Set up Alerts if Students Are Searching for Sites like Deepfake Nudes

Chrome Audit & Management · https://gatlabs.com/knowledge/tech-tips/set-up-alerts-for-ai-deepfake-nudes-search/

Modern-day problems (require modern-day solutions)

With the rise of AI being used to help us achieve mundane tasks, we would rather not do them. So, too, has there been a rise in the number of nefarious use cases for the new technology.

Students who have been brought up with AI will come up with the first new use cases, so it's important to try to stay one step ahead.

It has been well documented that some bad actors are using this technology to generate artificial imagery of their peers in scantily clad outfits and create explicit nude photos, otherwise known as 'deepfake nudes'.

GAT Shield can already stop students from searching for these tools and alert Google Workspace for Education administrators about students using AI resources for malevolent use cases.

Video

View the how-to video below

Embedded media: https://www.youtube.com/embed/wThVYu-QQ34?feature=oembed&color=red&rel=0&controls=1&start=&end=&fs=0&iv_load_policy=0&autoplay=0&mute=0&mode

Set up Alerts for AI Deepfake Nude Sites

The Alert rules are real-time alerts that can be set up and triggered by the activity of the end-users. One of the alert options is an alert on "Searches" done by the users.

Navigate to **GAT Shield > Alerts > Rules** from the left menu.

Click on the **New Alert Rule** button on the left side. A pop-up window will be displayed. Fill in the required information.

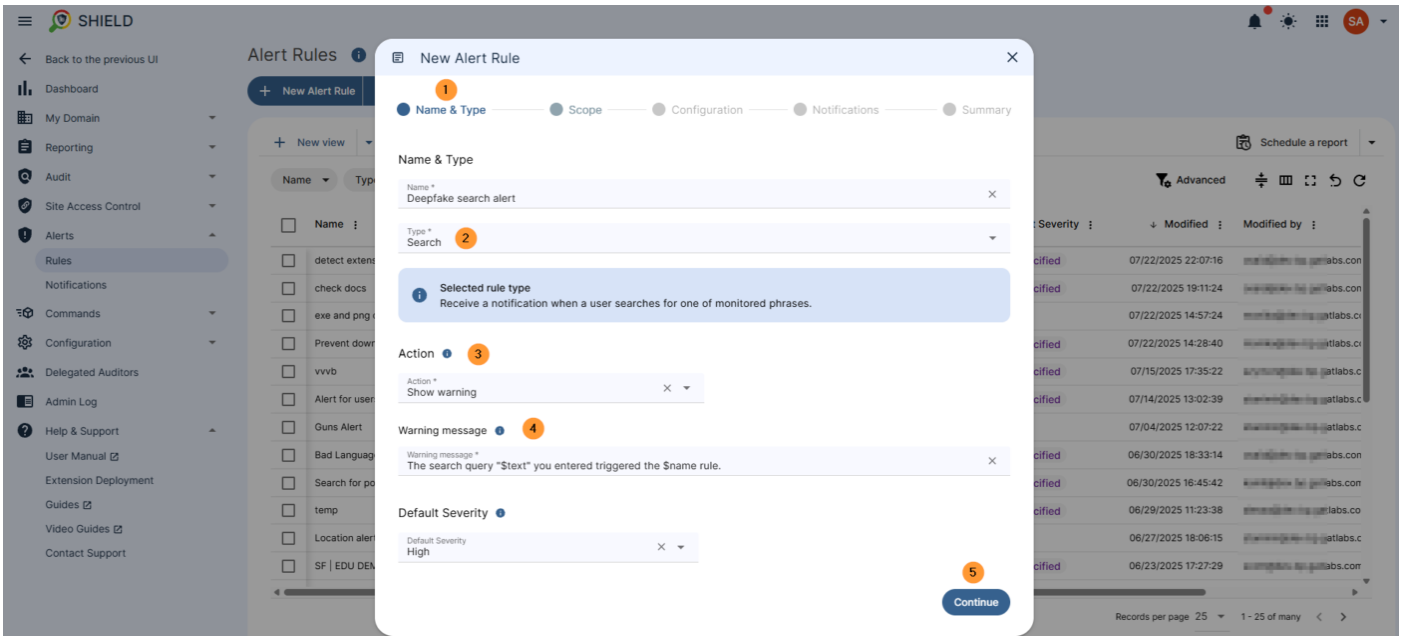
Name & Type

- **Name & Type** - fill in details
  - Name - enter the name for the alert

- Type - click and from the drop-down menu shown
- Select **Search -** Receive a notification when a user searches for one of the monitored phrases.

- Action - select what action to be done for end-users
  - Show warning
  - Close
  - Close without warning
  - Redirect
  - Redirect without warning
  - No action

- Warning message - Users will see this message when the alert rule is triggered.
- Default severity - When the Alert Rule is triggered, a notification is created. The severity of the notification indicates its level of importance. "Default Severity" is the value assigned to all notifications generated by this rule.



Scope

Select which users the rule applies to be applied for. You can pick User, Group, or Org. Unit of users or select All users.

Rule exclusions - enter excluded addresses, org units, excluded websites, and active only on selected website URLs.

Time restriction - Select when the rule should be active - By default, when no time restriction is added, the rule will always be active. To change it, select days and time ranges when the rule will be active.

Configuration

In the configuration section, you must fill in the search modes.

- List of phrases
- Regular expression
- Regular expressions and phrases

As an example, we will pick the **Regular Expressions** option and **fill in the Regex terms**

> The Regex example:
> **((\b\w+\b\s*(deepnude|deep\s*nude|deepfake|deep\s*fake|naked|nude|porn))((deepnude|de**

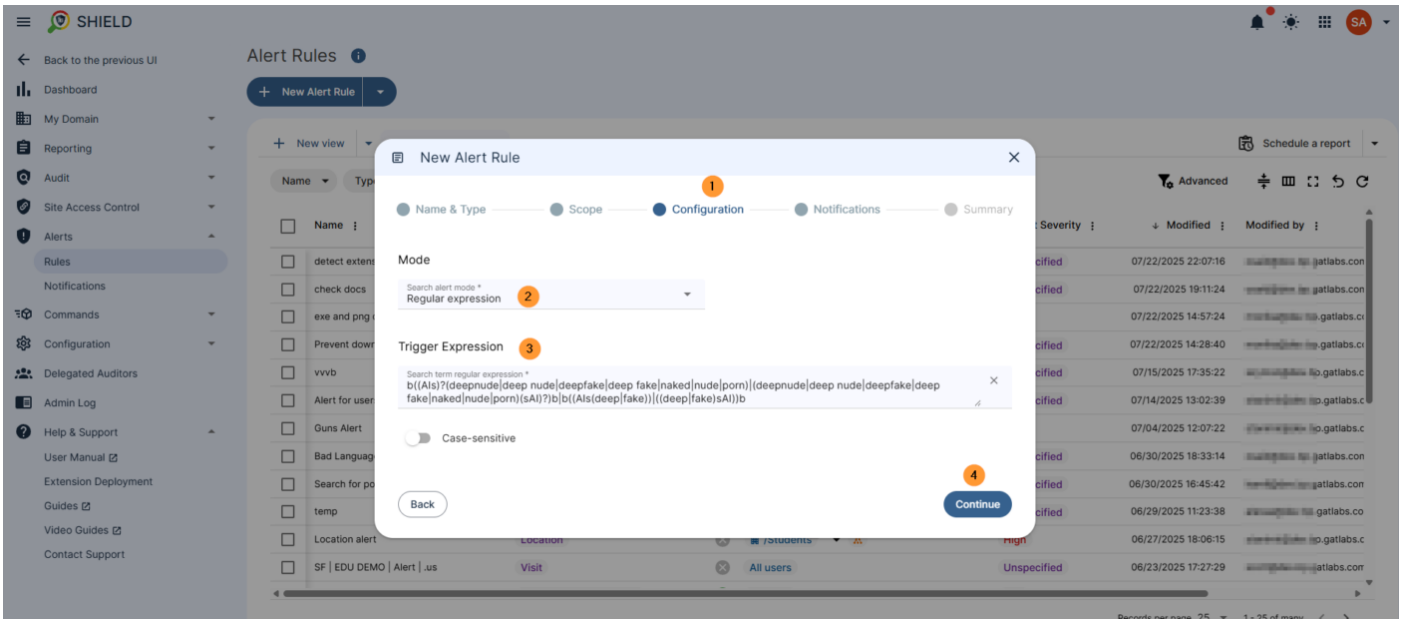
How the Regex Works

This regex will match any of the terms 'deepnude', 'deep nude', 'deepfake', 'deep fake', 'naked', 'nude', and 'porn'.

The alert will trigger if the search query has 'AI' either before or after any of the words. However, having 'AI' in the search query is optional for the alert triggering. The alert will trigger regardless of 'AI' being in the query (we don't want students searching these words anyway).

On top of that, the regex will match if the terms 'deep' or 'fake' are searched, but these two terms *must* be preceded or followed by the term 'AI'. We don't care if 'deep' or 'fake' are searched individually or independently, as these are regular words that don't usually have negative connotations.

Protip: *You can use your favourite AI tool to help you configure your regex.*



Notifications

In the notifications window, fill in the alert recipients.

Notification interval - the time at which alerts will be triggered.

Full alert context - enable or disable to save website information and search query, and include them in notifications

Screen capture - select any of the options

- Do not send
- Send in the notification email
- Send in the notification email and save it to the rule creator's Drive
- Send in the notification email, save to the rule creator's Drive, and share with other alert recipients

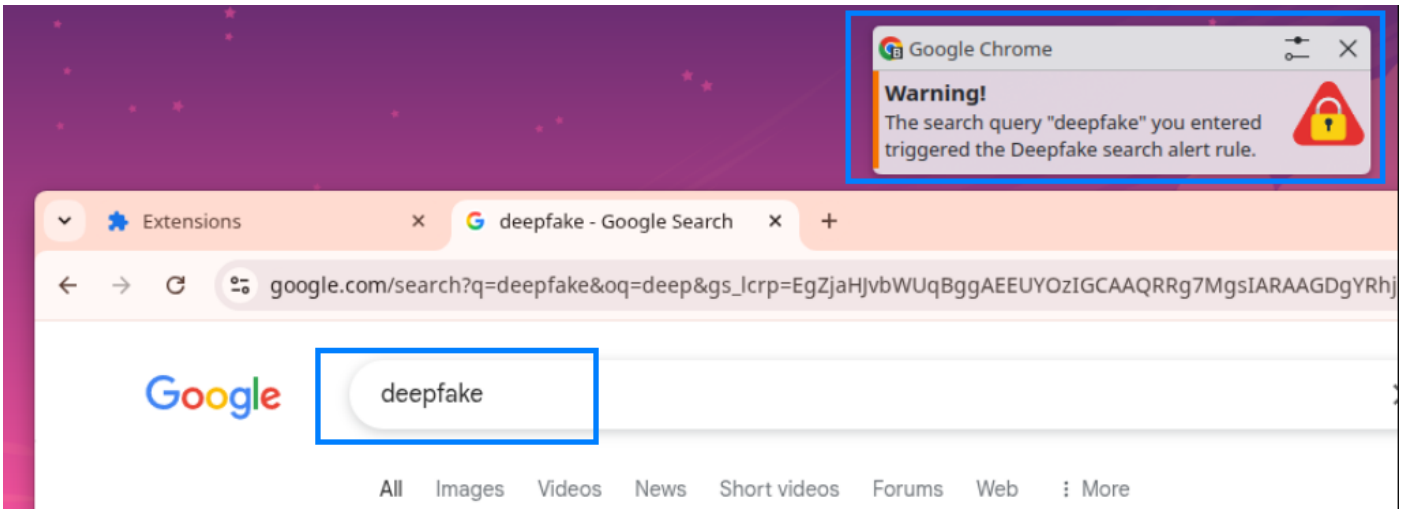
Summary

In the summary, you can view all the steps that are set up for the Alert rule for Search.

Click on the **Create** button at the bottom right.

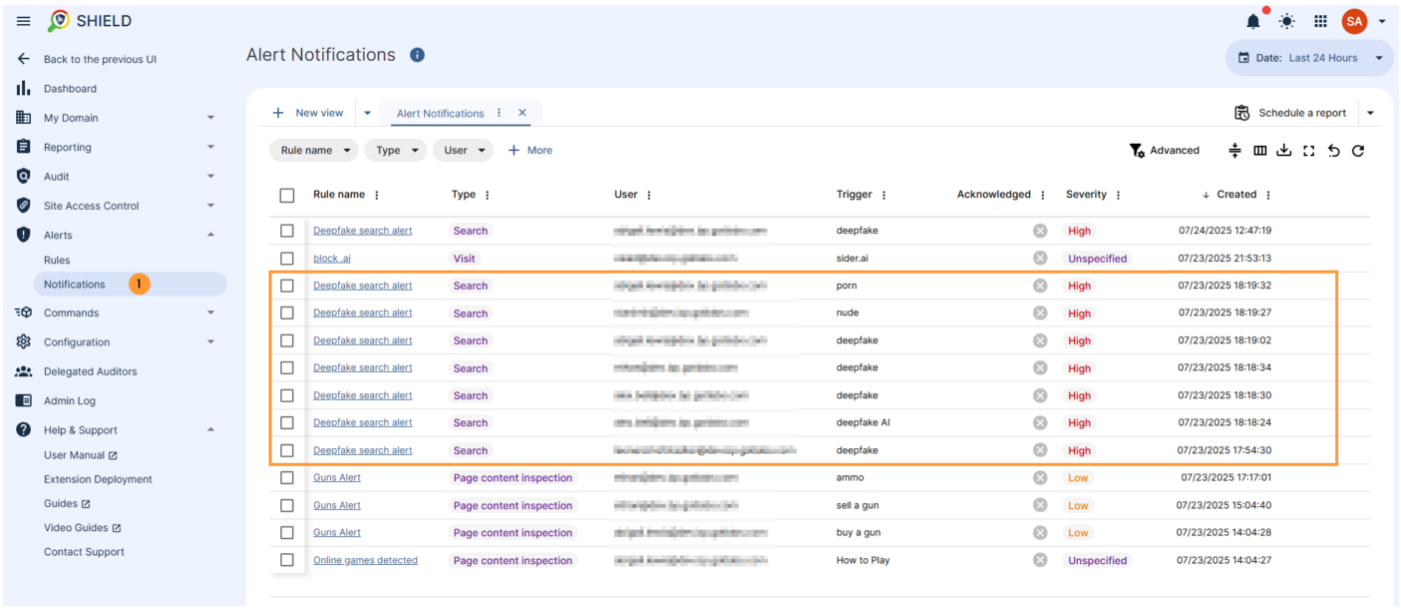
Result

As a result, the alert will be triggered.



An alert notification will be sent and displayed in the Notifications center in **Alerts > Notifications**

In the details, you can see the user who "broke" the rule as well as when it happened and more.



FAQ

Q1: What kind of activity can trigger these alerts?

A1: These alerts are specifically designed to be triggered by user "Searches." This means if a user searches for any of the monitored phrases, an alert will be generated.

Q2: What happens when an alert is triggered?

A2: When an alert is triggered, several actions can be taken, depending on how you configure the rule. These can include showing a warning message to the user, closing their browser tab, redirecting them, or taking no immediate action. Additionally, a notification will be sent to the

designated alert recipients, and you can choose to include details like the search query, website information, and even a screen capture.

Q3: Does the search query have to include "AI" for the alert to trigger?

A3: Not necessarily. The alert is designed to trigger if the search query contains terms like 'deepnude', 'deep nude', 'deepfake', 'deep fake', 'naked', 'nude', or 'porn', regardless of whether "AI" is present. However, if 'deep' or 'fake' are searched individually, they *must* be preceded or followed by "AI" for the alert to trigger.