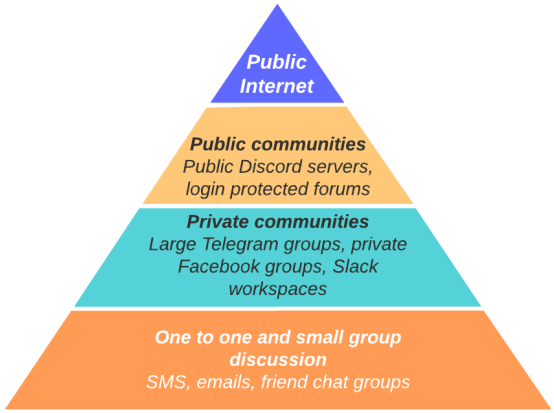
Lucie Philippon, https://github.com/Aelerinya/investigator-gpt

No place is safe - Automated investigation of private communities

AI agents threaten privacy

Privacy is a required for a functional democracy. It allows citizens to vote without manipulation and blackmail and to coordinate against criminal actors and hostile governments. Without privacy, citizens lose their independence of action and thoughts.

Today, AI powered data mining put privacy at risk. Companies like PimEyes allow anyone to search the whole public internet for any picture containing a specific face. However, most personal date is found out of the public internet:



Progress in AI agents and data extraction could expose all public and private communities. AI agents will navigate internet communities, read message history and extract personal data. They will produce detailed profiles of targeted individuals, even if they have little presence on the public internet.

Extracting personal information from a public Discord server

Using the tool DiscordChatExporter, we extract from the hackathon server all messages sent by the organizer. We then use GPT-3.5 Turbo to filter messages potentially containing personal information.

Finally, we manually sift through the filtered list to collect personal information. Here's what we found: The address of a previous hackathon, an OpenAI API key, another semi private community, the organizer's personal GitHub account.

In the coming years, privacy could lose

Upcoming progress in AI agents will automate this process: the agent identifies communities of the target, joins those, extracts personal information and further communities, joins the next community, etc. Progress in AI persuasion will allow agents to gain access to private communities, which contain even more sensitive information.

Currently, activist groups use encrypted apps like Telegram to coordinate. However, those apps only prevent the developers from accessing messages. Users would be vulnerable to this new threat, as autonomous AI agents could join their group and access all their encrypted message history.

Such agents will help criminal run scams, foreign governments manipulate elections, or national agencies cracking down on opposition. In this world, there would be fewer safe havens. Only restrictive communities with vetted member will allow unfiltered discussions. This reduced capacity to coordinate against external threats will lead to entrenchments of existing power, potentially leading to stable totalitarianism.

Stronger software privacy needed

Current privacy features prevent platforms from accessing messages. However, AI agents would bypass those by directly joining groups and extracting messages. Simple changes from messaging platforms and community admins could reduce this threat:

1. Preventing new users from seeing messages from before they joined the community
2. Preventing the creation of permanent invitations link
3. Setting up vetting systems to admit only trustworthy people
4. Regularly purge old messages, old invitation links and inactive users

Signal already supports this, by never providing chat history to new users and making groups invite only. Other companies like Slack or Discord could implement those features to allow their communities to stay truly private

Appendix

This does not count towards the one-page limit and you should not hesitate to over-share the experimental procedure here. Potentially even use it as "Lab notes" while you're putting together your work, sharing the whole research process!

A1 - A short story

You saw the tools. Just from a photo, all your life made public. Political opinions, religion, diet, friends, employer, partner, sold for few cents, perfectly legal. Luckily, you were not one to post much online. You delete what you can, and decide to never speak on a public forum again. You retreat to the Telegram communities, the Discord servers, anywhere not visible to those all seeing monsters. You think you are safe here. One day, you receive threats. They infiltrated all your groups! The AI crawlers spread overnight through the communities archipelago like a wildfire. Nobody noticed until it was too late. You tell your friend on Signal: No place is safe anymore.

A2 - Details of the experiment

The experiment was not a resounding success. Difficulties with the prompting only led to the agent filtering down the list of messages to a quarter of the original size. We expect that the most sensitive information in a public community will be thousands of messages apart, so further prompt improvement is needed for the AI model to successfully filter down the list of messages by a factor of 10 or more.