnels; seed-to-candidate explanations; retrieval source dashboards; latency budgets Lineage If it influenced production, expect to see experiment names around prompts, notification cadence, autoplay defaults, social-Behavioral-count visibility, cooldown timers, “time well spent,” “wellbeing,” “fatigue,” “regret,” science lineage “healthy session,” “night mode,” “wind down,” “social comparison,” “body image,” or “peer feedback” estimated_action_rate , total_value , quality_score , sequence-Ad-tech lineage learning feature stores, event-stream schemas, identity-resolution docs, cross-device path analyses, auction tuning memos, remarketing or audience-expansion model cards Government/ solicitations, contractor deliverables, social-media-monitoring procurement records, influence lineage meeting invites, switchboarding logs, RFI responses, data-sharing agreements, “insights” reports, election or health misinformation escalation channels Harm-internal memos on compulsive use, youth wellbeing, rabbit holes, repeated knowledge recommendations, self-harm or body-image guardrails, experiment-readout decks, lineage model-card risk sections, postmortems, and red-team or integrity reviews Some of these traces are directly suggested by the public sources; others are inferential targets derived from the disclosed architecture and from ordinary recommender-system practice.

Harm-knowledge lineage and minimal reproductions The harm-knowledge lineage is strongest where public sources show a transition from generic denial or abstraction to mechanism-level mitigation. YouTube’s 2023 teen-wellbeing post says repeated recommendations in categories involving physical-feature comparison, idealized fitness or body weight, and social aggression can be problematic for teens even if a single video is not. That is close to an admission that repeated delivery itself changes the harm profile. TikTok’s 2023 time-limit default likewise acknowledges a risk severe enough to justify default friction, even though the passcode design preserves choice and continuity. Meta’s 2026 public stance is different in tone but similar in structure: it says recent lawsuits oversimplify a complex issue while defending its record and listing safety tools. None of these positions amounts to abandonment of engagement-based feed architecture. They look more like safety overlays on top of it.

External records reinforce the inference that companies knew more internally than they disclosed publicly, though the evidentiary status varies. Frances Haugen’s written Senate testimony states that the documents she provided “prove” Facebook repeatedly misled the public about what its own research revealed regarding the safety of children, among other issues. State complaints against TikTok publicly allege that internal documents described the For You system as “addictive” and “content-neutral,” and that time spent on the platform was treated as a key success measure. Those complaint allegations are not final adjudications, but they are high-value documentary leads because they point to metric names, internal language, and experiment priorities likely to exist in underlying repositories. The regulatory direction of travel also supports a harm-knowledge interpretation. The European Commission’s 2024 request for information asked YouTube, Snapchat, and TikTok for detailed parameters used by recommender systems and their contribution to addictive behavior, rabbit holes, mental well-being

risks, and protection of minors. Regulators are thereby treating the ranking layer itself as a material risk generator, not merely user-posted content. Hypothesized minimal reproductions The recipes below are not reconstructions of proprietary code. They are minimal algorithmic reproductions that fit the public architecture and would be sufficient to produce the observed harm cluster if deployed at scale. Minimal reproduction Required labels and training data Loss / reward Experiment Predicted harm mix regime cluster Session maximizer Social-evaluative maximizer impression logs; multitask loss on click starts; watch duration; click, long-watch, completion, completion; next-view transitions; next-view, and return; score emphasizes reopen within expected watch 1h/24h; recent-time plus session user state continuation likes, comments, weighted reshares, negative feedback, actions plus social-graph affinity and creator-viewer freshness A/B tests on autoplay, source weights, rerank freshness, and compulsive use, bedtime drift, passive notification overconsumption, timing; winner persistent low-level chosen by arousal, fast habit session length, formation long-watch, and return experiments on public counters, ragebait/ commentbait thresholds, and repetition frequency self-comparison, peer-status dependence, cyberbullying salience, identity threat, persistent sadness/ hopelessness shares, follows, combination of feed ranking by profile taps,

peer affinity, relation features all of the above plus ad impressions, ad Monetization-clicks, coupled continuity maximizer conversions, sequence history, cross-device path data, ad quality, bid value feed score experiments on coupled to ad ad load, ad attention value and re-spacing, fragmentation, entry probability; notification-to-sellable attention ads use total-value auction session effects, states, stronger sequence-incentives to with estimated learning feature preserve arousal and action rate and windows, and prevent exits quality churn tolerance These recipes are directly compatible with the public disclosures about co-visitation graphs, multi-stage ranking, multi-objective prediction, watch-time optimization, ad total value, sequence learning, and cross-device modeling. The “harm” mechanism does not require a label called depression . It only requires

reward mixes that concentrate exposure around emotionally sticky, socially evaluative, self-referential, or difficult-to-disengage stimuli. Documentary leads, falsification tests, and research plan Prioritized documentary leads Priority lead Why it is high

Likely custodians Date

range Search terms / request terms Recommender metric dictionaries and dashboard screenshots fastest route to ranking teams; the actual data science; reward surface; product analytics; reveals what growth; integrity; 2009– present teams watched youth-safety daily groups session_length , long_watch , lctr , next_view , return_24h , quality_penalty , total_value , healthy_session , time_well_spent autoplay, infinite scroll, push notifications, Experiment shows what registry exports interventions and A/B test readouts were tried and what metric won experimentation platforms; product managers; ML infra; trust and safety 2010– repeated present recommendations, social counts, teen defaults, “not interested,” rabbit holes Feature-store schemas and model cards for ranking models reveals telemetry windows, labels, ML platform teams; and features responsible AI; actually in production ranking infra 2015– present Safety-tradeoff harms were youth wellbeing, identifies what memos and harm-review decks measured integrity, policy, internally and legal, escalation 2016– present whether they teams changed ranking Government-platform meeting records and “insights” reports highest-value public-sector lead for recommender relevance and influence utility HHS, CDC, Surgeon General, White House, FBI, CISA 2020– 2023 feature_store , event stream , sequence length , timestamp encoding , long horizon , return label , candidate source weight body image, self-harm, suicidality, addictive use, compulsive use, repeated recommendations, teen wellbeing recommendation algorithms, amplification, misinformation policies, content demotion, non-recommendable, Covid Insights

Priority lead Why it is high

Likely custodians Date

range Search terms / request terms DARPA / IARPA maps public-solicitations, contractor deliverables, transition reports sector influence research and possible technical spillover shows whether Ads-ranking and feed and monetization monetization design docs objectives were Youth-account and age-estimation policy docs Litigation discovery and state-AG exhibit repositories Procurement records for coupled reveals whether how likely source of internal language unavailable elsewhere practical route to federal social-contractor media analytics identification tools DARPA, IARPA, contracting offices, performers SMISC, Narrative Networks, 2009– OSI, sentiment, narrative present analysis, opinion mining, social media monitoring ads ranking, auction, monetization, growth science 2014– present minors were youth product, explicitly policy, trust and modeled and safety, privacy 2021– present estimated action rate, total value, ad quality, sequence learning, people-centric infra, cross-device conversions age estimation, teen account, supervision, repetitive content, wind down, passcode, bedtime prompts plaintiff steering addictive, content neutral, committees, AG 2021– time spent, engagement-offices, court present based design, compulsion, dockets body image, teen harm agency procurement offices, FPDS, SAM.gov, IGs 2009– present social media monitoring, sentiment analysis, influence operations, narrative analysis, public indicators, viral detection The agencies and institutions above are grounded in the official records already cited; the requests

themselves are investigative recommendations based on those records. Falsification tests and counterfactuals The lineage hypothesis weakens materially if the following findings appear. If production dashboards and experiment registries show that the dominant objectives were short-horizon relevance or explicit stated-preference satisfaction, while session continuation, long watch, return frequency, and monetization were secondary or absent, then the continuity-maximization reconstruction is too strong. Public YouTube evidence pushes the other way, but this remains falsifiable. If internal harm reviews show that negative weights for compulsive use, sleep disruption, body-image salience, and repeated recommendation risk were introduced early, enforced hard, and measurably changed traffic allocation before the public-health deterioration intensified, then the case for persistent

harmful objective selection weakens. Current public evidence shows later guardrails, not early redesign.

If causal or quasi-experimental audits demonstrate that the harm cluster is unchanged when autoplay, notifications, social-count visibility, and return nudges are disabled while ranking remains constant, then the mechanism shifts away from the hidden reward mix and toward interface-level prompts alone. The present epidemiology and platform disclosures do not resolve that question cleanly. If government records show only content moderation contact and no access to ranking parameters, behavioral telemetry, model-evaluation dashboards, or recommender design discussions, then the government/influence lineage should be narrowed to downstream moderation pressure rather than upstream recommender influence. The public record currently supports that narrower view more strongly than a stronger one. Open questions and limitations Official platform disclosures are selective and architecture-level. They are highly useful for reconstructing system classes, but they do not expose production weights, objective tradeoffs, or rollout decisions.

Court complaints and whistleblower statements are major documentary leads, not final adjudications. Where this report relies on such materials, it treats them as allegations or disclosures to be tested against underlying exhibits and internal records. The public-health record is strong enough to support serious concern and intervention, but it does not uniquely identify the exact coefficients of the hidden reward function. It identifies the outcome cluster and the most plausible mechanism classes.

Actionable research plan The first FOIA tranche should target HHS, the Surgeon General's office, CDC, FBI, and CISA for platform-meeting calendars, email threads, slide decks, "insights" reports, recommendation-algorithm discussions, misinformation-policy RFIs, and any documents containing the terms recommender , recommendation algorithm , amplification , demotion , non-recommendable , rabbit hole , engagement-based , watch time , or return frequency . The date focus should be 2020 through 2023 because that is the period of documented high-intensity federal-platform contact in the Murthy record.

The second tranche should target DARPA and IARPA for solicitations, performer lists, technical reports, evaluation criteria, transition documents, and closeout materials related to SMISC, Narrative Networks, and OSI, plus procurement records for social-media analytics, narrative analysis, sentiment analysis, and public-indicator forecasting. The aim is not to prove state authorship of consumer recommenders. It is to map overlap in methods, vendors, feature vocabularies, and influence objectives. The first discovery tranche in litigation should request experiment registries, ranking dashboards, model-card repositories, youth-wellbeing review decks, feature-store schemas, and source-code comments containing metric names associated with session continuation and return probability. The shortest path to the hidden objective is not the algorithm source code. It is the metric and experiment vocabulary that tells engineers what "better" meant operationally. Public sources already point to candidate names: long watch,

session length, time until first long watch, long-term value, estimated action rate, total value, repeated recommendations, and non-recommendable classes. The archive priority should be: official platform engineering posts and blogs; court dockets and unsealed exhibits in social-media-addiction and AG cases; FOIA releases from HHS and other agencies; DARPA/IARPA program pages and archived performer materials; and EU DSA recommender-system proceedings. That archive order maximizes the odds of recovering the hidden objective function through metrics, experiments, and operational traces rather than through speculation alone.
https://research.google/pubs/deep-neural-networks-for-youtube-recommendations/ https://research.google/pubs/deep-neural-networks-for-youtube-recommendations/
https://www.supremecourt.gov/opinions/23pdf/23-411_3dq3.pdf
https://www.supremecourt.gov/opinions/23pdf/23-411_3dq3.pdf
https://bytes.usc.edu/cs572/s25-555-sear-ch/lectures/YouTube/docs/The_YouTube_video_recommendation_system.pdf
https://bytes.usc.edu/cs572/s25-555-sear-ch/lectures/YouTube/docs/The_YouTube_video_recommendation_sys
https://www.cdc.gov/mmwr/volumes/73/su/su7304a3.htm
https://www.cdc.gov/mmwr/volumes/73/su/su7304a3.htm

https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf
https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf

https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/

https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/

https://www.darpa.mil/research/programs/social-media-in-strategic-communication https://www.darpa.mil/research/programs/social-media-in-strategic-communication

https://behaviordesign.stanford.edu/resources/fogg-behavior-model https://behaviordesign.stanford.edu/resources/fogg-behavior-model

https://engineering.fb.com/2021/01/26/core-infra/news-feed-ranking/ https://engineering.fb.com/2021/01/26/core-infra/news-feed-ranking/

https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/ https://blog.youtube/news-and-events/youtube-now-why-we-focus-on-watch-time/

https://behaviordesign.stanford.edu/ethical-use-persuasive-technology https://behaviordesign.stanford.edu/ethical-use-persuasive-technology

https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

https://engineering.fb.com/2023/08/09/ml-applications/scaling-instagram-explore-recommendations-system/

https://engineering.fb.com/2023/08/09/ml-applications/scaling-instagram-explore-recommendations-system/

https://blog.youtube/inside-youtube/continued-support-for-teen-wellbeing-and-mental-health-on-youtube/

https://blog.youtube/inside-youtube/continued-support-for-teen-wellbeing-and-mental-health-on-youtube/ https://www.apa.org/news/apa/2022/social-media-children-teens

https://www.apa.org/news/apa/2022/social-media-children-teens https://support.google.com/analytics/answer/7668466?hl=en

https://support.google.com/analytics/answer/7668466?hl=en https://www.commerce.senate.gov/wp-content/uploads/media/doc/

Frances%20Haugen%20Written%20Testimony.pdf https://www.commerce.senate.gov/wp-content/uploads/media/doc/Frances%20Haugen%20Written%20Testim

https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-youtube-snapchat-and-tiktok-recommender-systems-under-digital

https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-youtube-snapchat-and-tiktok-recommender-systems-under-digital