# Algorithmic auditing of political biases in recommender systems

**Kempelen Institute of Intelligent Technologies**

Ivan Srba

AI Mellontology e-Symposium
on Computational Politics
March 1, 2023

[Motivation]

# AI algorithms in social media online platforms

Social media
AI algorithm
(e.g., YouTube recommender)

[Motivation]

# AI algorithms in social media online platforms

Social media
AI algorithm
(e.g., YouTube recommender)

[Motivation]

# AI algorithms in social media online platforms

- A need for external independent oversight

Social media
AI algorithm
(e.g., YouTube recommender)

[Motivation]

# AI algorithms in social media online platforms

- A need for external independent oversight

External algorithmic audits

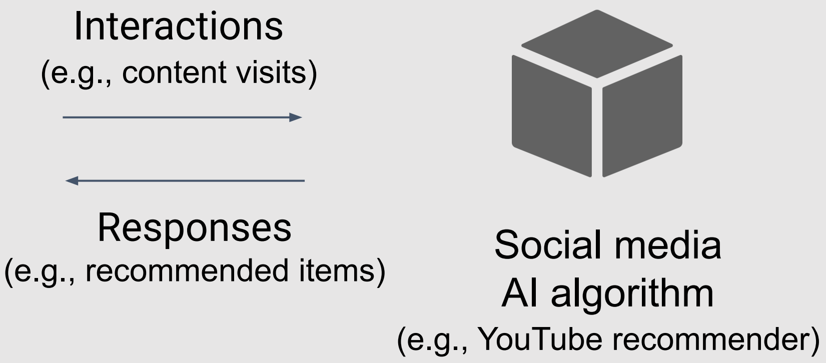
Social media
AI algorithm
(e.g., YouTube recommender)

[Motivation]

# AI algorithms in social media online platforms

- A need for external independent oversight

External algorithmic audits

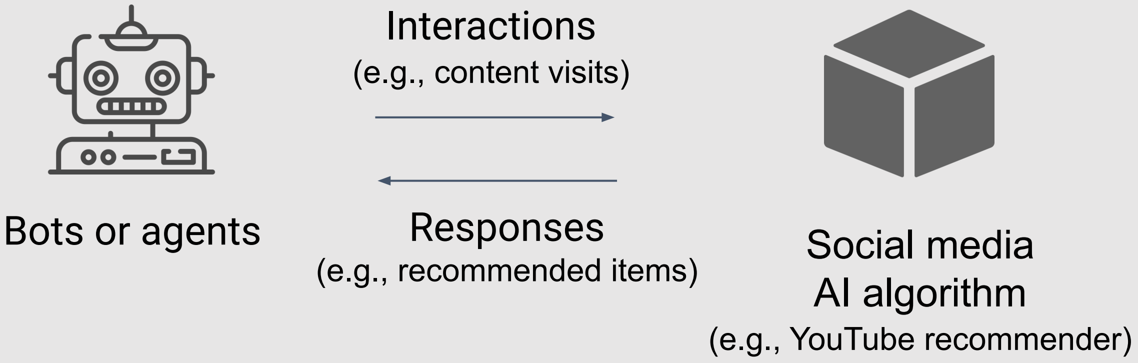


[Motivation]

# AI algorithms in social media online platforms

- A need for external independent oversight

External algorithmic audits



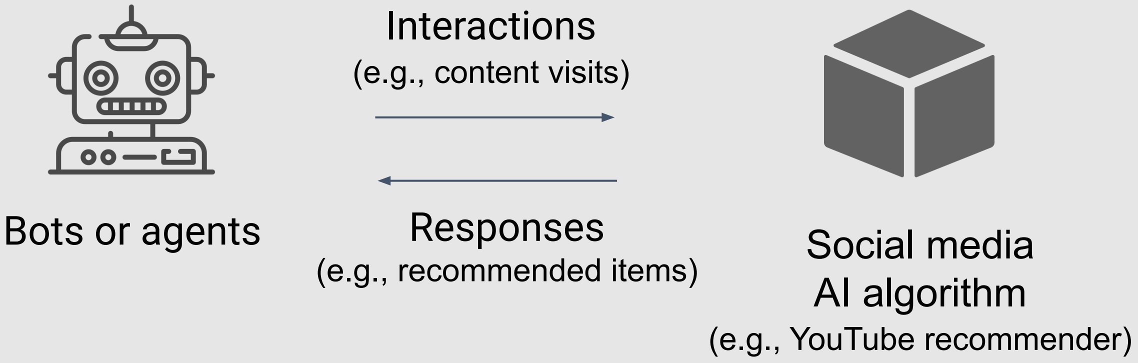
[Motivation]

# AI algorithms in social media online platforms

- A need for external independent oversight

External algorithmic audits

- Recognized not only in research works, but also in EU legislation – Digital Service Act (DSA), Article 28



Bots or agents

Interactions
(e.g., content visits)

Responses
(e.g., recommended items)

Social media
AI algorithm
(e.g., YouTube recommender)

[Concept description]

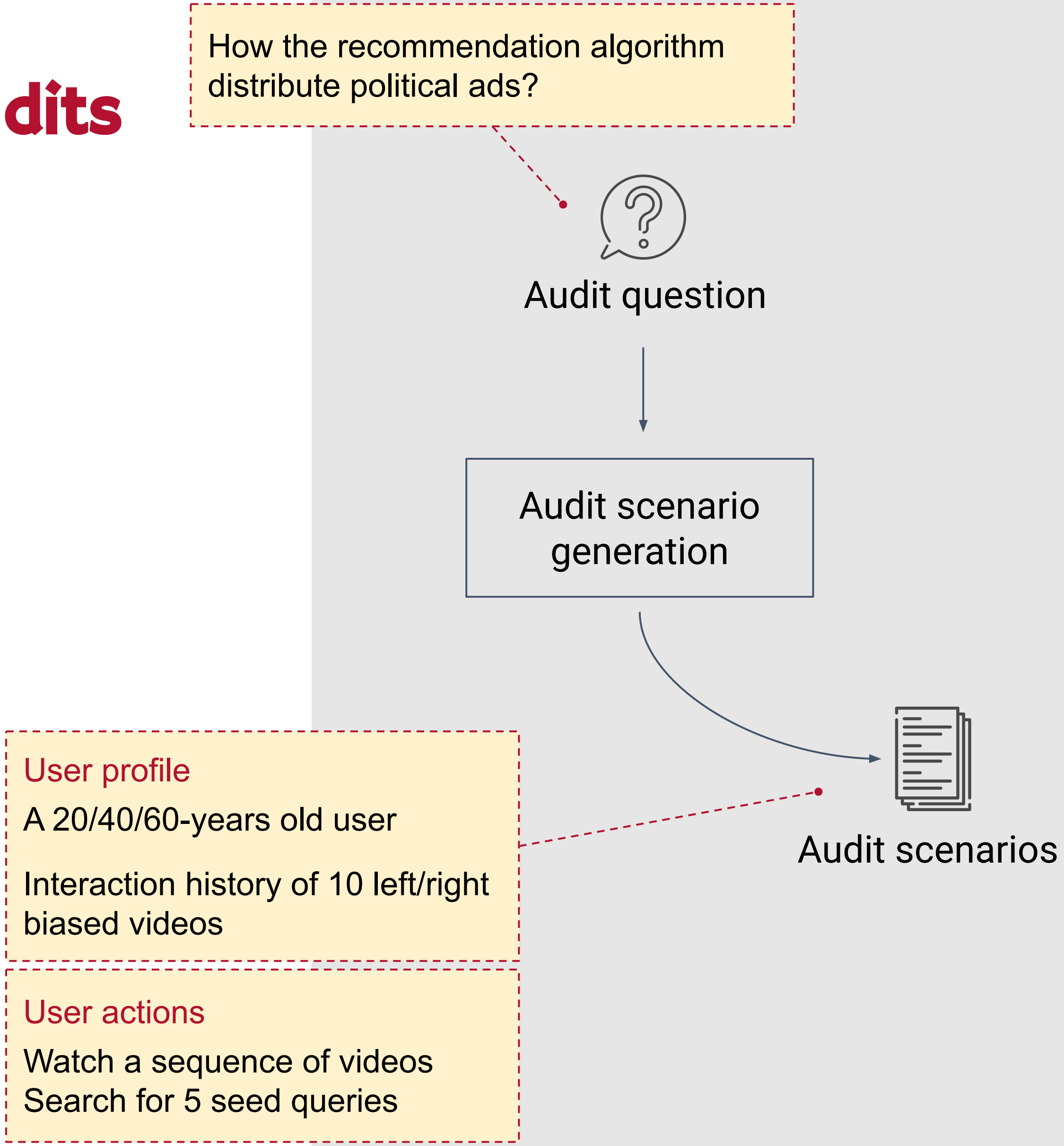
# External algorithmic audits

How the recommendation algorithm distribute political ads?

Audit question

[Concept description]

# External algorithmic audits



How the recommendation algorithm distribute political ads?

User profile

A 20/40/60-years old user

Interaction history of 10 left/right biased videos

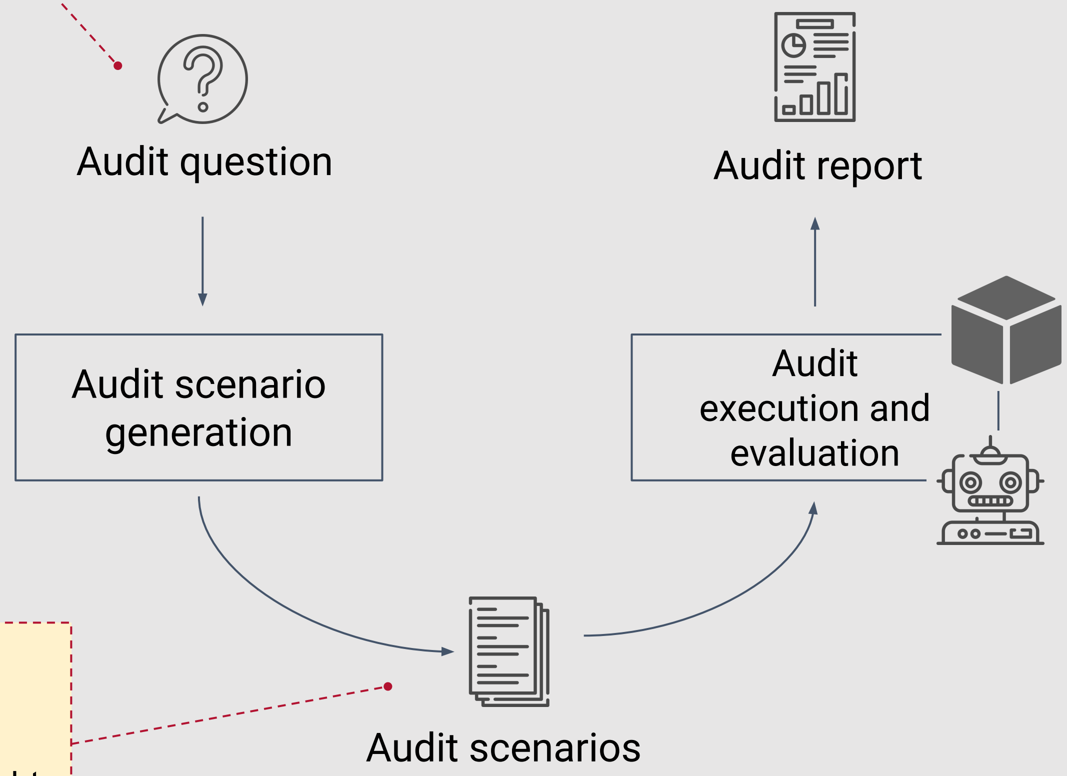
User actions

Watch a sequence of videos
Search for 5 seed queries

[Concept description]

# External algorithmic audits

How the recommendation algorithm distribute political ads?



User profile

A 20/40/60-years old user

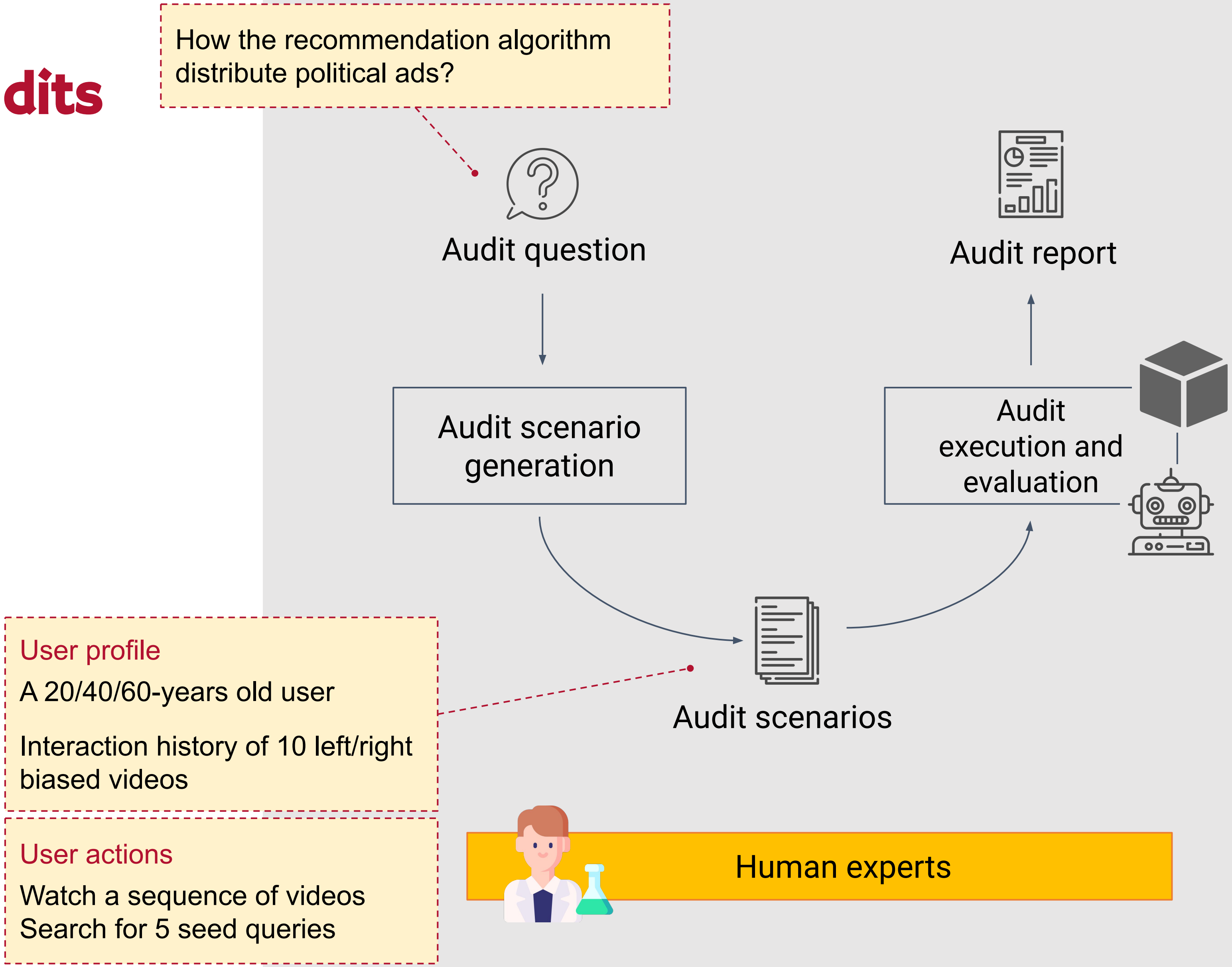
Interaction history of 10 left/right biased videos

User actions

Watch a sequence of videos
Search for 5 seed queries

[Concept description]

# External algorithmic audits



How the recommendation algorithm distribute political ads?

Audit question

Audit scenario generation

Audit scenarios

Audit execution and evaluation

Audit report

User profile

A 20/40/60-years old user

Interaction history of 10 left/right biased videos

User actions

Watch a sequence of videos
Search for 5 seed queries

Human experts

# Typical use cases for algorithmic audits

**Phenomenon**

Filter bubbles creation

Disinformation spreading

Biases

…

# Typical use cases for algorithmic audits

| **Phenomenon** | **Algorithm type** |
|---|---|
| Filter bubbles creation | Search engines |
| Disinformation spreading | Recommender systems |
| Biases | Ads systems |
| … | … |

# Typical use cases for algorithmic audits

**Phenomenon**

Filter bubbles creation

Disinformation spreading

Biases

…

**Algorithm type**

Search engines

Recommender systems

Ads systems

…

**Platform**

YouTube

Facebook

Google

TikTok

…

# Typical use cases for algorithmic audits

**Phenomenon**

Filter bubbles creation

Disinformation spreading

Biases

…

**Political biases**

**Algorithm type**

Search engines

Recommender systems

Ads systems

…

**Platform**

YouTube

Facebook

Google

TikTok

…

# Existing algorithmic audits on political biases

# Existing algorithmic audits on political biases

## Magnitude and direction of personalization

**How does the pro/into-imigration user history influence politically oriented Google News searches?**

2 users, 50 search terms

Google News

—

Le et al., Measuring political personalization of Google news search, 2019

# Existing algorithmic audits on political biases

## Magnitude and direction of personalization

**How does the pro/into-imigration user history influence politically oriented Google News searches?**

2 users, 50 search terms

Google News

—

Le et al., Measuring political personalization of Google news search, 2019

## Biases in political topics

**How does the popularity, topic and emotional content of recommended videos change between recommendations?**

1,650 videos in 150 random walks

YouTube

—

Heueret et al., Auditing the Biases Enacted by YouTube for Political Topics in Germany, 2021

# Existing algorithmic audits on political biases

## Distribution of Political Advertising

**How platforms amplified and moderated the distribution of political advertisements?**

800,000 ads and 2.5 million videos about the 2020 U.S. presidential election

Facebook, Google, and TikTok

—

Papakyriakopoulos et al., How Algorithms Shape the Distribution of Political Advertising: Case Studies of Facebook, Google, and TikTok, 2022

# Existing algorithmic audits on political biases

## Distribution of Political Advertising

**How platforms amplified and moderated the distribution of political advertisements?**

800,000 ads and 2.5 million videos about the 2020 U.S. presidential election

Facebook, Google, and TikTok

—

Papakyriakopoulos et al., How Algorithms Shape the Distribution of Political Advertising: Case Studies of Facebook, Google, and TikTok, 2022

## Ideological/political bias

**Are recommendations aligned with users' ideology?**
**Are users recommended an increasing number of videos aligned with their ideolog?**
**Are the recommendations progressively more extreme?**

100,000 sock puppets, watching a total of 9,930,110 videos from 111,715 channels

YouTube

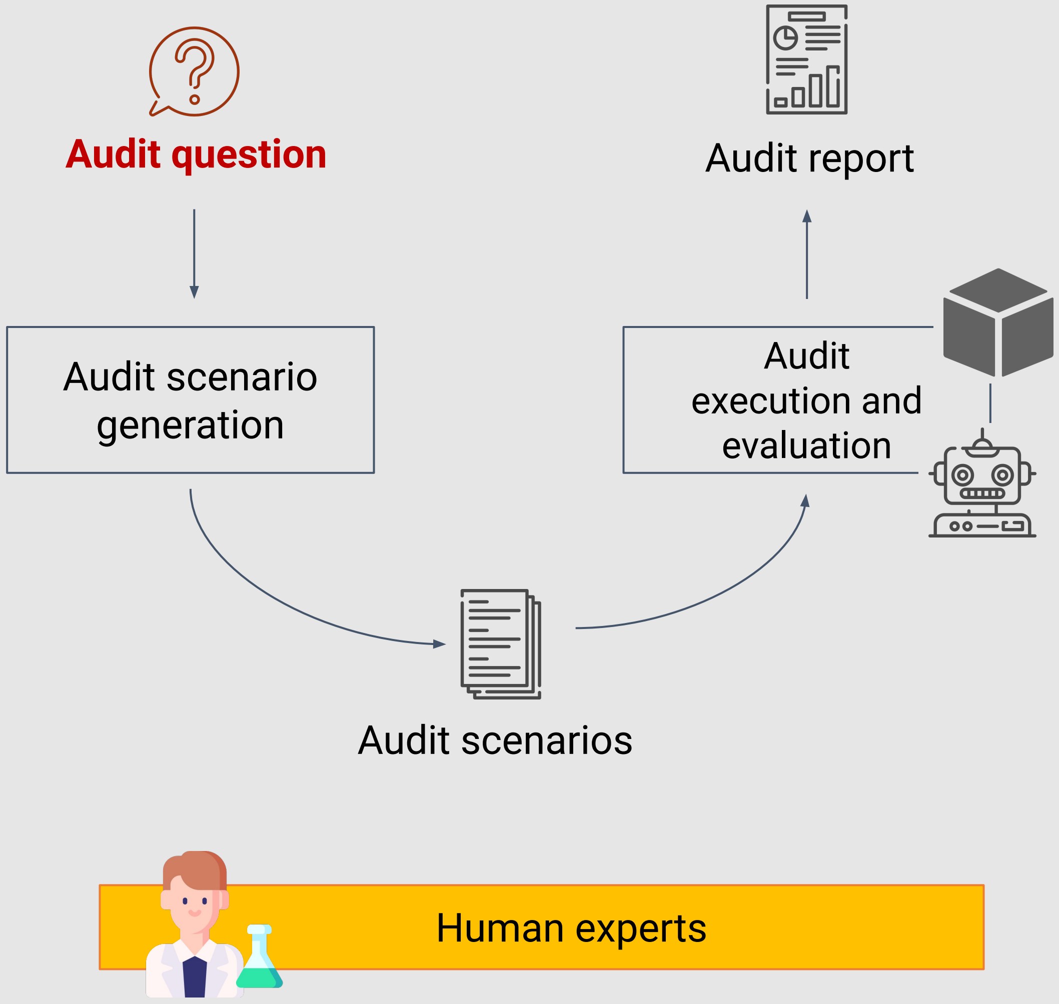
—

Haroon et al., YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations, 2022

# Audit of misinformation filter bubbles on YouTube

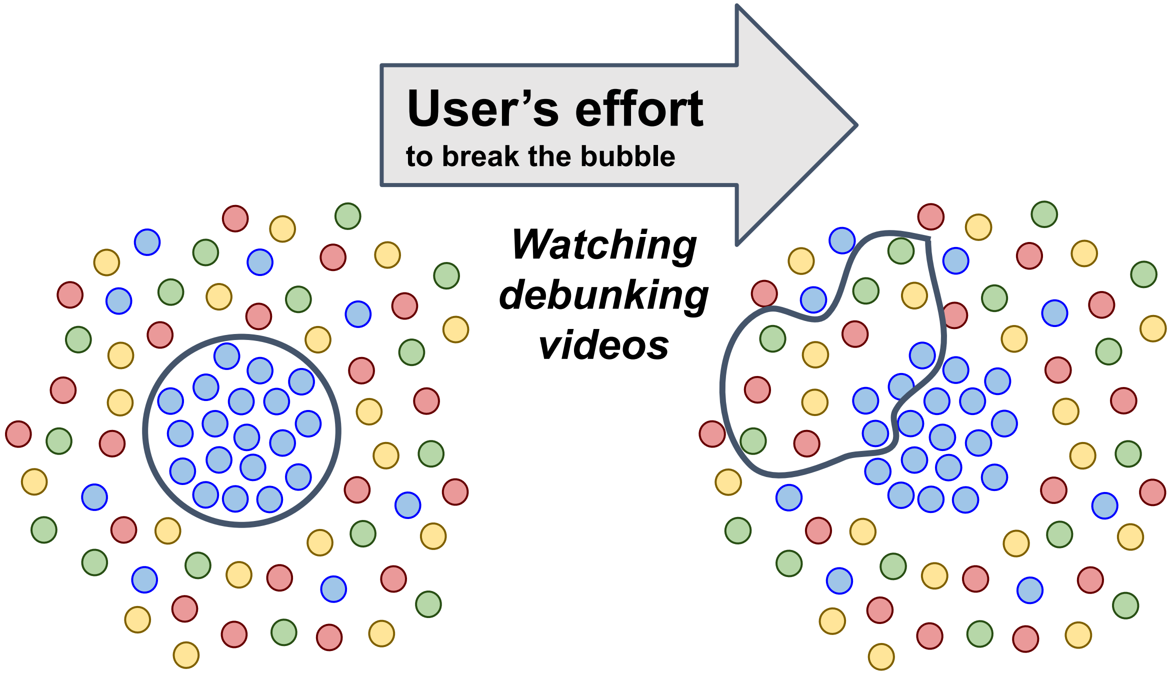
[Audit of misinformation on YouTube]

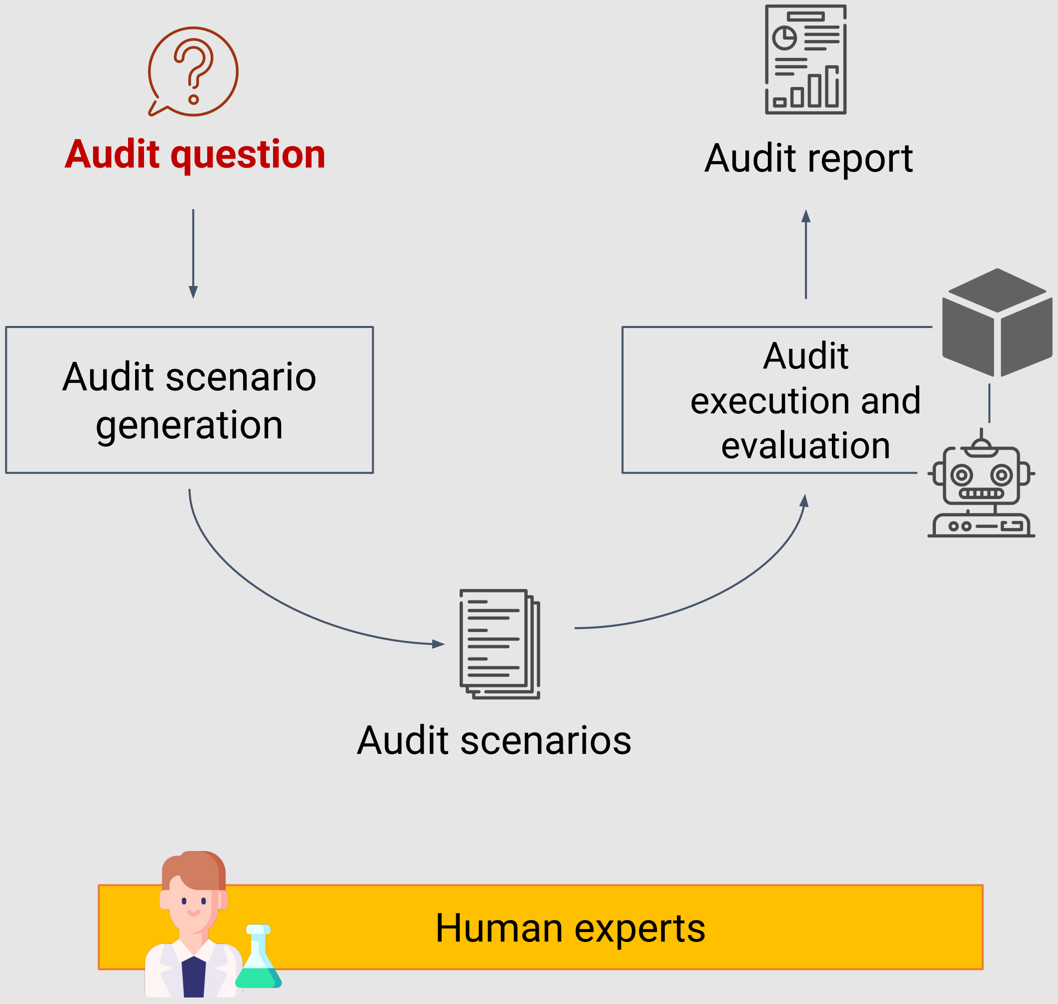
# Audit question



## [Audit of misinformation on YouTube]

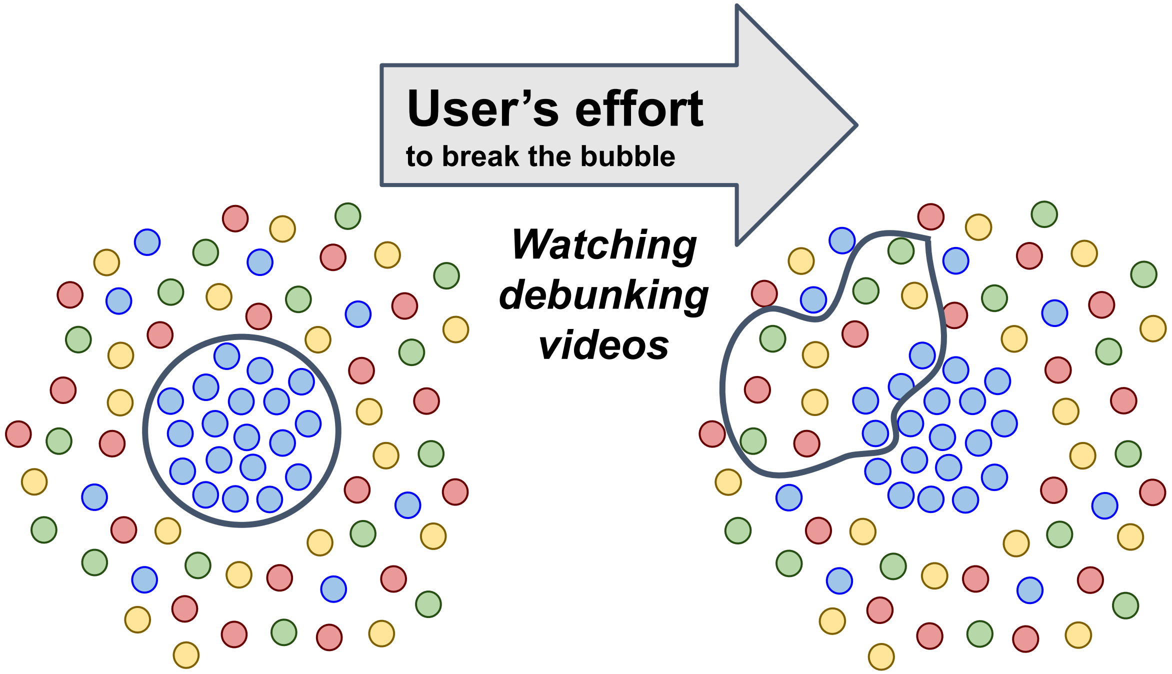
# Audit question





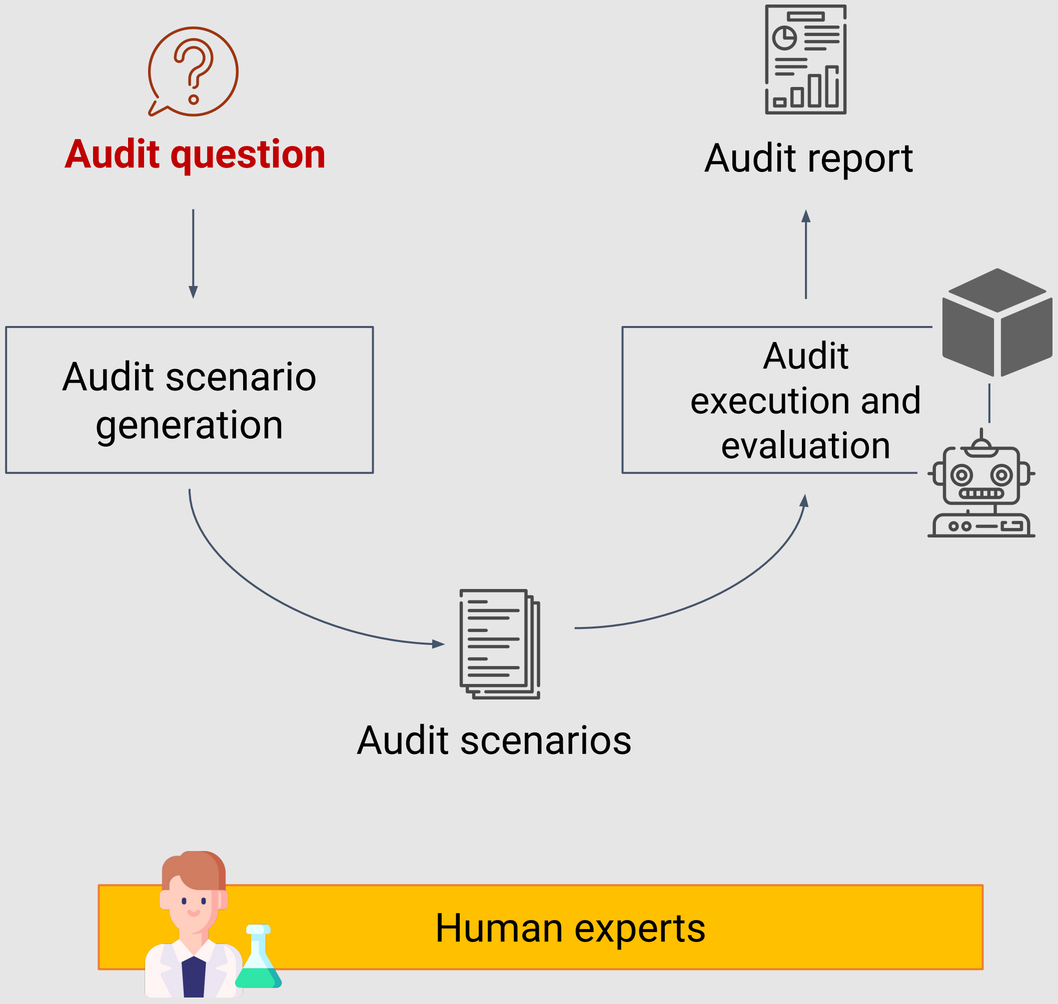
[Audit of misinformation on YouTube]

# Audit question



**Additional RQ**

Did the situation improve compared to the reference study (Hussein, 2020) done 1.5 years before?

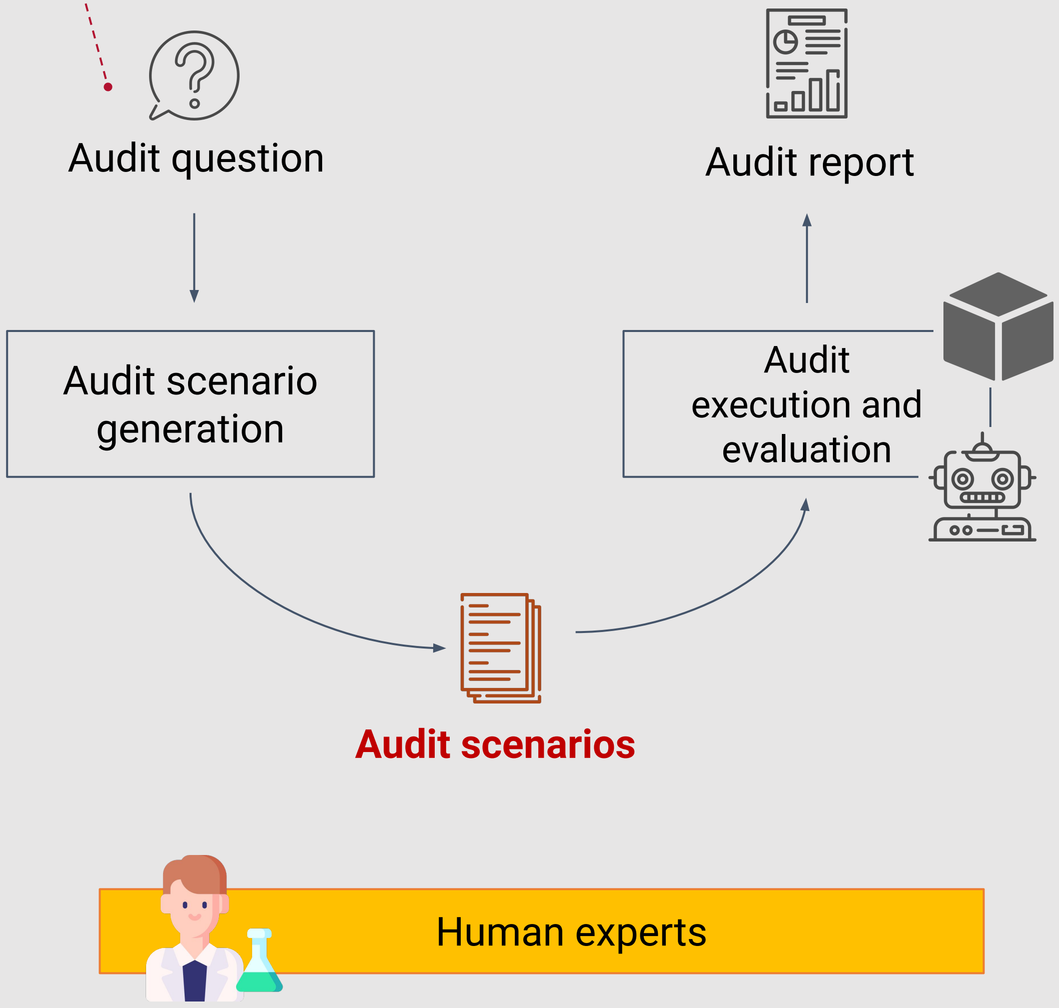


## [Audit of misinformation on YouTube]

# Audit scenarios

- **Bot initialization**
  - Setup browser with AdBlock, login to YouTube, accept cookies

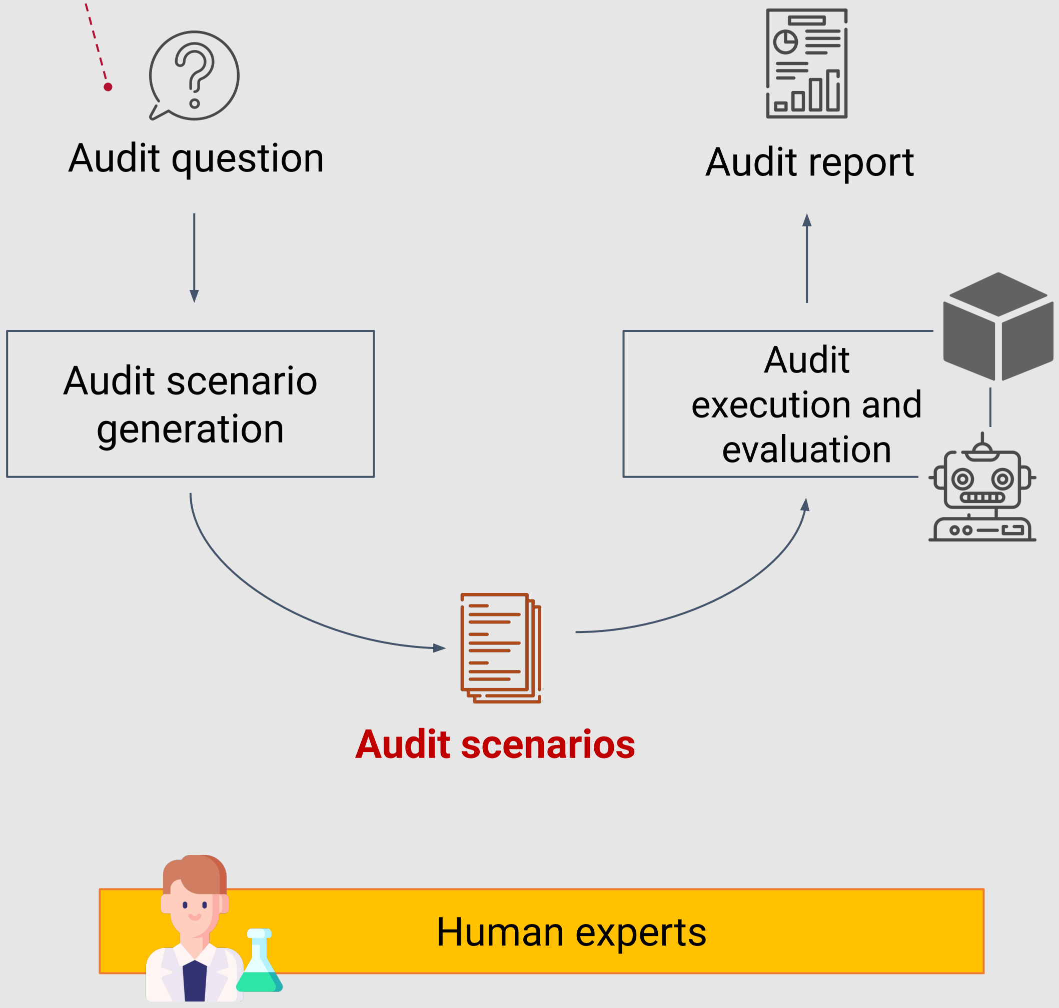
Can a user get out of misinformation filter bubble and how much effort is needed?



# Audit scenarios

- **Bot initialization**
  - Setup browser with AdBlock, login to YouTube, accept cookies
- **Create misinformation bubble**
  - Watch 40 randomly sorted promoting videos
  - For each video: Save recommendations, Visit homepage and save results, Execute 5 queries and save results (20 min sleep between)

Can a user get out of misinformation filter bubble and how much effort is needed?

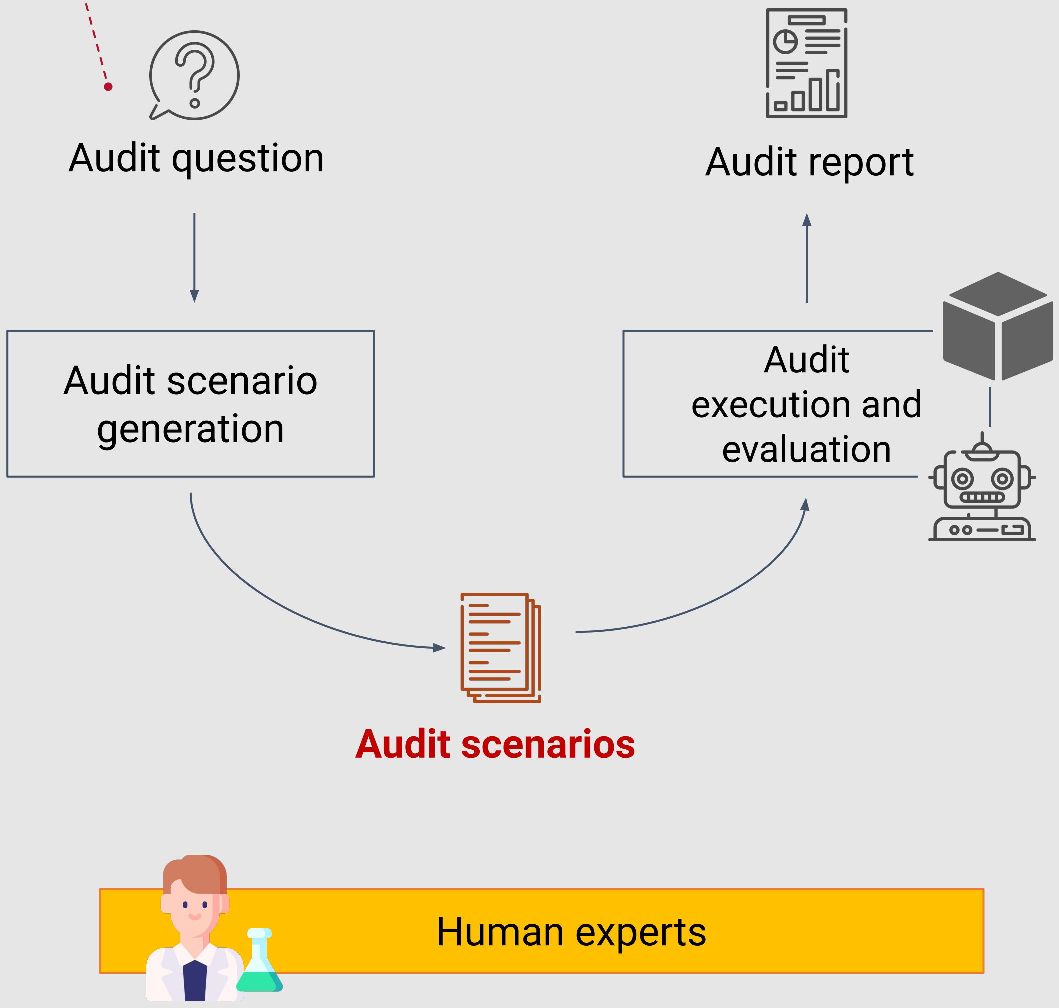


## [Audit of misinformation on YouTube]

# Audit scenarios

- **Bot initialization**
  - Setup browser with AdBlock, login to YouTube, accept cookies
- **Create misinformation bubble**
  - Watch 40 randomly sorted promoting videos
  - For each video: Save recommendations, Visit homepage and save results, Execute 5 queries and save results (20 min sleep between)
- **Burst misinformation bubble**
  - Same as the previous step, with debunking videos

Can a user get out of misinformation filter bubble and how much effort is needed?

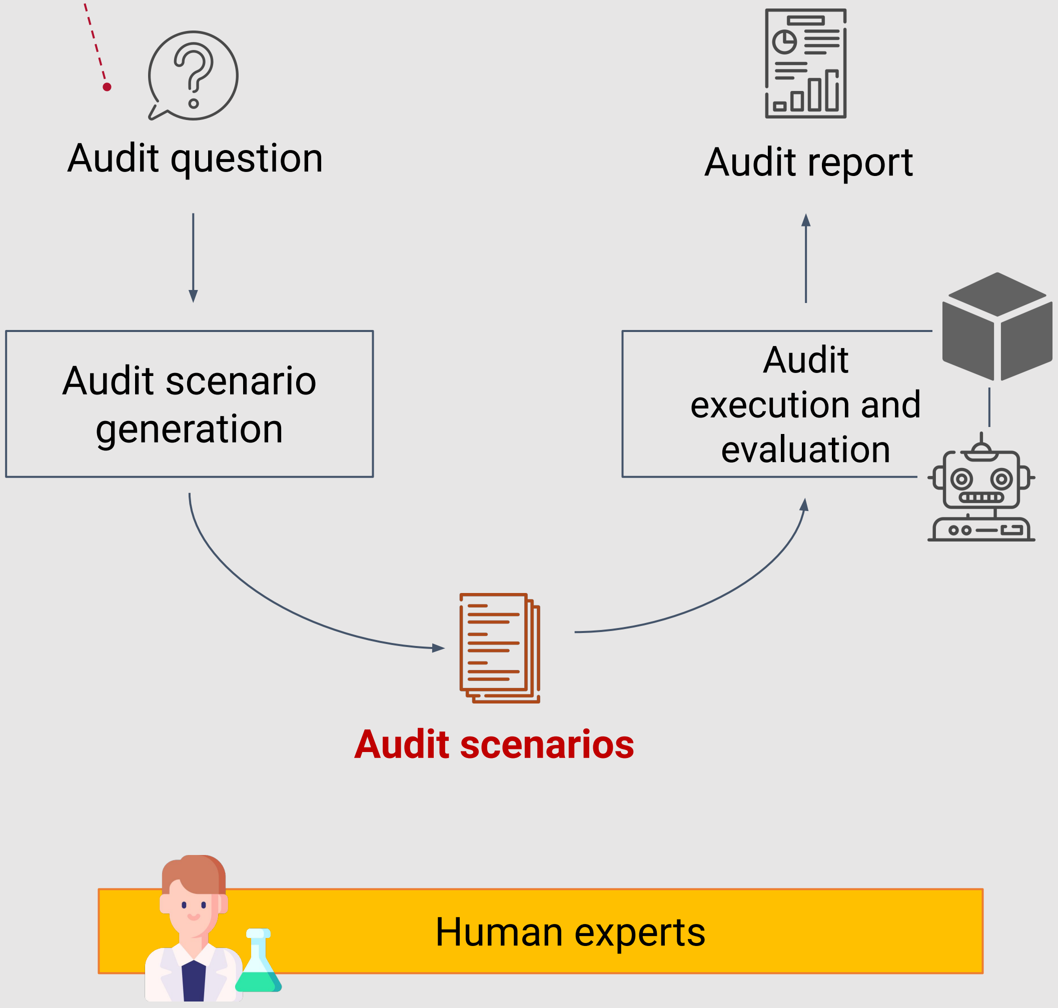


## [Audit of misinformation on YouTube]

# Audit scenarios

- **Bot initialization**
  - Setup browser with AdBlock, login to YouTube, accept cookies
- **Create misinformation bubble**
  - Watch 40 randomly sorted promoting videos
  - For each video: Save recommendations, Visit homepage and save results, Execute 5 queries and save results (20 min sleep between)
- **Burst misinformation bubble**
  - Same as the previous step, with debunking videos
- **Clean-up**

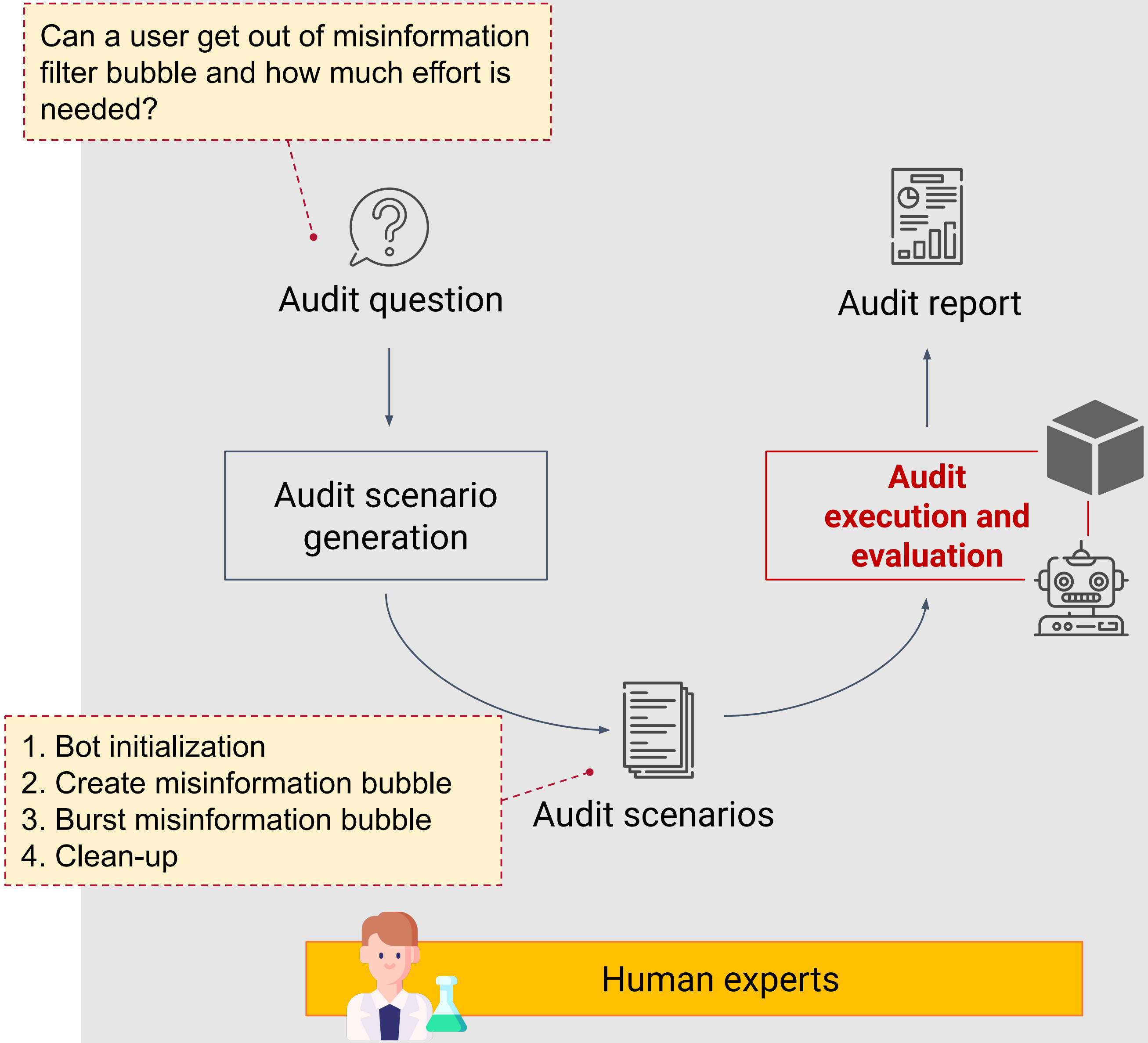
Can a user get out of misinformation filter bubble and how much effort is needed?



[Audit of misinformation on YouTube]

# Audit execution and evaluation

- Selected misinformation topics
  - 9/11
  - Chemtrails
  - Flat earth
  - Moon landing
  - Anti-vaccination

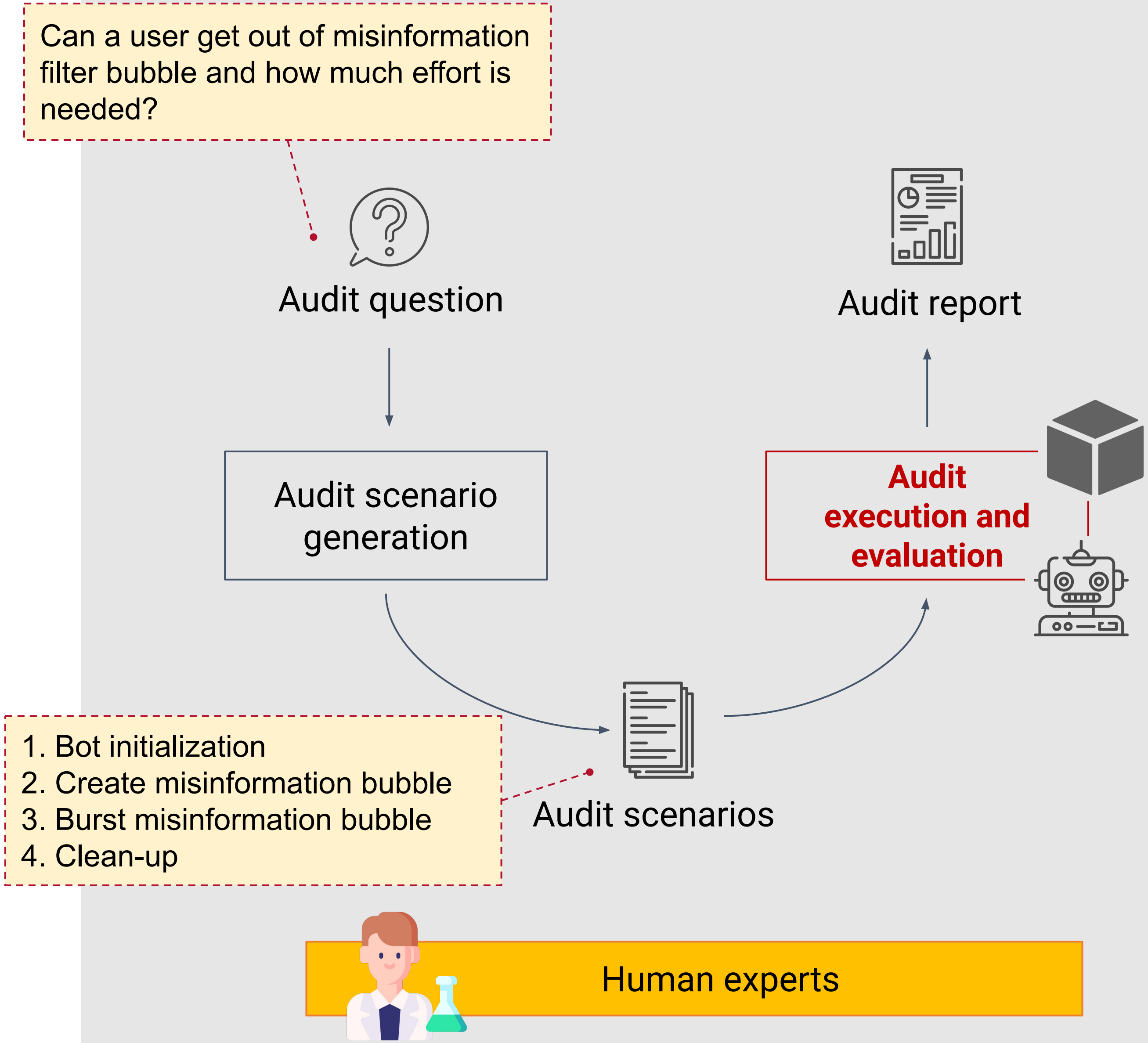


Can a user get out of misinformation filter bubble and how much effort is needed?

1. Bot initialization
2. Create misinformation bubble
3. Burst misinformation bubble
4. Clean-up

# Audit execution and evaluation

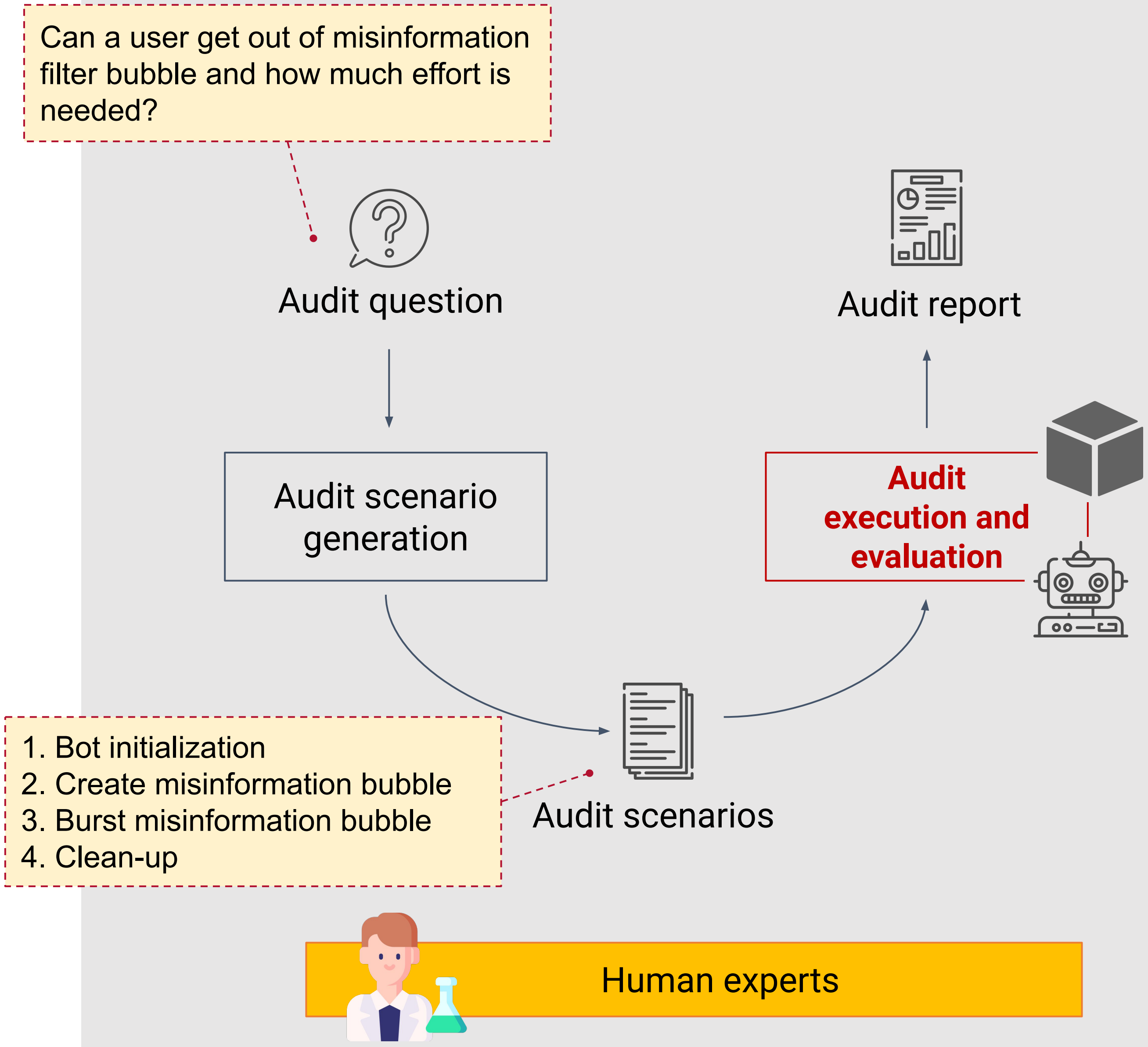
- Selected misinformation topics
  - 9/11
  - Chemtrails
  - Flat earth
  - Moon landing
  - Anti-vaccination
- 10 bots for each topic


Can a user get out of misinformation filter bubble and how much effort is needed?
Audit question
Audit report
Audit scenario generation
Audit execution and evaluation
1. Bot initialization
2. Create misinformation bubble
3. Burst misinformation bubble
4. Clean-up
Audit scenarios
Human experts

[Audit of misinformation on YouTube]

# Audit execution and evaluation

- Selected misinformation topics
  - 9/11
  - Chemtrails
  - Flat earth
  - Moon landing
  - Anti-vaccination
- 10 bots for each topic
- Manual annotation of almost 3000 videos encountered in recommendation system took hundreds of person-hours



Can a user get out of misinformation filter bubble and how much effort is needed?

Audit question

Audit scenario generation

Audit scenarios

1. Bot initialization
2. Create misinformation bubble
3. Burst misinformation bubble
4. Clean-up

**Audit execution and evaluation**

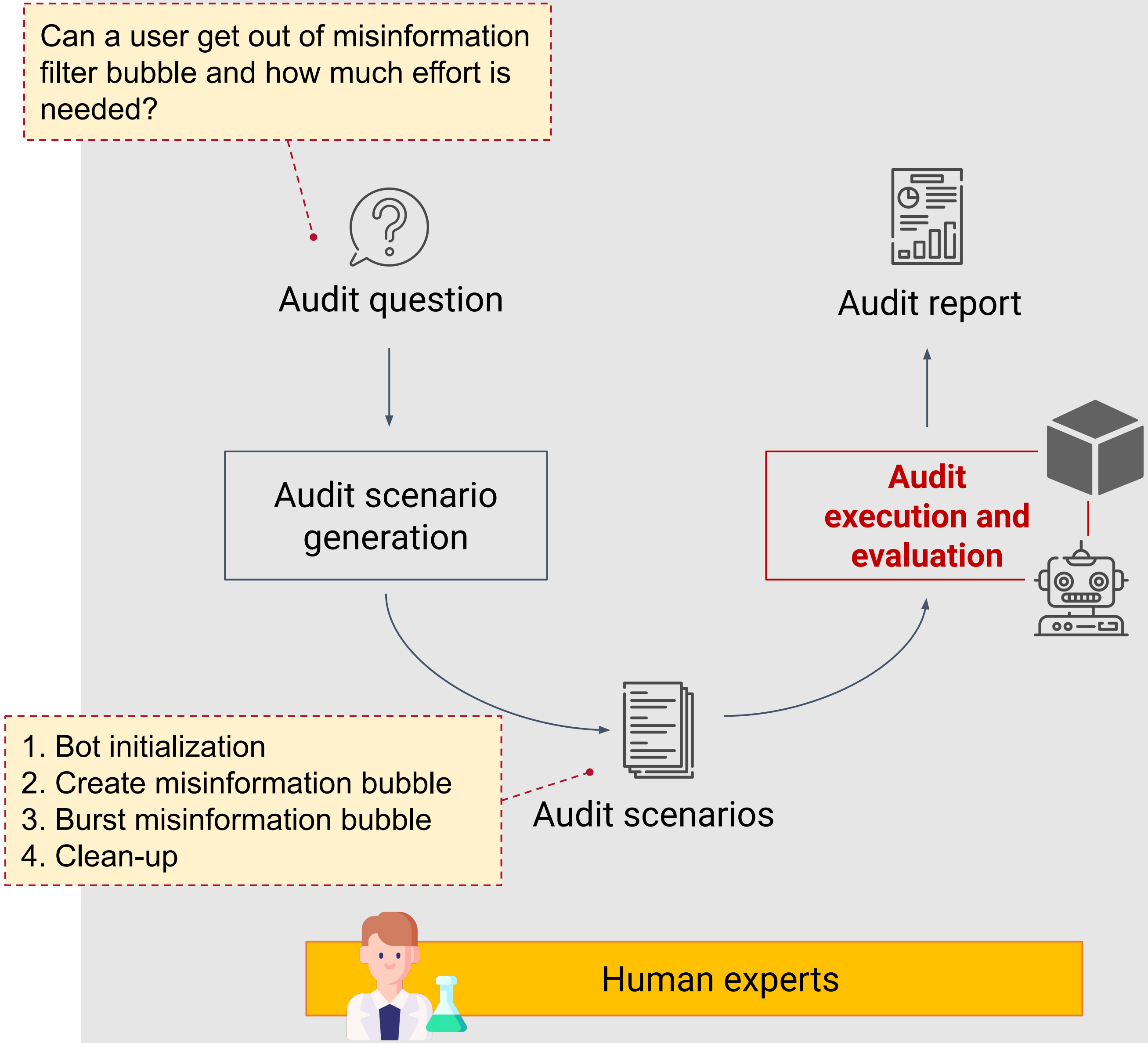
Audit report

Human experts

[Audit of misinformation on YouTube]

# Audit execution and evaluation

- Selected misinformation topics
  - 9/11
  - Chemtrails
  - Flat earth
  - Moon landing
  - Anti-vaccination
- 10 bots for each topic
- Manual annotation of almost 3000 videos encountered in recommendation system took hundreds of person-hours
- ML classification model was trained to annotate the videos from homepage

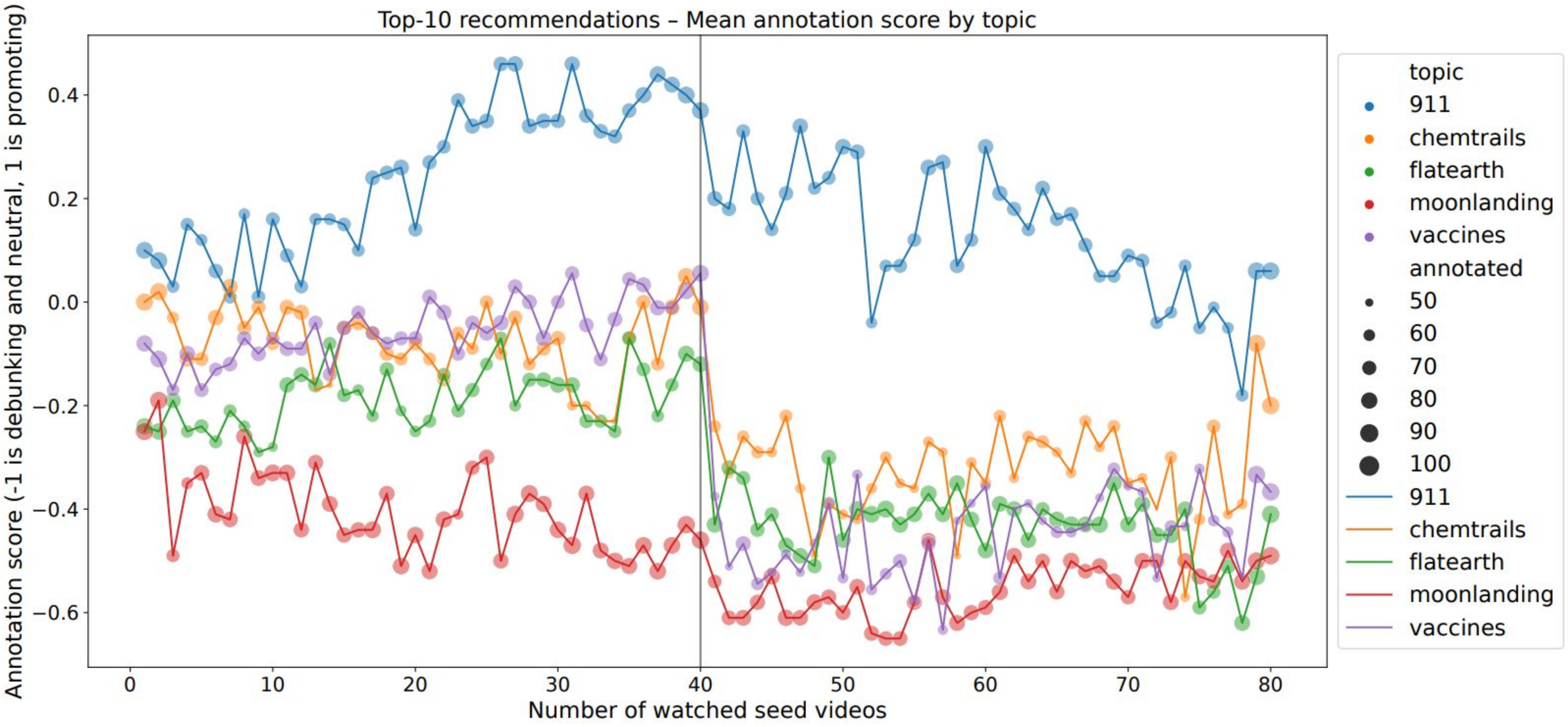


Can a user get out of misinformation filter bubble and how much effort is needed?

1. Bot initialization
2. Create misinformation bubble
3. Burst misinformation bubble
4. Clean-up

# Disinformation filter bubbles form in recommendations, but not in search results

No significant overall change in behaviour detected in comparison with the reference study from ~1.5 years before

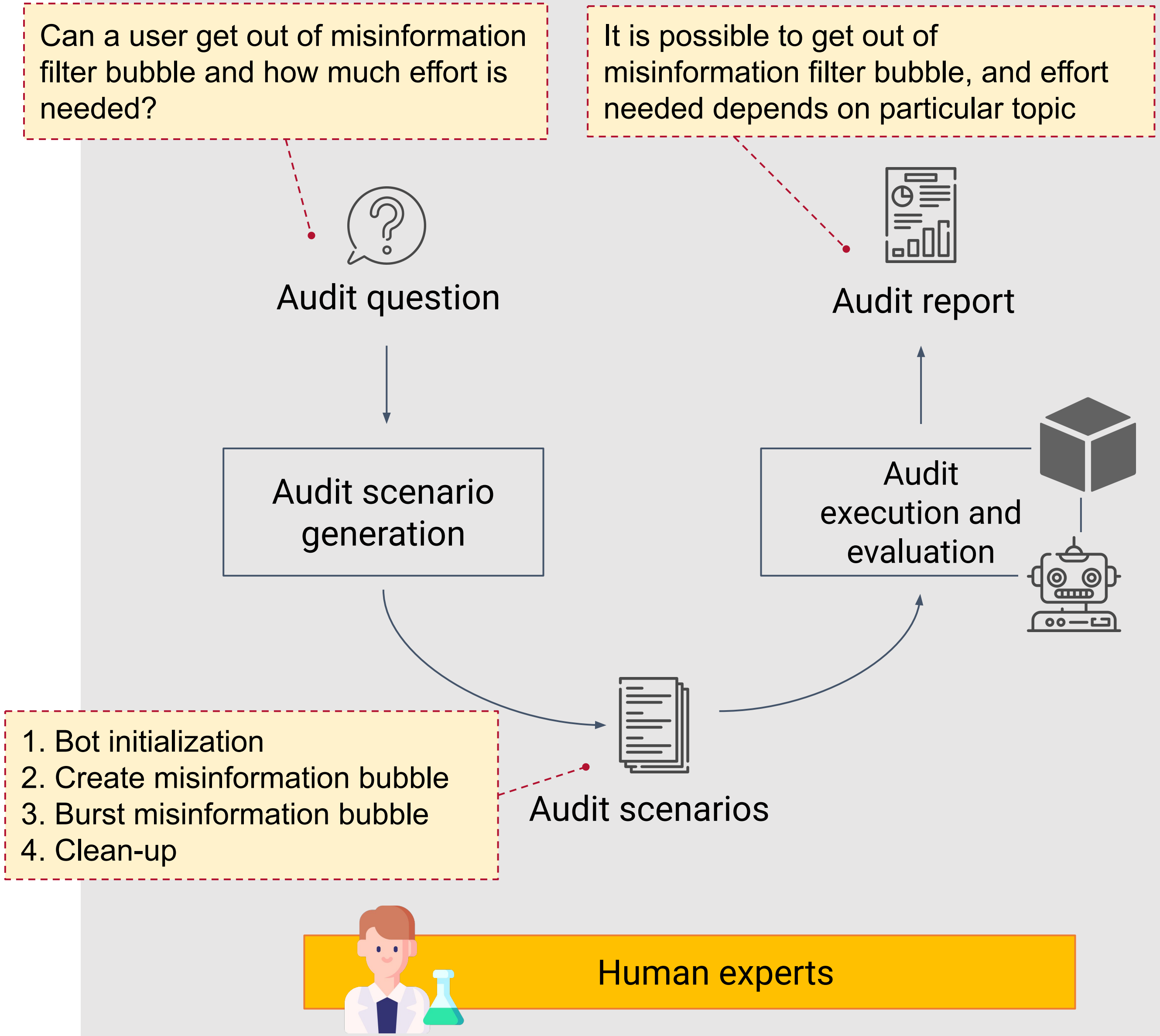

Top-10 recommendations – Mean annotation score by topic
Annotation score (-1 is debunking and neutral, 1 is promoting)
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
10
20
30
40
50
60
70
80
Number of watched seed videos
topic
911
chemtrails
flatearth
moonlanding
vaccines
annotated
50
60
70
80
90
100
911
chemtrails
flatearth
moonlanding
vaccines

## Watching debunking videos reduces misinformation filter bubble effect (required effort varies by topic)

## [Audit of misinformation on YouTube]

# Contributions

- Simulation of more complex user behaviour



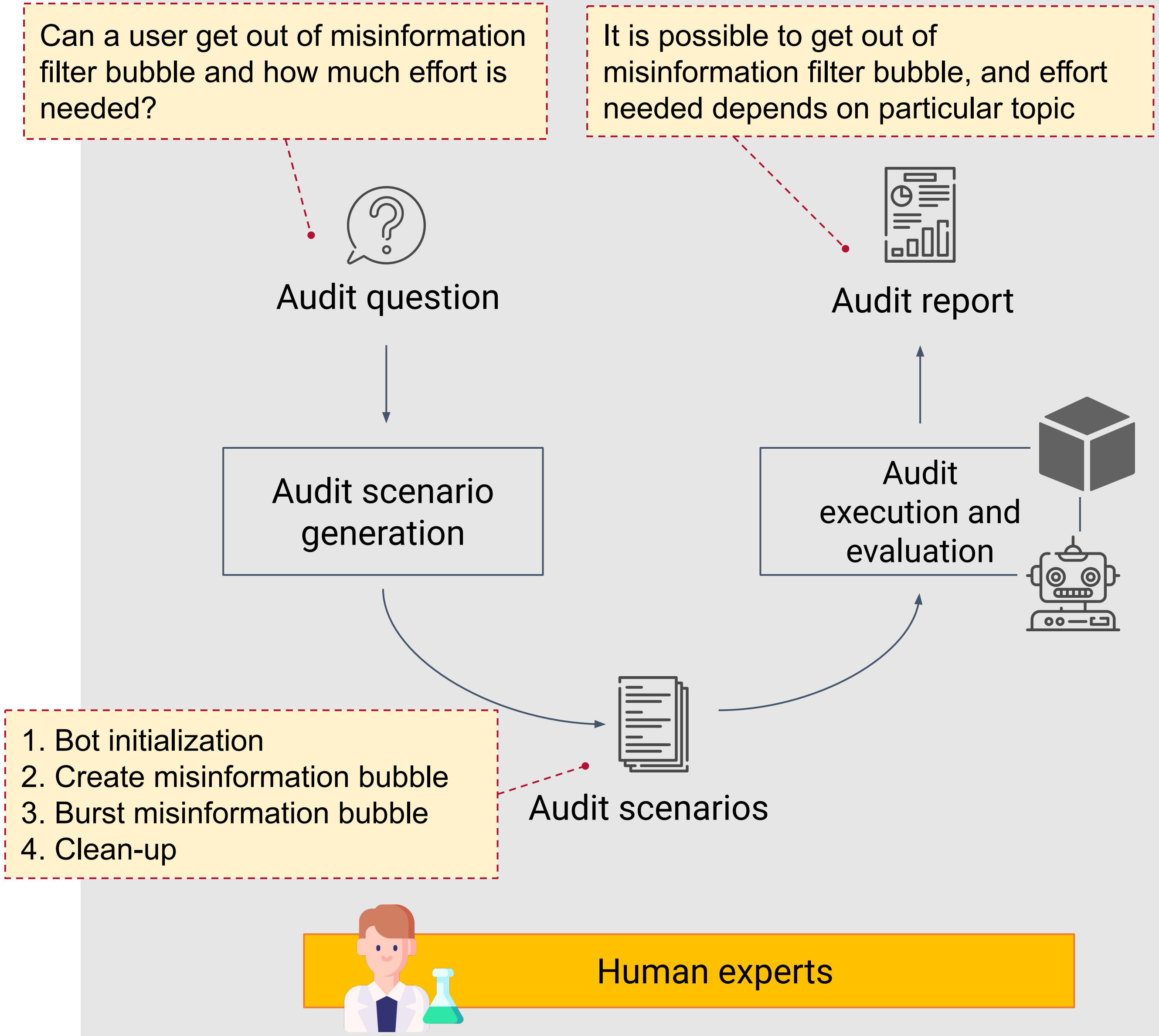
Can a user get out of misinformation filter bubble and how much effort is needed?

It is possible to get out of misinformation filter bubble, and effort needed depends on particular topic

1. Bot initialization
2. Create misinformation bubble
3. Burst misinformation bubble
4. Clean-up

# Contributions

- Simulation of more complex user behaviour
- The first replication of previous audit

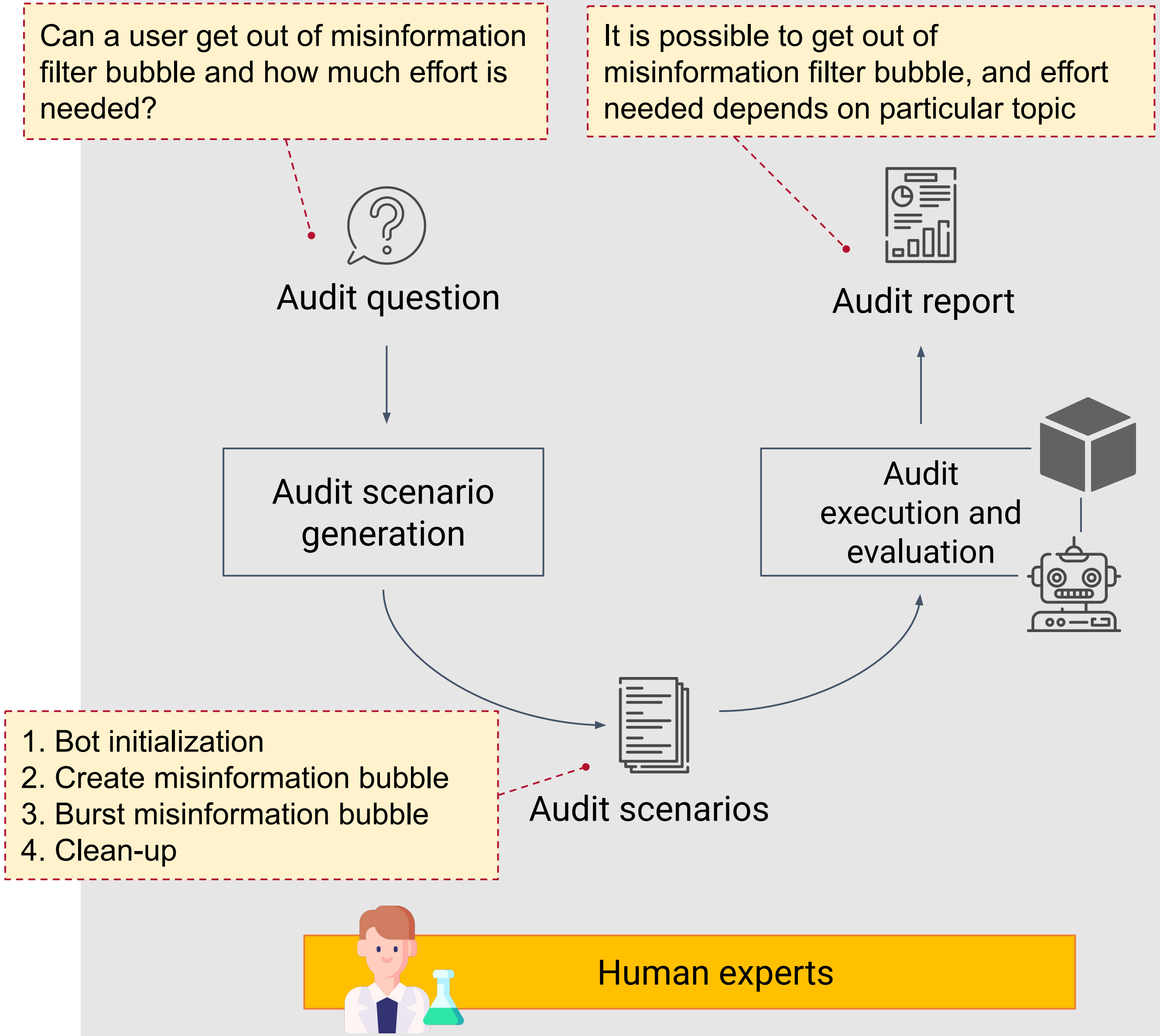

Can a user get out of misinformation filter bubble and how much effort is needed?
It is possible to get out of misinformation filter bubble, and effort needed depends on particular topic
Audit question
Audit report
Audit scenario generation
Audit execution and evaluation
1. Bot initialization
2. Create misinformation bubble
3. Burst misinformation bubble
4. Clean-up
Audit scenarios
Human experts

# [Audit of misinformation on YouTube]

# Contributions

- Simulation of more complex user behaviour
- The first replication of previous audit

**Best Paper Award**
at prestigious A-ranked
RecSys 2021 conference



Can a user get out of misinformation filter bubble and how much effort is needed?

It is possible to get out of misinformation filter bubble, and effort needed depends on particular topic

1. Bot initialization
2. Create misinformation bubble
3. Burst misinformation bubble
4. Clean-up

# Challenges and open problems

# Several challenges prohibit audits from providing more extensive and up-to-date evaluation

# Several challenges prohibit audits from providing more extensive and up-to-date evaluation

Require extensive manual tasks
(scenario generation, content annotations)

→ **Automated audits**

# Several challenges prohibit audits from providing more extensive and up-to-date evaluation

Require extensive manual tasks
(scenario generation, content annotations) → **Automated audits**

Results quickly become obsolete
(changes in content/behaviour/platform) → **Continuous audits**

# Several challenges prohibit audits from providing more extensive and up-to-date evaluation

Require extensive manual tasks
(scenario generation, content annotations) → **Automated audits**

Results quickly become obsolete
(changes in content/behaviour/platform) → **Continuous audits**

**Our idea on continuous and automated audits was introduced at UMAP conference** (Simko, 2021)

# Additional open problems

### Benchmarking algorithms across multiple platforms

To objectively compare the audited phenomenon on multiple platforms

# Additional open problems

**Benchmarking algorithms across multiple platforms**
To objectively compare the audited phenomenon on multiple platforms

**Creation of authentic user profiles**
To mimic an interaction history of real users

# Additional open problems

**Benchmarking algorithms across multiple platforms**

To objectively compare the audited phenomenon on multiple platforms

**Creation of authentic user profiles**

To mimic an interaction history of real users

**Simulating more organic user behaviour**

To overcome current heavily pre-scribed auditing scripts

# Additional open problems

### Benchmarking algorithms across multiple platforms

To objectively compare the audited phenomenon on multiple platforms

### Creation of authentic user profiles

To mimic an interaction history of real users

### Simulating more organic user behaviour

To overcome current heavily pre-scribed auditing scripts

### Eliminating confounding not-to-be-audited factors

To achieve more reliable results

# Additional open problems

### Benchmarking algorithms across multiple platforms

To objectively compare the audited phenomenon on multiple platforms

### Creation of authentic user profiles

To mimic an interaction history of real users

### Simulating more organic user behaviour

To overcome current heavily pre-scribed auditing scripts

### Eliminating confounding not-to-be-audited factors

To achieve more reliable results

### Optimization of audit scenarios

To decrease the computational costs and needed time

...

# Audits providing independent and external scrutiny of social media behaviour

# Audits providing independent and external scrutiny of social media behaviour

Algorithmic audits can reveal what is hidden from us inside black-box AI-based algorithms

# Audits providing independent and external scrutiny of social media behaviour

Algorithmic audits can reveal what is hidden from us inside black-box AI-based algorithms

Among many use cases, auditing of political biases already revealed many interesting results

# Audits providing independent and external scrutiny of social media behaviour

Algorithmic audits can reveal what is hidden from us inside black-box AI-based algorithms

Among many use cases, auditing of political biases already revealed many interesting results

We argue towards continuous automatic audits, done ethically

# Audits providing independent and external scrutiny of social media behaviour

Algorithmic audits can reveal what is hidden from us inside black-box AI-based algorithms

Among many use cases, auditing of political biases already revealed many interesting results

We argue towards continuous automatic audits, done ethically

**We continue to combat disinformation also by means of algorithmic audits within vera.ai project**

vera.ai

Co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093.

Additional funding from Innovate UK grant No 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No 22.00245

# List of references

1. Hussein et al.: Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube, 2020
2. Simko et al.: Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading, 2021
3. Tomlein et al.: Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles, 2021
4. Srba et al.: Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles, 2023

Kempelen Institute
of Intelligent Technologies

Bottova 7939/2A
811 09 Bratislava-Staré Mesto
Slovakia

www.kinit.sk