AIF CODEBOOK

The AIF Codebook for Agentic AI Accounts

AI FORENSICS

Credits

Author: Natalia Stanusch

The contribution from AI Forensics is funded by core grants from Open Society Foundations, Luminate, Omidyar Network, and Limelight Foundation.

Email: info@aiforensics.org

This codebook was developed for research published in the report titled "Prompt, Upload, Repeat: How Agentic AI Accounts Flood TikTok With Harmful Content" also available at: https://doi.org/...

Table of Contents

The AIF Codebook for Agentic AI Accounts

This codebook provides a consistent coding scheme for the types of content shared by Agentic AI Accounts (AAAs). The main challenge of coding AAAs is that the coder has to consider the whole feed of each AAA, meaning their entire history of posted content on the platform. This is necessary to establish whether an account qualifies as an AAA in the first place, and then to assess what content strategy it has adopted.

The following coding scheme focuses on dominant trends, meaning it considers the quantity (number of posts displaying certain characteristics) and quality (most popular posts) in order to assign a specific content or format category to each account. Hence, it accounts for dominant trends rather than each fluctuation in content strategy. Such content fluctuations include, for example, a mono-topic AAA "trying out" a new content and form across three posts in its feed, only to return to its original niche specialization. This means that the categories below are indicative rather than accounting for each posted piece of content.

This codebook accounts for a spectrum of synthetic content: moving images, still images, and deepfakes. The exact definitions, as well as visual analysis strategies to detect such content, can be found in the Human Guide to Detecting Generative AI Imagery. What follows is a brief description of each label, alongside remarks on particularly attention-demanding content types for the coder. Examples of coding can be provided upon request. The following coding categories were derived from an initial exploratory analysis of the dataset. This typology is non-exhaustive.

Account labels

1. Type of account

The type of AAA is assigned based on its content-creation strategy, which can be traced and analyzed in its posting history accessible via the chronological profile feed view. There are three mutually exclusive (an account can be coded as one type only) types of AAAs: mono-topic, poly-topic, and hybrid AAAs.

1.1. Mono-topic AAAs - specialize in one thematic convention or type of content, which can be observed both as a specific subject matter and/or as formal qualities that apply to most or all of the account's uploaded posts.

1.2 Poly-topic AAAs - attempt various formal and subject matter conventions, often following memetic trends. No rigid or coherent content specialization can be identified.

1.3 Hybrid AAAs - use AI-generated, stock, as well as found (non-AI) imagery intertwined with AI-generated voiceover. The presence of AI imagery can be found in some posts but not others, or can be consistently used in each post while being intertwined with non-AI content. The AI audio is present across most uploaded posts.

2. Origin of account

The origin of an account is a binary coding category that identifies whether, at the time of analysis, the account in question showed signs of uploading AI content from its inception already, understood as the oldest posts available on the account's feed.

2.1 AAAs - At the time of analysis, the account had no available posts that indicated non-AI content posted at its inception.

2.2. Non-AAAs - Account used to post non-AI content, meaning non-AI generated moving or still images, and then "switched" to posting AI content.

3. Account status

Account status labels define the status of the account based on the probable reason behind its feed being no longer accessible.

3.1 Deleted - Account's feed is inaccessible and appears to be deleted.

3.2. Empty - Account's feed is inaccessible as it no longer contains any uploaded content; an indication of previous content can be visible in the high number of account likes displayed, despite seemingly no posts available on the profile.

3.3. Possible username migration - Account is no longer accessible via the original profile link and appears to be deleted; however, searching for a specific post reveals the same previous content (old feed) having been uploaded under a new username.

3.4. Private - Account's feed is no longer accessible and displays a message that the account viewing setting is now "private."

4. Account characteristics

Account characteristics describe particularities of the shared content *aside from* its subject matter and formal qualities. An account can show multiple characteristics (the labels are not mutually exclusive). Such characteristics include:

4.1. Ads/sponsored content - Presence of disclosed and undisclosed sponsored content and advertisements.

4.2. Niche specialization - Account occupies a unique thematic and/or formal niche compared to other accounts in the dataset; usually applicable to particular mono-topic AAAs.

4.3. Reactivated after Sora - Account did not post prior to the release of Sora 2, or posted very rarely, and then began posting with an increased regularity or quantity of content per day/week after the release of Sora 2. The use of Sora 2 is either visible through the watermark within the content frames or in the hashtags/descriptions of posts.

4.4. Switched to Sora - Previously active account switched its preferred generative AI tool to Sora 2; this change is reflected in the unprecedented presence of Sora 2 watermarks within the content frames or in the hashtags/descriptions of posts.

4.5. Topic transition detectable - Account has preserved in its profile feed history a shift from one particular thematic specialization to another, allowing one to trace and see the content rupture between the subject matter and/or formal qualities of posts. This topic transition usually applies to mono-topic AAAs.

4.6. Veo tag - Account posts only or mostly content that contains the Veo 3 tag.

Content labels

5. Trend-based content

Trend-based content includes both AI-specific trends as well as broader viral formats and memes that circulate on TikTok and other platforms. Such trends share the same format and/or subject matter. In this coding, we focus on three particular trends present across AAAs on TikTok which gained popularity between July and August 2025.

5.1 "Antisemitic" trend - This trend gained traction in early July 2025. Its focus lies in satirical hyperbole encompassing issues around Jewish faith and the state of Israel, presented (sporadically) within the context of the war in Gaza. The format shows photorealistic generative AI scenes where Orthodox Jews claim that Cyprus (or, as the trend develops, any other location) was promised to them 9,000 years ago and is therefore now part of Jewish-owned territory. Aside from satiric messages, a lot of content shared on TikTok quickly took on an antisemitic turn, focusing on depicting derogatory Jewish stereotypes.

5.3. "Metro" trend - This trend displays AI-generated scenes taking place in metro compartments, where visibly young and stereotypically attractive women stand or sit in a crowd. Depending on the severity of the emotions and situations displayed, the context ranges from women in physical proximity of a (male) crowd seemingly at ease, to scenes of sexual assault conducted by a man or men from the crowd on the woman.

5.4. "Mother-son" trend - This trend displays an AI-generated scene, where an attractive woman (as per Western stereotypes) is in physical proximity to a young male-looking child. Several hashtags and descriptions across many posts following this trend suggest that the woman is the "mother." The scene is most often situated indoors (next to a desk, or on a bed, sofa, or carpet), or occasionally outdoors. The woman is sometimes dressed in a sexualized manner. The physical contact between the woman and the child is mostly suggestive and sometimes sexually explicit (embraces, kisses on the lips, etc.).

6. Format

The following labels describe the dominant format of content present across the account's feed. The labels are based on the formal characteristics (e.g., background, foreground, objects, main characters, setting, stylistic convention) that define the *format* of the post, independent of its content. Format labels are not mutually exclusive.

6.1. Polls-like engagement - Posts that function as polls, asking for engagement from users. They mostly include writing (or subtitles) as their visual focal point, occupying the center of the frame. The writing includes a question followed by a call to action, such as "answer/let me know in the comments."

6.2. Social situations - Posts relying on a depiction of situations across public and domestic spaces, often including satirical or irrational turns to conventional interactions between colleagues, strangers, employees, family members, etc. This content format has been popularized by the Veo 3 release and often appears in conjunction with a Veo tag.

6.3. Synthetic (citizen) journalism - Posts that purposely use professional news media conventions to appear as plausible news stories or reporting based on factual events. These conventions include settings and characters based on stereotypical cable television news programs and include interviews and reporting on particular events. Some posts show fake or real news media logos (e.g., ABC or CNN) within the video frame or on journalists' gear (e.g., microphones).

6.4. Synthetic interviews - Posts based on the format of street interviews, where a journalist/interviewer with a microphone asks a

question(s) to an interviewee in an exterior setting, usually on a street or other urban space. This content format has been popularized by the Veo 3 release and often appears in conjunction with a Veo tag.

6.5. Synthetic point of view (POVs) footage - Posts employing the convention of POV vlogs, where the main character visibly holds the camera in one hand and usually narrates the scene.

6.6. Influencer - Posts displaying usually one main AI character that is visually very similar across most of the shared content. This character poses as an "influencer," and often includes "I" in sentences included both in post descriptions and in subtitles within the posts' frames.

7. Subject matter

Subject matter refers to the dominant content (Who/what is portrayed? In what context? What is the message conveyed? What is the aim of the scene shown?) of the posts across the account's feed. An account can display multiple subject matters; consequently, the labels are not mutually exclusive.

7.1. Animals - A significant number of posts display animals as the main or significant characters.

7.2. Animated - None of the posts include any photorealistic content, but appear as cartoons, animated series, and the like. This label can also be used when the account contains predominantly animated content.

7.3. Anti-immigrant - Content that encourages anti-immigrant narratives and sentiments. Examples might include AI-generated synthetic scenes showing immigrants "flooding" the entry routes to a given country (e.g., a raft filled with people making their way towards a shore), or AI-generated fake interviews and POV recordings where immigrants are shown reiterating, inflating, and spreading derogatory stereotypes about themselves (e.g., taking away jobs, living off social benefits, engaging in crime). This content can also appear across more abstract formats, such as POVs that are "set in the future," e.g., claiming to represent an "immigration-driven" dystopian vision of a Western country in 2050.

7.4. ASMR - Content focused on the Autonomous Sensory Meridian Response (ASMR), displaying visually or sonically "pleasing" images, sounds, and actions. ASMR mostly includes object and food related

actions such as cutting, breaking, eating, etc., accompanied by a strong visual and sound element.

7.5. Potentially harmful (political) content - Content that implicitly or explicitly spreads potentially harmful, ideology-driven messaging, often in a political context. This includes content spreading racist, misogynistic, or other harmful messages. It also includes AI content that directly engages and disrupts political discourses through the use of plausible deepfakes or by sharing generative AI content in support of/against a political candidate in the election period. It also includes content that displays, justifies, or downplays violent behaviors.

7.6. Authority satire - Content focused on displaying celebrities, politicians, and other recognizable people in positions of power engaging in ridiculous, impossible, or unlikely activities for satirical purposes. Content which appears as caricature is also an example of such satire. Most often, this content does not try to be plausible and, instead, is visibly exaggerated.

7.8. Fantasy - Content displaying scenes, characters, and objects inspired by the canon of fantasy (fairies, unicorns, magic, dragons, etc.). This also includes scenes and stories from various mythologies.

7.9. Female body - Content primarily focused on the physical appearance of the female body dependent on (and reproducing) clichés, stereotypes, and conventions around "attractive" physical attributes. These conventions blend the history of commercial photography and historical representation of women across pop culture, influencer culture, and dance-based TikTok trends. These AI women are always stereotypically attractive, with often sexualized attire or cleavage.

7.10. Health - Content focused on issues related to human health, mostly related to "healthy" diet and strategies to stay "healthy."

7.11. Illusive historicizing - Content attempting to visualize particular historical events or imagined situations, resulting in historical distortions. It shows photorealistic renderings of historical scenes, characters, or locations to illustrate actual historical facts. Such images and videos can also be organized in the "POV" format, often opening with the text "POV: ..." in the posts' description or within the frame. Such content is meant to visualize "what it would be like to be"

in a specific historical place or during a historical event, often emulating a first-person perspective.

7.12. Jesus/Bible - Content displaying the character bearing the stereotypical attributes and look of Jesus Christ and/or content based on illustrating a story based on the Bible.

7.13. (Natural) disasters/events - Content displaying dramatic or catastrophic events taking place in nature (e.g., volcano eruption) or near cityscapes and urban settings (e.g., fires, tsunami, plane crash). The content is photorealistic and depicts the scene directly, with no POVs or news media-like formal elements (journalists, interviews, etc.).

7.14. Sports - Content focused on imagery related to sports and the Olympics.

7.15. Storytelling clickbait - Content meant to "illustrate" various sensational stories, focusing mostly on contemporary (mostly fake) events. Examples include murder mysteries and conspiracy theories, with taglines often including "you'll never believe..." and "did you know that..." as opening sentences.

7.16. Toddlers - Content focused on showing toddlers and newborns, often engaging in physically unrealistic activities.

7.17. Young girls - Content focused on AI female-like characters with child-like features, appearing as either children or likely underage girls. This is a specific variation of the **female body** label, which includes the same content-specific characteristics in terms of conventional objectification and potential sexualization of the female body, *in addition to* the AI females in question appearing as either children or underage girls.

7.18. Other AI slop - Other types of AI content.