AI-Auditology

D3.1
Audit scenario co-creation methods

Project Title	Model-based Auditing of Social Media AI Algorithms and their Tendencies to Spread Harmful Content
Contract No.	09I03-03-V03-00020
Project start date	August 2024
Duration	23 months

Funded by
the European Union

RECOVERY
AND RESILIENCE
PLAN

Grant agreement no.:	09I03-03-V03-00020
Project acronym:	AI-Auditology
Project website:	https://kinit.sk/project/AI-Auditology
Project full title:	Model-based Auditing of Social Media AI Algorithms and their Tendencies to Spread Harmful Content
Project start date:	August 2024 (23 months)
Work package:	WP3 – Scenario creation
Version:	1.0
Delivery date	June 30, 2026

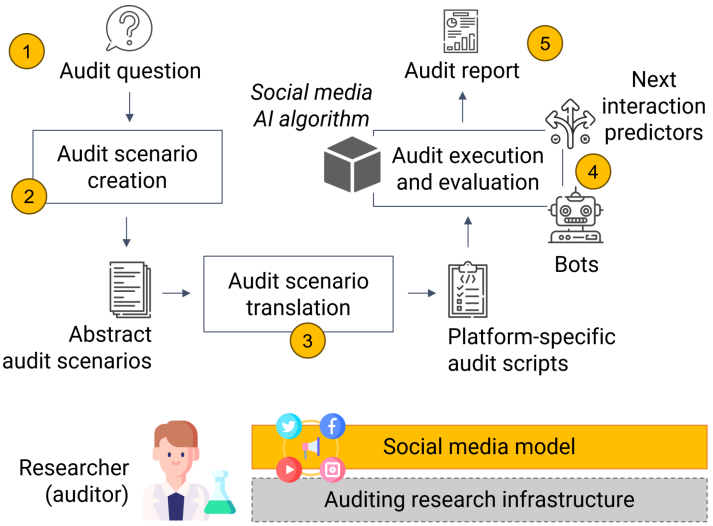
Project funded by VAIA - Research and innovation authority, under the call 09I03-03-V03, Grant agreement no. 09I03-03-V03-00020		
Dissemination Level		
PU	Public	x
NP	Non-public, only for members of the consortium (including the Agency Services)	

Table of Contents

1. Introduction

This document concludes research and development activities conducted within AI-Auditology work package 3 (WP3).

The process of model-based algorithmic auditing (the primary novelty of the AI-Auditology project) follows these steps (for more information, please, see our concept paper published at EWAF 2025):



1. A researcher (an auditor) formulates the audit question, for example: "how does the prevalence of persuasive antivax false claims in YouTube's search engine differ for various age groups".
2. The researcher interactively defines a set of abstract audit scenarios. In the process, she queries the social media model to determine user profiles (e.g., to achieve a representative age/gender/location distribution corresponding to the audited platforms) and to construct user actions (e.g., to select phrases a user should search for, such as "vaccination causes autism"; or to determine how many videos should be watched/skipped/up-voted to faithfully mimic the behavior of users on the audited platforms).
3. Next, such abstract audit scenarios are translated to platform-specific audit scripts (e.g., a user profile is converted to a real YouTube user account, topics are matched to the actual YouTube videos, user actions are mapped to their corresponding platform implementations).
4. During audit execution, the bots follow the prescribed audit scripts. User interactions can be either pre-determined by the script itself or automatically and dynamically predicted with utilization of next interaction predictors. The reactions of the platforms' AI algorithms (e.g., recommended videos) are observed and recorded.
5. The recorded reactions are automatically annotated for the presence of the audited phenomenon (e.g., the presence of disinformation claims in the recommended videos) and quantified in the resulting audit report.

The proposed audit scenario co-creation methods (and this deliverable) covers the second step of the outlined process. The subsequent steps of audit scenario translation (3), audit execution and evaluation (4) are covered in the subsequent deliverable D4.1.

This deliverable is structured as follows: In Section 2, it summarizes the *methods and experiments on audit scenario derivation and selection* (the outcome of T3.1 and T3.2). Section 3 consequently describes the *audit co-creation tool* providing the necessary user interface and support for the auditor (the outcome of T3.3).

Individual methods and solutions are demonstrated throughout this deliverable on two specific use cases (selected from the audits conducted during the AI-Auditology project implementation), which are briefly introduced below (for further details, please, refer to the corresponding published papers):

Use Case 1: Algorithmic Audit of Personalization in Polarising Topics on TikTok

Our first use case demonstrates the audit scenario co-creation on the study of personalization dynamics in polarising topics on TikTok (the study was published as a full paper accepted to UMAP 2026). In this study, synthetic user accounts were created and exposed to controlled sequences of content representing different positions (supporting and opposing) on selected polarising topics (such as politics, climate change, vaccines, and flatearth conspiracy theories). By systematically simulating interaction patterns in line with the user's assigned interest, such as likes, watch time, and skips, we were able to observe how the platform adapted its recommendations over time.

The collected data revealed: 1) a *preference-aligned drift* (showing a strong personalisation towards user interests), 2) a *polarisation-topic drift* (showing a strong neutralising effect for misinformation-themed topics, and a high preference and reinforcement of interest of US politic topic); and 3) a *polarisation-stance drift* (showing a preference of oppose stance towards US politics topic and a general reinforcement of users' stance by recommending items aligned with their stance towards polarising topics).

These findings highlight several risks. First, the rapid amplification of initial preferences creates opportunities for adversarial manipulation. Malicious actors could exploit this sensitivity by injecting targeted content early in a user's interaction history, effectively shaping their future information environment. Second, the reinforcement of homogeneous content streams may contribute to the formation of echo chambers, reducing exposure to diverse perspectives and increasing susceptibility to disinformation. Importantly, the auditing approach allowed us to quantify these effects systematically. Rather than relying on anecdotal evidence, we were able to provide reproducible measurements of personalization dynamics and their potential consequences.

Use Case 2: Algorithmic Audit of Advertisement Delivery to Minors on TikTok

The second use case presented in this deliverable focuses on the auditing of advertisement delivery to minors on TikTok (the study was published as a full paper accepted to FAccT 2026 and awarded the Best Paper award). This study specifically examined the platform's regulatory compliance with Article 28(2) of the European Digital Services Act (DSA). This regulation prohibits profiling-based advertising targeted at minors, reflecting concerns about adolescents' limited ability to recognize and resist commercial persuasion in digital environments.

In this study, we conducted an algorithmic audit using controlled sock-puppet agents simulating both minor and adult users with identical interest profiles (representing interest in 4 topical categories: beauty, fitness, gaming, and politics). These agents interacted with the platform in a systematic manner, allowing us to observe how TikTok's recommendation and

advertisement systems responded to user characteristics. The collected content was automatically annotated and categorized into four types: *formal* (explicitly disclosed paid) advertisements, *disclosed influencer/promotional content*, *undisclosed commercial content*, and *non-advertising content*. Additionally, we analyzed the extent to which recommended content aligned with the predefined interests of the accounts.

The results reveal a nuanced and concerning picture. On the one hand, TikTok appears to comply with the formal requirements of the DSA by limiting profiling-based targeting in clearly labeled advertisements delivered to minors. On the other hand, this compliance is largely confined to narrowly defined *formal advertisements* (covering only advertisements purchased through a platform's ad system for which a platform is directly paid). In contrast, both disclosed and undisclosed promotional content, particularly influencer marketing, exhibit strong personalization effects aligned with user interests. Notably, the degree of profiling observed in such content is substantially higher than that seen in formal advertisements delivered to adult users.

The most pronounced effects were identified in *undisclosed commercial content*, where creators or brands fail to properly label promotional material. In these cases, the platform neither enforces disclosure nor restricts personalized delivery to minors. As a result, minors are still exposed to highly targeted commercial messaging, effectively bypassing the intended protection of the regulation.

These findings highlight a critical regulatory gap that can be interpreted as a form of systemic vulnerability. While safeguards are implemented at the level of formally recognized advertisements, functionally equivalent commercial content can circumvent these protections due to definitional limitations. This creates an exploitable pathway for targeted influence operations aimed at minors, whether for commercial or potentially malicious purposes.

This use case demonstrates how algorithmic auditing can uncover discrepancies between formal compliance and actual system behavior. By systematically analyzing different categories of content and their personalization dynamics, the approach provides concrete evidence of how protections may be undermined in practice. More broadly, it underscores the importance of considering not only technical system design but also regulatory definitions and enforcement mechanisms when assessing the security and safety of online platforms.

2. Audit Scenarios Derivation and Selection

The central concept of algorithmic auditing is an *audit scenario.* In the context of sockpuppet algorithmic auditing, an audit scenario refers to a controlled, reproducible experimental setup designed to evaluate the behavior, outputs, or impacts of algorithmic systems.

An audit scenario formally specifies:

1. an *audit question – a research objective* determining what aspect of the algorithm is being audited (e.g., radicalization pathways or a personalisation drift),
2. the *platforms* and their *algorithms* to be audited – which algorithmic system is being probed (e.g., recommendation system behind For you page at TikTok platform),
3. the *user profiles* to be simulated – specifying the sockuppets' characteristics (e.g., age, gender, location, interests), including the initial content, accounts, keywords, or topics used to seed the experiment,
4. the *user actions* – an interaction protocol specifying actions to be performed under the given conditions by the sockpuppet accounts (e.g., searches, clicks, likes, follows, watch time when a content is/is not of user's interest),
5. the *observation points and metrics* – what data are collected and how outcomes are measured (e.g., recommendations received, content characteristics to be annotated, metrics to be calculated),
6. the *temporal structure* – specifies the timeline of the scenario – when accounts are created, when they are seeded with their interests, how long the audit is executed, as well as user sessions (e.g., how many sessions should happen during each day, how long such sessions should be).

Out of these elements comprising the audit scenarios, specifying particularly two require extensive, time- and expertise-demanding input from a researcher (an auditor): 1) *user profiles*, and 2) *user actions*. As a result, in the previous algorithmic audits, audit scenarios were typically created ad-hoc on intuition of researchers, remained incomplete (covering only a small subset of relevant user/content/interaction space) and inauthentic (heavily prescribed or too random).

In contrast with the previous audits, the AI-Auditology introduces a novel concept of the *abstract audit scenarios*, which employs a platform-, content- and time-agnostic abstraction at two levels (which also give the structure to the rest of this section):

1. Instead of specific user accounts, abstract audit scenarios define a (representative) **user population** which consequently determines the profile space from which specific **user profiles** are drawn (e.g., the audit will employ 100 bots, 56% of them are female, 44% are male). The user population definition ensures that sockpuppet accounts are not arbitrary, but grounded in a realistic demographic and behavioral space.

2. Instead of specific user actions with the predetermined platform- and time-specific content (e.g., a bot should like/follow these manually preselected videos or authors determined by a platform-specific URL), abstract scenarios define modular/reusable **audit phases** further specifying the platform-agnostic user actions (e.g., up-vote) and

more latent content representations (e.g., keywords to be used to identify currently available/popular content).

As such abstract scenarios are not directly executable, audit scenario translation is performed before each audit run and converts scenarios to the executable platform-specific scripts (see D4.1 for more details). This approach makes it possible to execute the “same audit” on multiple platforms, repetitively over longer time periods, and in multiple languages; as well as to make the audit outcomes directly comparable.

The AI-Auditology project aimed to explore how the algorithmic auditing process can be supported by *audit scenario co-creation methods*. Specifically, it explored the potential use of the social media model (as previously introduced in deliverable D3.2) — the key underlying element of the proposed model-based algorithmic auditing paradigm.

2.1 User Population and User Profiles

At first, AI-Auditology co-creation methods support the definition of the representative user population, which allows us to answer the posed audit question. Depending on the input audit question, a user population can be defined either: 1) manually by a researcher (an auditor) in order to validate a specific finding (e.g., whether there is a difference in personalisation between genders); or 2) by the social media model (please refer to D2.1 and D2.2 for more details) to have a user population more representative of the specific platform.

In both cases, the user population is defined by the following user characteristics: gender of the user (male/female), age range, location and interests. Individual characteristics can be signaled to the platform either during the creation of the user account (gender, age), by technological means (location – determined by proxy servers situation in the desired country or language of the browser used during the audit execution) or by behaviour (interests – signalled by simulating the user history at the beginning of the audit execution).

The interests of the individual users are the most challenging for defining the user population, as these are integral to answering the audit question – forming the first grouping of the users based on the interests. For example, if we are interested in analysing the spread of misinformation on the platform, the interests of the users will be in misinformation topics. In specific cases, the interests are further expanded with a stance towards the interest or other details. For example, if we work with misinformation, some users might believe the individual misinformative topics (have a supporting stance) while others might disagree with them (opposing stance). When defining the user interests, we always follow the audit question, focusing on currently popular topics in the specific country – for example, climate change misinformation. This also holds for more neutral topics, where opting for more niche topics could bias the observed findings.

Depending on the character of the audit question, the AI-Auditology supports two methods how to distribute population across the user characteristics:

1. *Uniformly* across genders, age groups and locations, while specifically focusing on having more than 1 user for each combination of interest-gender-age-location (this is to guarantee the statistical significance of the observed findings, as we cannot draw conclusions from a single user). This approach is suitable especially when an overall

population size is small and the audit question aims to systematically/uniformly cover all relevant user groups.
2. *Stochastically* generated from the social media model, achieving a more representative population that reflects the real-world age/gender/location distribution. This approach is suitable especially when an overall population is large enough and the audit question aims to simulate more natural conditions. Similarly, the representative interests of users can be determined by the social media model, as it contains the real-world topic distributions.

When the same audit is executed on multiple platforms, the researcher (auditor) can decide whether the same user population will be used on all of them, or each platform will have a different distribution of users across user characteristics (what may be necessary especially when platforms significantly differ in their real-word characteristics' distribution, e.g., there are significant differences in age distribution).

From the resulting user population, user profiles (with assigned specific name, age, gender, location, interest) are drawn. To this end, the following actions are taken:

1. The user is assigned a name, surname and an email address that will be used to confirm the creation of the user.
2. The user is randomly assigned a specific age (or a corresponding birthdate) from his/her age group defined in the audit scenario.
3. The user is assigned with a specific gender defined in the audit scenario.
4. The user is assigned a location through proxy servers. From the list of available proxy IP addresses, one is assigned to the user until the end of the study to guarantee consistency and avoid any (shadow)banning.
5. The specific user profile is created on the specific platform. For this purpose, the registration on the platform is used, using the email address for automatic confirmation of the creation. For the registration process, the user is already created with the assigned proxy IP address.

The proposed co-creation approach also allows two critical features. At first to *neutralize the effect* of other not-to-be-investigated confounding factors that may influence the behaviour of the AI algorithms (e.g., users' demographic). It is possible either to make such factors neutral (e.g., using gender-neutral names, omitting information about gender, selecting the average/means age of users in the cluster), or generate enough users (following the real-word distribution of such factors) that will sufficiently eliminate any undesired side effects.

Second is the *audit scenario selection*. Especially, when a user population is stochastically generated from the social media model, the total size of the population can be quite high. In practice, the final number of simulated users is also a subject of the technical feasibility, such as the availability of proxy servers or the maximum number of concurrent users that can be managed at the same time. The proposed approach allows the researcher (auditor) to adjust the population distribution over user characteristics to decrease the overall population size (while preserving the necessary coverage of user characteristics) to make the audit technically feasible.

Last but not least, as already explained, the user population is independent of the platform(s) that is being audited and as such, it does not provide information about specific user accounts (e.g., user's name, surname, username, password, birthdate/age). Specific account information is determined as a part of the translation of the user population to specific platforms, which is further described in D4.1.

Below, we provide use cases of user populations from the individual auditing studies we have conducted as part of the AI-Auditology project. In both cases, the selection of user characteristics (e.g., age groups) and a distribution of user population across these characteristics were informed by the social media model.

Use Case: Algorithmic Audit of Personalisation Drift in Polarising Topics on TikTok

There were 68 sock-puppet accounts created for this audit. All accounts were set to the 18–24 age range to represent young adults and a 50:50 split between male and female profiles was used to control for gender-based differences. Accounts were operated using proxy servers located in the USA to standardize geographic context.

This population was structured into three distinct groups, each designed to test different aspects of the recommendation algorithm: *Neutral + Polarising Group* (32 accounts), *Polarising-Only Group* (32 accounts) and *Mixed Polarity Group* (4 accounts).

The purpose of the *Neutral + Polarising Group* was to test how TikTok's algorithm balances neutral (e.g., cooking) and polarising topics (e.g., flat earth, vaccines, climate change, US politics). Each account was seeded with both a neutral topic (cooking) and a polarising topic (e.g., flat earth with a "support" or "oppose" stance) with 4 accounts per combination of polarising topic and stance (e.g., 4 accounts for flat earth/support, 4 for flat earth/oppose, etc.).

Polarising-Only Group was designed to test the algorithm's behavior when users are only exposed to polarising content during the seeding phase. Accounts were seeded exclusively with polarising topics (e.g., US politics/support or oppose) therefore but interacted with neutral content (cooking) during the interaction phase as well. 4 accounts per combination of polarising topic and stance were in place in this case.

Lastly, the *Mixed Polarity Group* was created. Here we tested how the algorithm handles users with mixed or neutral stances on polarising topics. Accounts were seeded and interacted with both stances (e.g., support and oppose for US politics) simultaneously. All 4 created accounts were applied to the US politics topic only.

Use Case: Algorithmic Audit of Advertising and Minor Profiling on TikTok

The audit used 4 pairs of accounts (4 minors + 4 adults), totaling 8 sock-puppet accounts. Minors were explicitly defined as users below 18 (aligned with EU guidelines), while adults were 18 or older. All accounts were set to the European Union, specifically with a location determined by proxy servers situated in Germany. The pairs included both male and female profiles to control for gender-based differences. Accounts were configured to express

identical interests (e.g., fitness, beauty, gaming and politics) to isolate the impact of age on ad recommendations.

2.2 Audit Phases

Second, the abstract audit scenarios consist of definitions of audit phases. Each study is composed of multiple phases that can be loosely grouped into:

1. *login phase*, where the user account is created (manually by a researcher) and the first login is automatically performed;
2. *seeding phase*, where all the steps required for setting up the agent are executed (e.g., users are seeded with their interests);
3. *interaction (main) phase*, where the is main part of the audit is executed and where the platform responses (e.g., recommendations, search results, advertisements) are collected and prepared to be further analysed in order to answer the audit question;
4. *cleanup phase,* where the account is removed, the actions executed by the bot are reversed (e.g., likes/shares/bookmarks are removed).

In the *seeding phase*, the specific user accounts on different social media are created. Using the user population defined for the audit, the individual users are translated into the real users, e.g., by creating specific user accounts (with specific name, gender, email and login credentials). Also all the required setup steps are conducted in this step (e.g., if the platform requires further setup, it is done as part of this step). After the user accounts are created, the second step is to set up the accounts for the audit itself. This includes creating a browsing history for the users or setting up their interests. For example, if we are interested in whether a specific platform is profiling minors for advertisements in beauty products, this step will seed the user's interest by watching videos from the beauty domain. Overall, after the seeding phase is completed, the user is prepared for the audit execution.

The *interaction phase* represents the main phase of every audit. For every study, the phase has different characteristics based on the audit question we want to answer. This includes characteristics, such as for how many days the audit will be running, how many sessions will be in a single day or how long each session will take. During this phase, the created user interacts with recommended videos according to the audit scenario (e.g., if the user encounters a video that belongs to his/hers interest, the user watches it, otherwise skips it), in order to further build the user interests and profile. In addition, the design of the interaction phase specifies how all the encountered posts and videos are collected, along with all the metadata, in order to allow for analysis. After the interaction phase is finished, the collected data are prepared to be further evaluated for the purpose of producing the final audit report.

The *cleanup phase* is performed after the whole audit study is over and user accounts are not needed any more. They are removed from the platform altogether with all user interactions that can be reversed. This process is performed in line with the ethical and legal guardrails prepared specifically in the AI-Auditology project (please, refer to the Risk list with identified mitigation techniques).

Both the seed and interaction phase can be designed to use the so-called user interaction predictor which is one of the novel contributions of the AI-Auditology project. As opposed to

the existing previous audit studies that implement heuristics and simple, static rules, we design the user interaction predictor that provides dynamic decisions regarding the interactions, while still providing the interaction in real time for the users. The behaviour of the user interaction predictor is simple – it takes as input the encountered video or post along with all of its metadata and outputs the interaction that the user should take (like, share, watch, etc.). For more information, how user interaction predictors are implemented for each platform, please, refer to D4.1.

Proposals of the seeding and interaction phase require a large scope of decisions to be made by the researcher (auditor), such as to determine which phrases should be used to signal the users' interests, or how a user should interact when presented to a specific searched/recommended content (i.e., how a user interaction predictor should work). AI-Auditology provides necessary co-creations methods to support these decisions. To this end, the auditor can query the social media model, e.g., to determine the average number of sessions per day a typical user perform on social media, the length of such sessions, the likelihood that a content will be liked/watched/bookmarked (determined by the ratio of implicit/explicit feedback used by real-world users).

Similarly as for a user population, when the same audit is executed on multiple platforms, the researcher (auditor) can decide whether the same audit phases will be performed on all platforms or there will be some specifics for individual platforms (e.g., platform-specific length of sessions). In general, to achieve unbiased comparability between platforms, the aim is to have the audit phases as similar as possible.

The selected specific use cases illustrate the co-creation process of the audit phases done in the AI-Auditology project.

Use Case: Algorithmic Audit of Personalisation Drift in Polarising Topics on TikTok

In this study, two primary phases (the seed phase and the interaction phase) were designed to simulate realistic user behavior and measure how TikTok's recommendation algorithm personalizes and drifts content over time.

Seed phase was an iterative and adaptive process with the aim to establish and reinforce the initial interest profiles of sock-puppet accounts, while adapting to TikTok's reactions to previous audit methodologies (e.g., bot detection or algorithmic resistance). Accounts were designed to be seeded multiple times to ensure the platform's algorithm consistently recognized their assigned topics and stances. For each iteration:

- Accounts searched for videos related to their assigned topic and stance (e.g., flat earth/support, US politics/oppose, or cooking for neutral accounts) using pre-defined search queries.
- The first 51 videos from each search query were evaluated using an LLM-driven User Interaction Predictor (GPT-4.1) to determine relevance.
- Relevant videos were watched in full, liked, and bookmarked to signal strong interest in TikTok's algorithm.

Adaptive part of the seeding process was different for every of 3 major users groups:

- Neutral + Polarising Group: Initially seeded with 25 neutral videos (cooking) followed by 25 polarising videos. After 3 days of interaction, an additional 25 polarising videos were seeded to reinforce the profile.
- Polarising-Only Group: Seeded with 25 polarising videos to maximize polarity.
- Mixed Polarity Group: Seeded with 25 videos for each stance (e.g., support and oppose for US politics) to simulate a neutral or conflicted user.

The seeding phase was designed to take 1 day per iteration, with a 1-day wait period after each iteration to allow TikTok's algorithm to stabilize as the iterative seeding accounted for TikTok's evolving responses to sock-puppet behavior, ensuring that the algorithm consistently recognized and adapted to the simulated user profiles.

The interaction phase was proposed to measure personalisation drift by tracking how the proportion of recommended videos (by topic and stance) changed over time on TikTok's For You Page (FYP). Accounts log in daily (using saved cookies) and scroll through the FYP for ~1 hour per day, simulating realistic user sessions (according to the social media model the average TikTok user spends about 59 minutes at the platform per day). Each recommended video is evaluated by the user interaction predictor. If the video matches the account's topic and stance (or neutral topic for the first two groups), it was watched in full, liked, and bookmarked. If irrelevant, the video was skipped after 1–2 seconds to avoid detection as a bot. All video metadata (URL, title, description) and predictor classifications were recorded for analysis.

Duration of the Interaction Phase was 15 days for the Neutral + Polarising and Polarising-Only groups and 9 days for the Mixed Polarity group. Over 80,000 unique videos were encountered and analyzed during this time.

The iterative seeding phase was critical to counteract TikTok's bot-detection mechanisms and ensure the algorithm treated the sock-puppet accounts as legitimate users. The additional seeding after 3 days for the Neutral + Polarising group reinforced the user profiles, addressing potential algorithm decay or shifting recommendations. The study's design explicitly accounted for lessons learned from previous audits, where platforms like TikTok had adapted to detect and mitigate artificial behavior.

Use Case: Algorithmic Audit of Advertising and Minor Profiling on TikTok

The whole process was split into a few phases. Before the audit itself, preliminary analysis was necessary to gain initial insights into TikTok's commercial content ecosystem and disclosure practices to inform the quantitative audit design. Videos from influencers popular among minors were examined to understand how commercial content was labeled (or not).

In the seed phase initial interest profiles for the sock-puppet accounts are established to ensure they received relevant content for the audit. This involved liking, watching, and engaging with videos related to the chosen interests to "train" TikTok's algorithm. The goal was to create a consistent baseline for each account pair (minor + adult) so that their "For You Page" (FYP) would reflect the same interests. For each interest profile, we defined a set of domain-specific search queries (e.g., makeup, skincare, or cosmetics for the beauty topic). The agents executed these queries and selected videos from the search results to

interact with based on a semantic relevance check. To decide how a user should interact with each of the returned videos, we utilized a user interaction predictor.

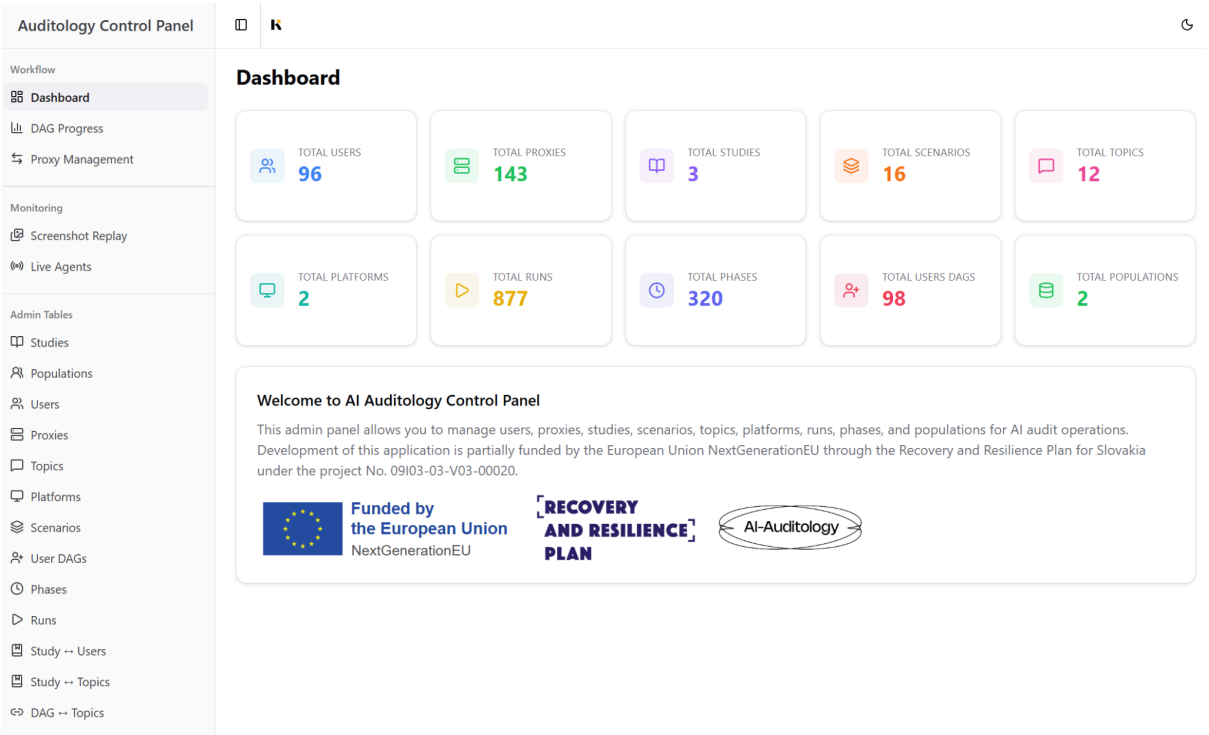
Following the seeding phase, we conducted the main interaction phase over a 10-day longitudinal period. Each bot interacted with the TikTok for 1 hour in one session each day, while each adult-minor pair was running in parallel. The duration of 1 hour corresponds to an average time spent by TikTok users on the platform per day. By running bots contemporaneously, we secured that there are no time-of-day-specific effects that may cause potential undesired biases. Unlike passive observation studies, our agents maintained their interest profiles through conditional interaction, simulating a user who selectively engages with relevant content while ignoring irrelevant recommendations. For every video presented in the “For You” feed, the system performed a real-time relevance assessment using the same user interaction predictor based on the video’s textual metadata. If the video matched the agent’s assigned interest, the agent executed the engagement routine (watch, like, and bookmark). If the content was unrelated, the agent immediately skipped the video. Crucially, we captured comprehensive metadata – including descriptions, hashtags, and visual frames – for all videos appearing in the feed, regardless of whether they were watched or skipped. This ensured we captured the platform’s unfiltered recommendation stream and ad delivery logic rather than just the user’s interaction history.

3. Audit Scenario Co-creation Tool

To streamline the auditing process, the AI-Auditology project designed and developed the complex software infrastructure. Within this infrastructure, the *Control Panel* represents the end-user-facing interface that ties all platform's features together and serves as a main frontend for any researcher and operator of an audit study. More specifically, it produces and maintains the audit configurations (audit scenarios), hands them over to the execution layer, and brings the resulting observations back into one place for review.

- The first part of the Control Panel supports audit scenario co-creation. It allows users to easily create, update and manage audit studies and persist them into the database.
- The second part of the Control Panel provides monitoring capabilities, exposing logs and screenshots for debugging and monitoring purposes. Live agent pods can be inspected via embedded VNC. Screenshot and log archives uploaded to AWS S3 can be browsed, extracted, and replayed frame-by-frame from inside the same UI.

The Control Panel is designed to be platform agnostic and easily extensible, currently supporting both TikTok and YouTube Shorts platforms.

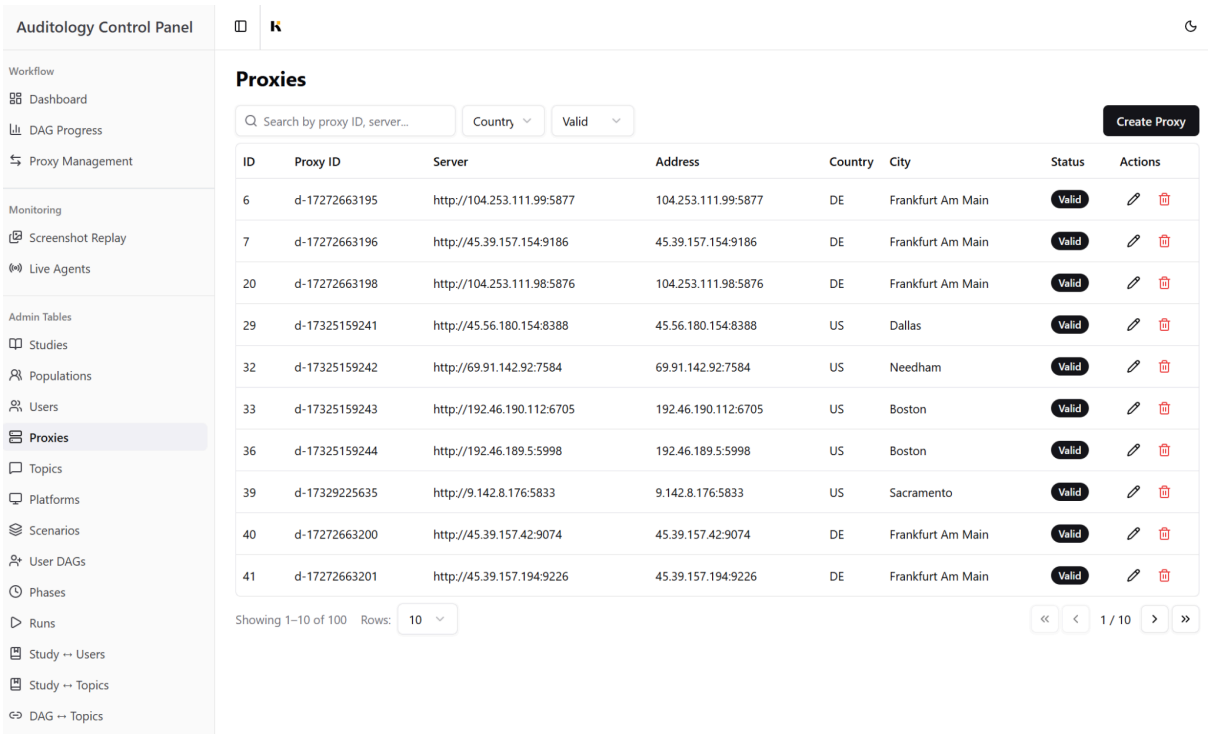


The rest of this section describes the first part of the Control Panel supporting audit scenario co-creation. The second part is further described in the subsequent deliverable D4.1.

3.1 Management of Platforms, Topics, and Proxies

At first, the Control Panel provides a researcher necessary features to manage enumerations that are later used to setup the audit study, namely:

- Social media *platforms* supported by the audit execution engine (currently TikTok and YouTube shorts).
- *Topics* of content to be audited and of users interests to be simulated. Topics are the content lens through which the population interacts with the platform. A topic is more than a label. It stores the queries the audit pod uses to actively search for related content, especially during seeding. It also stores a structured prompt in JSON form. The prompt is passed to the interaction predictor, which then decides whether observed content matches the topic and how the sock puppet should react. At runtime, topics are not taken from the study in the abstract. They are taken from the specific User DAG. This matters because different sock puppets in the same study can carry different topic subsets.
- *Proxies* used to simulate users' location.



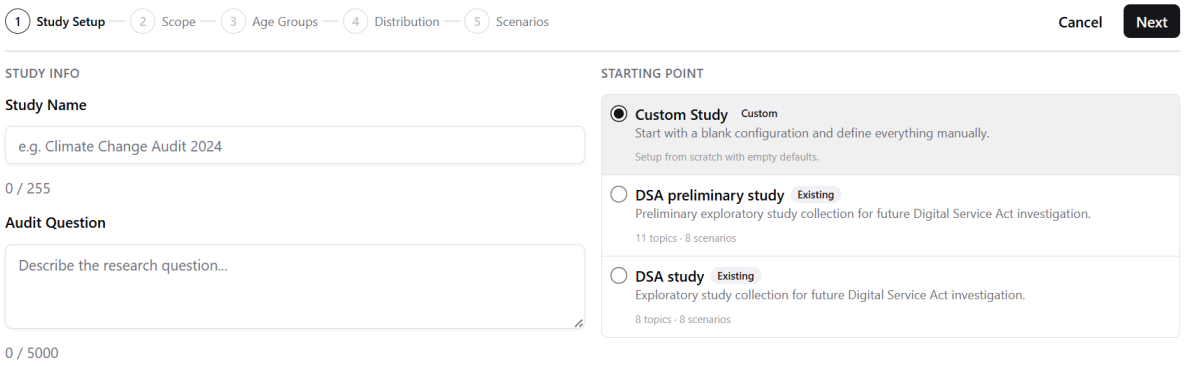
3.2 Study Designer

Designing the user population is the first thing a researcher does after deciding what to audit. In the Control Panel, this happens through the *Study Designer*. The designer is a five-step interactive wizard reached from the Studies page.

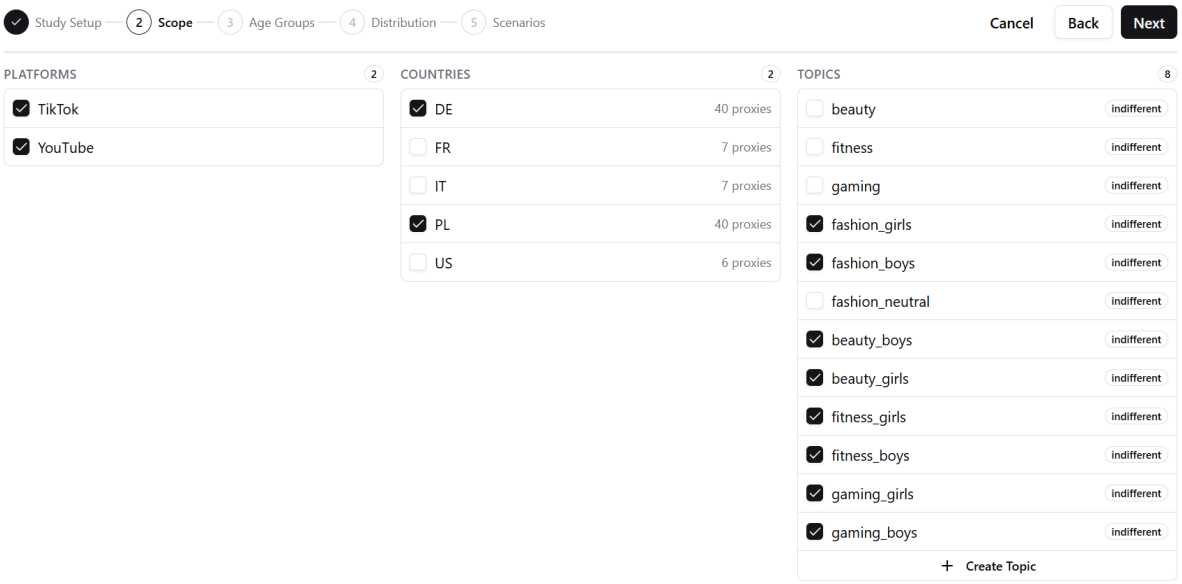
The wizard guides a researcher through five steps as follows:

1. The first step is the *Study Setup*. The researcher names the study, writes the audit question, and chooses a starting point. The starting point is either a blank Custom Study or an existing study reused as a template. Selecting an existing study loads its linked topics, the platforms used by its scenarios, and the scenarios themselves. It also loads the countries, the total population, and the age groups from that study's

first population. This effectively turns the designer into a study-based template system.

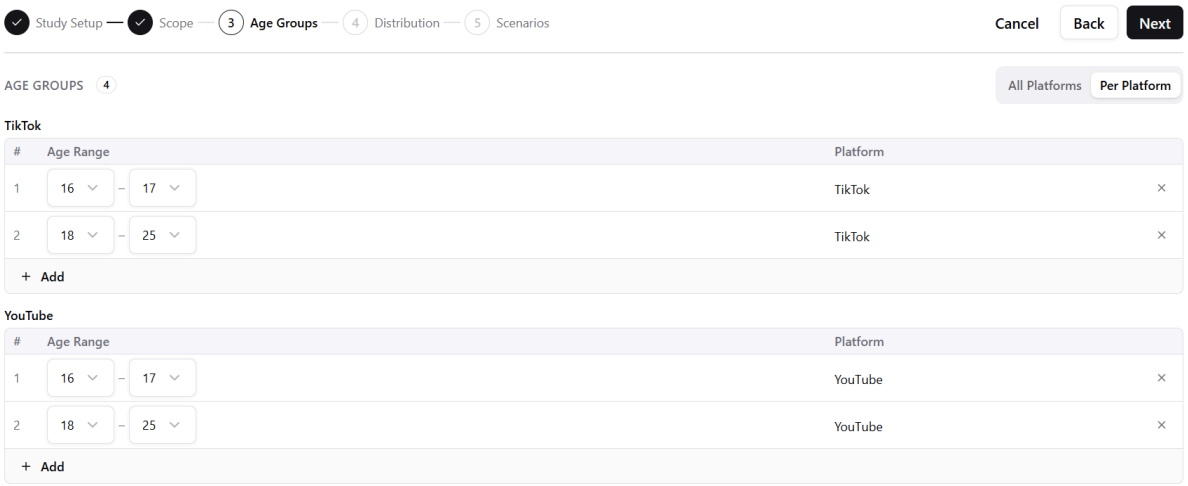


2. The second step is *Scope definition*. A researcher chooses three things. First, the platforms the study will target (for example, TikTok or YouTube). Second, the countries the sock puppets will operate from. Third, the topics the study cares about. Only countries that already have proxies registered in the Control Panel are selectable, and each entry shows how many proxies are available. This is what enforces geographical realism. A sock puppet cannot be allocated to a country in which the platform has no IP-level presence. Topics are picked from the global topic catalogue, or created inline. Each topic carries a label, a stance (support, oppose, or indifferent), a JSON prompt used later by the interaction predictor, and a list of queries used during seeding.

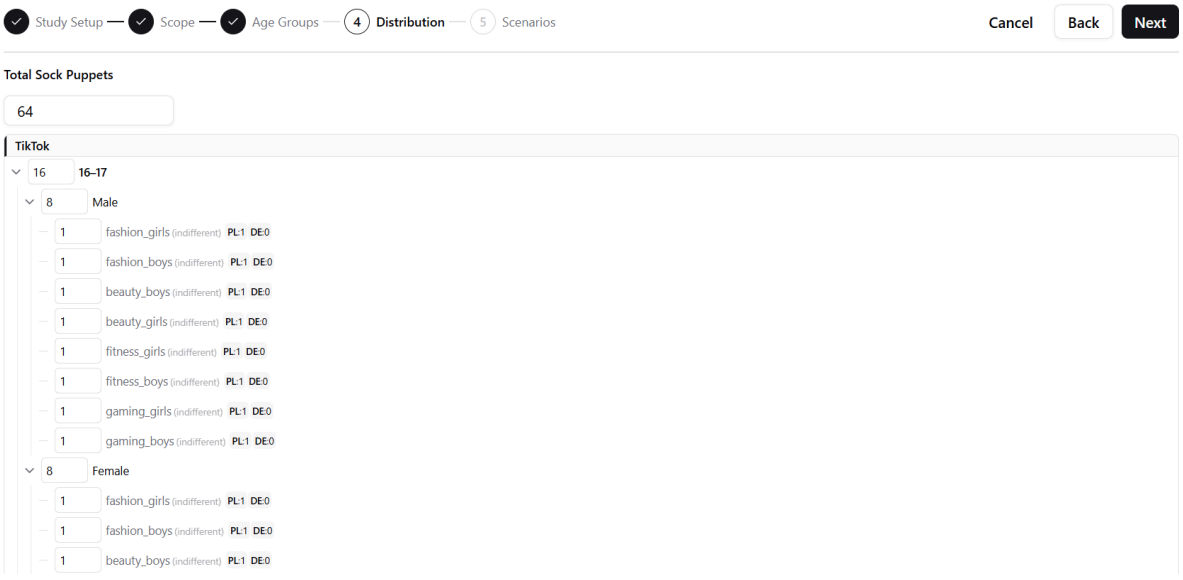


3. The third step defines *Age Groups*. A researcher defines the age ranges the study population will cover. Age groups can be shared across all selected platforms, or

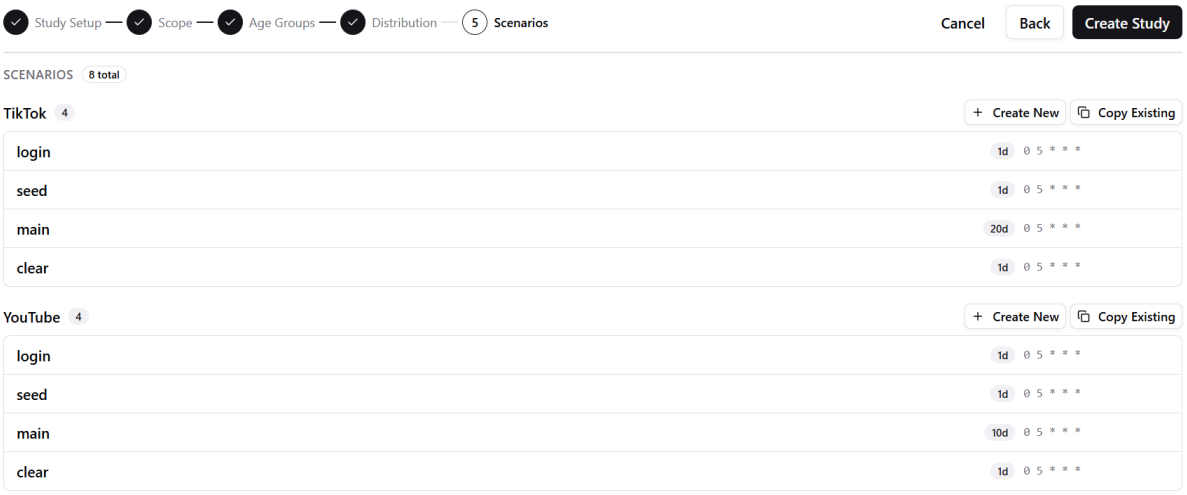
defined per platform. Per-platform mode is used when the audit needs platform-specific age targeting.



4. The fourth step is the definition of *Distribution*. This is where the abstract design becomes a concrete population. The wizard exposes a hierarchical distribution tree (Platform, Age Group, Gender, Topic, Country) and a total Sock Puppets count at the top. Changes at any level redistribute counts downwards. Changes at lower levels roll back up. The researcher can therefore drive the design top-down ("I want 68 accounts split evenly across these topics") or bottom-up ("I need exactly 4 accounts per topic-stance combination"). A footer continuously reports country usage as used / available, so the population can never silently exceed the proxy supply.



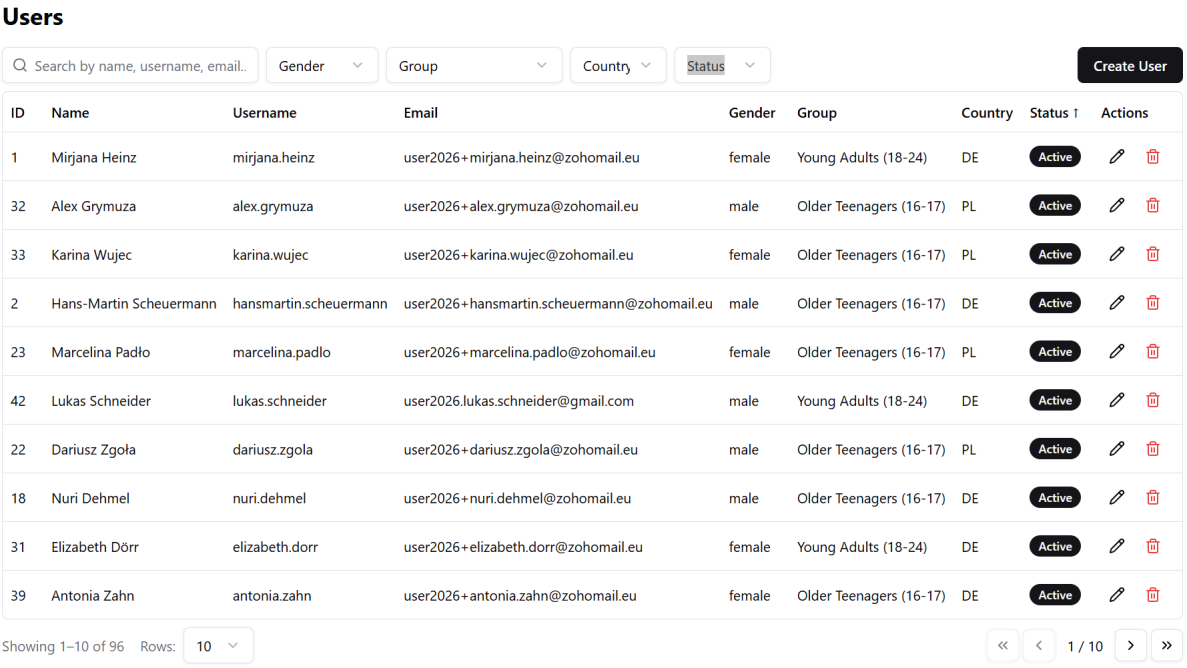
5. The fifth step is the definition of *Scenarios*. A researcher can define for each platform a list of the executable audit phases, their order and length of execution. These scenarios define what a sock puppet agent does during the browsing session.



When the researcher confirms the ‘Create Study’ action, the Control Panel does four things at once. It writes the study record, links the selected topics to the study, creates a single user population configuration describing the demographic target, and persists the scenarios.

3.3 Populations

After completing the Study Designer wizard, the researcher is then redirected to the *Populations* page, which allows to generate specific user profiles from the user population specified in the study design.

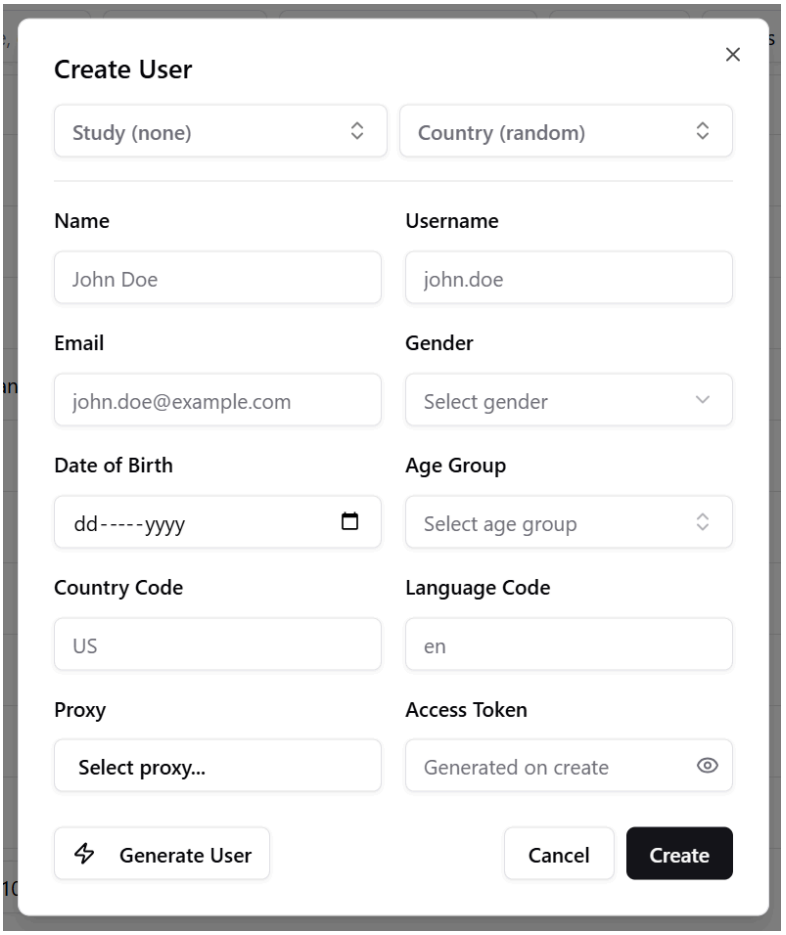


The Study Designer deliberately separates designing the user population from deriving the actual user profiles. The population is a target makeup, for example: "68 users, USA located, age 18 to 24, evenly split between male and female, allocated across four polarising topics and one neutral topic with the proportions shown above". The user profiles themselves carry

their own names, emails, dates of birth, access tokens, and assigned proxies. They only get materialised when the researcher opens the population and clicks “Generate Users”.

The Generate Users modal supports two complementary paths:

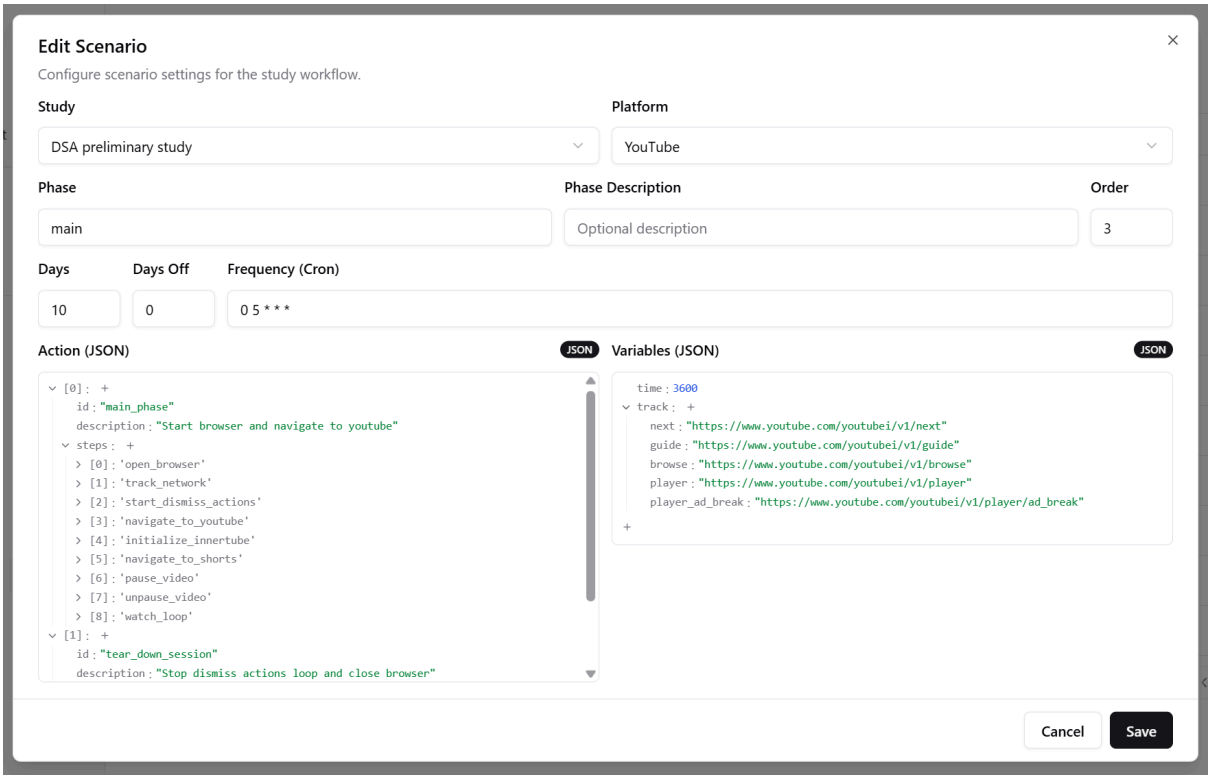
1. *Batch generation*. When the study has no enrolled users yet, the modal opens with an auto-generated preview list. The target count comes from the population configuration. Each preview row shows name, username, email, gender, country, age group, and the proxy the Control Panel has suggested for that country. The researcher can regenerate any row, edit any field, or remove a row. Clicking “Enroll All” then performs several actions at once. It creates the users, enrols them into the study, links them to topics, creates one User DAG per study platform, and pre-creates the first phase for each DAG. From that point on, the population is executable. Airflow has everything it needs to launch the first day of audit runs.
2. *Single-user creation*. A researcher can also open the Users page and create a user manually. When a study is selected at the top of the dialog, the country and age-group pickers narrow to those defined by the study's population configuration. The Generate User helper then fills in study defaults. After the user is created, a follow-up Enroll Created User modal opens. The researcher picks the platform (restricted to the study's platforms) and a subset of the study's topics. This path is used for ad-hoc additions, and for studies that need a specific, manually controlled topic assignment per user.



3.4 Audit Phases Design

A population on its own is inert. To turn it into an audit, the researcher must also describe what each sock puppet should do every day. In the Control Panel, that job belongs to Scenarios. The running state of those scenarios for each sock puppet is captured by User DAGs and Phases.

A scenario is a single execution phase of a study, defined for a single platform. Scenarios are created in two places. The first is the Study Designer's final step. The second is the standalone Scenarios page. The standalone page supports filtering by study, by platform, and by phase name. It is used for both inspection and fine-grained editing. Each scenario carries the following fields:



- *Study and Platform:* the scenario always belongs to exactly one study and one platform, so the same phase name (e.g. login, seed, main) can exist multiple times across the system for different combinations.
- *Phase*: a short identifier (login, seed, main, etc.) used in admin views, in run logs, and in replay metadata.
- *Phase Description*: optional free text describing the purpose of the phase, embedded later into the generated experiment YAML.
- *Phase Order*: an integer that defines the order in which the scenarios execute within the study. The User DAG starts on the lowest-ordered scenario and advances to the next one once the current one is complete.
- *Days*: the number of successful runs that must accumulate before the scenario is considered finished. For example, a main scenario with days = 15 corresponds to the 15-day Interaction Phase of the UMAP Neutral+Polarising group (Section 2.2). When

the successful run count reaches days, the phase is marked success and the DAG advances.
- *Days Off*: kept as metadata on the scenario record for future scheduling extensions.
- *Frequency (Cron)*: the schedule cadence at which the scenario runs, expressed as a standard cron string (typically once per day, but tunable per phase).
- *Action*: a structured JSON workflow describing what the audit pod will do during this phase. This is the most important field describing the translated platform-specific interaction steps and is described in detail in the deliverable D4.1.
- *Variables*: a JSON object of runtime parameters merged into the experiment's variables and made available to every step (time budgets, loop limits, network-tracking endpoint maps, thresholds, and similar tuning values).

4. Conclusion

In this deliverable, we presented the summary of a wide range of activities we undertook in the AI-Auditology project in order to overcome the drawbacks and challenging open problems of the current first-generation of algorithmic audits. Current approaches require a lot of effort and expertise as they excessively rely on manual inputs, particularly during specification of audit scenarios. As a result, audit scenarios remain created ad-hoc on intuition of researchers, incomplete (covering only a small subset of relevant user/content/interaction space) and inauthentic (heavily prescribed or too random). Consequently, current audits are oversimplified and artificial, and do not reflect the complexity of the real-world social media environment.

While introducing a novel paradigm of model-based algorithmic auditing as well as shifting to a concept of abstract audit scenarios, we pioneered the first audit scenario co-creation methods towards large-scale next-generation audits that will use an appropriate level of automation to overcome the limitations of the current audits.

As a result of our research efforts, and as demonstrated on the selected two use cases, we verified the potential of model-based auditing, which makes the audits: 1) more representative – they more comprehensively and systematically cover social media environment and more authentically replicate user behavior in comparison with the previous audit studies; and 2) cross-platform, longitudinal and multilingual – they are able to perform highly demanded comparison/benchmark of AI algorithms across multiple platforms and temporal dimension (continuously over time - to capture trends and changes in an AI) and spatial dimension (over multiple languages - to identify whether AI algorithms do not struggle with some specific languages, especially minor ones).

More specifically, we introduced methods for designing representative user populations (from which the individual user profiles are drawn), as well as specification of audit scenarios by means of audit phases that are platform-, time- and content-independent, thus allowing them to be deployed on multiple platforms and repetitively over time.

To provide appropriate means supporting a researcher (an auditor) during the audit scenario co-creation process, we developed the necessary software infrastructure. Besides the social media model (previously described in deliverables D2.1 and D2.2), the platform provides a Control Panel - a user facing web application that guides the users step-by-step through the audit design process. Within the Control Panel, their co-creation core design features, namely the Platform/Topics/Proxies enumerations, the Study Designer and the Populations management, let a researcher take an audit question and produce abstract audit scenarios in a few minutes through a guided UI. For example, the question "How does TikTok personalise content for 18 to 24 year-old US users seeded with polarising material?" yields a concrete demographic plan, a set of named sock puppet users with allocated proxies and topics, and a database state from which Airflow can immediately start launching audit runs.