AUTOMATING IMAGE ABUSE

DEEPFAKE BOTS ON TELEGRAM

October 2020

www.sensity.ai

October2020

Contact: info@sensity.ai

Authors: Henry Ajder, Giorgio Patrini, Francesco Cavalli

Graphic design: Eleanor Winter

Suggested citation: *Automating Image Abuse: Deepfake bots on Telegram*, Henry Ajder, Giorgio Patrini and Francesco Cavalli, Sensity, October 2020

About Sensity

Sensity (formerly known as Deeptrace) is the world's first visual threat intelligence company. Sensity provides individuals and organizations with solutions for detecting, monitoring, and countering threats posed by deepfakes and other forms of malicious visual media.

Acknowledgements

We would like to thank the following individuals for providing valuable feedback during the writing of this report:

Mahsa Alimardani, Danielle Citron, Adam Dodge, Mary Anne Franks, Sam Gregory, Nathan Hamiel, Ariel Herbert-Voss, Miguel Hernandez, Ayushman Kaul, Jean-Marc Rickli, Will Pierce, Max Rizzuto

Executive Summary

This report outlines Sensity's investigation into a newly uncovered deepfake ecosystem on the messaging platform Telegram. The focal point of this ecosystem is an AI-powered bot that allows users to photo-realistically "strip naked" clothed images of women.

Compared to similar underground tools, the bot dramatically increases accessibility by providing a free and simple user interface that functions on smartphones as well as traditional computers. To "strip" an image, users simply upload a photo of a target to the bot and receive the processed image after a short generation process.

Our investigation of this bot and its affiliated channels revealed several key findings:

- Approximately 104,852 women have been targeted and had their personal "stripped" images shared publicly as of the end of July 2020. The number of these images grew by 198% in the last 3 months.

- Self-reporting by the bot's users indicated that 70% of targets are private individuals whose photos are either taken from social media or private material.

- A limited number of bot-generated images shared publicly across affiliated channels featured targets who appeared to be underage.

- The bot and its affiliated channels have attracted approximately 101,080 members worldwide, with 70% coming from Russia and ex-USSR countries.

- The bot received significant advertising via the Russian social media website VK, which itself features related activity across 380 pages.

These findings also allude to broader threats presented by the bot. Specifically, individuals' "stripped" images can be shared in private or public channels beyond Telegram as part of public shaming or extortion-based attacks. Given the sensitive nature of this investigation, we have omitted key information to protect victims and avoid publicizing identifying information for the bot and its surrounding ecosystem. All sensitive data discovered during the investigation detailed in this report has been disclosed with Telegram, VK, and relevant law enforcement authorities. We have received no response from Telegram or VK at the time of this report's publication.

Bot Functionality and User Base

How does it work?

The core functionality of the bot is likely provided by an open-source version of DeepNude software (see DeepNude databox on page 6). DeepNude uses deep learning, specifically generative adversarial networks (GANs), to "strip" images of clothed women by synthetically generating a realistic approximation of their intimate body parts. The latest versions of the software use an implementation of pix2pix GANs that learns to progressively select the clothes to be removed, mark the points representing the anatomical body parts, and synthesize those body parts in the final image.

Once trained on a set of images of clothed and naked women, the software can be used indefinitely to "strip" photos of previously unseen targets with no need for further fine-tuning. While running the software efficiently would normally require users to have access to a computer with a graphics processing unit (GPU), the software embedded in the Telegram bot is powered by external servers. This removes the processing restriction for users, most notably on smartphones, and significantly lowers the barrier of use compared to predecessors of the technology.

Operating the bot

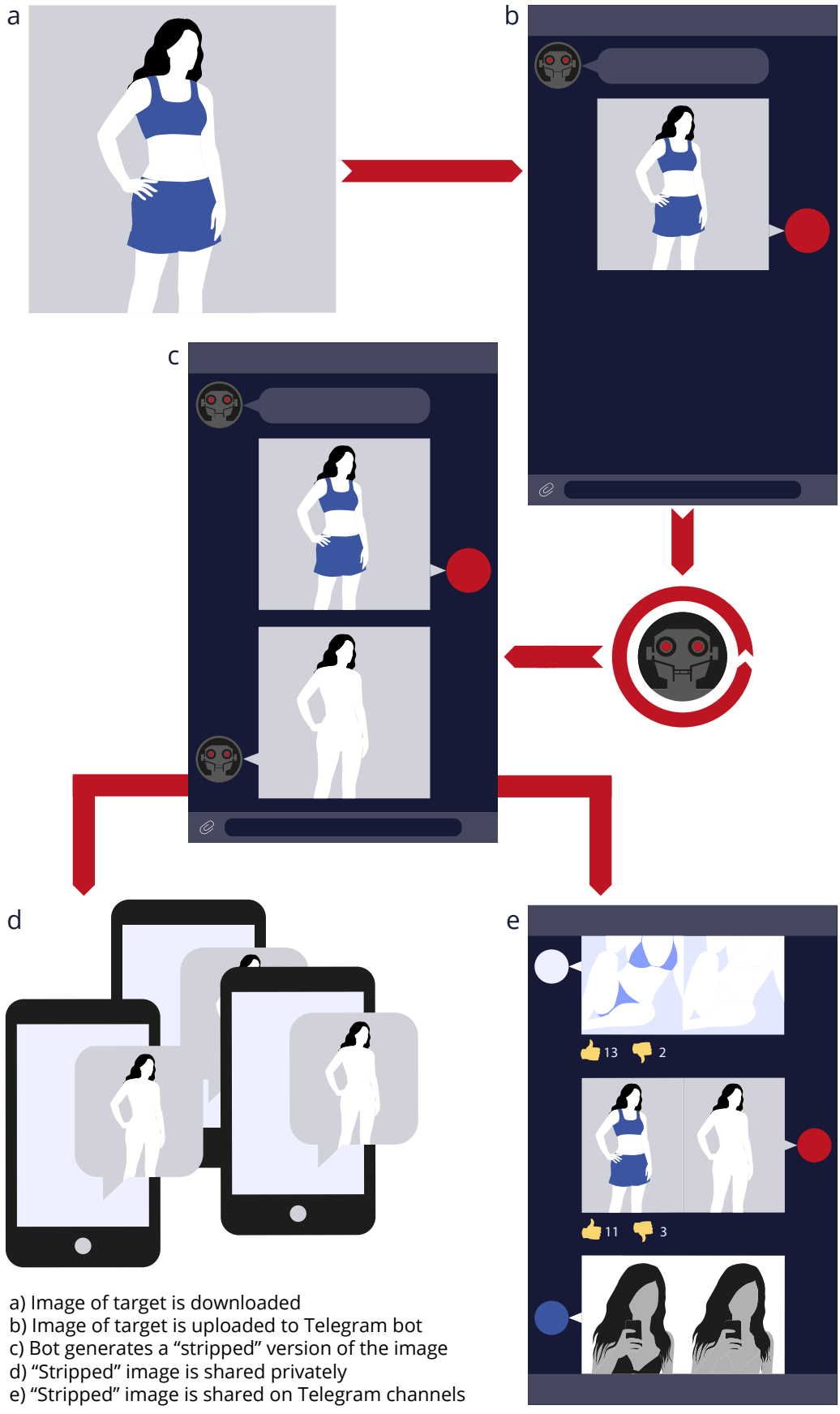
To generate a "stripped" image of a target, users upload a source image to the bot via an interface resembling a standard instant messaging app. The image is then automatically processed by the bot after a short waiting period, with the synthetic "stripped" image delivered to the user for download or forwarding within Telegram. The bot can only successfully perform this process on images of women.

While the bot is free to use, users can pay for "premium coins" that remove watermarks from generated images and skip the free user processing queue. The base price is 12 coins for 100 roubles (approx $1.50), with the price per coin dropping on larger bulk purchases.

At the time of this report's analysis, Telegram was banned in several countries, including Russia, China, and Iran. Posts made by the bot's creator encouraged users from these countries to use a VPN via Germany to circumvent the ban, while iOS Telegram users were similarly required to switch off Telegram's "safety mode" to gain access.

Affiliated channels & members

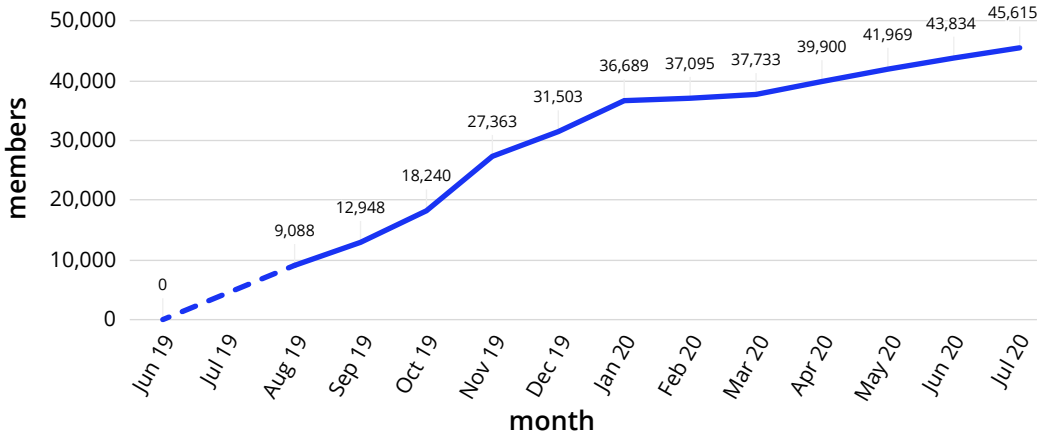
The bot's surrounding ecosystem of 7 affiliated Telegram channels had attracted a combined 103,585 members by the end of July 2020. While this figure does not account for the likelihood that many members are part of multiple channels, the 'central hub' channel alone attracted 45,615 unique members.



a) Image of target is downloaded
b) Image of target is uploaded to Telegram bot
c) Bot generates a “stripped” version of the image
d) “Stripped” image is shared privately
e) “Stripped” image is shared on Telegram channels

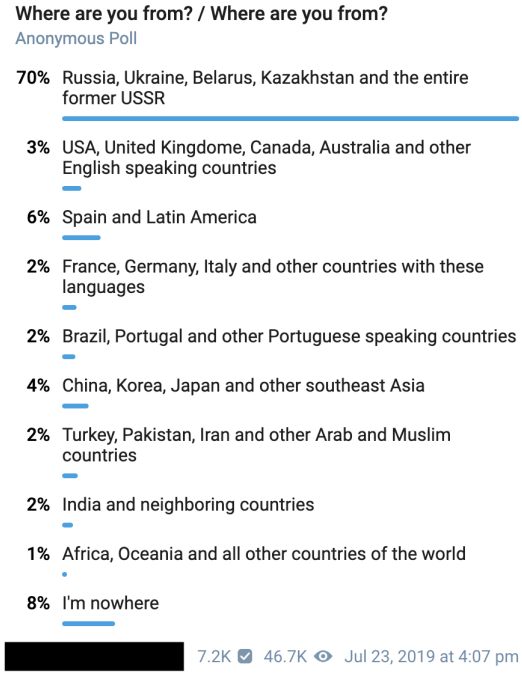
These channels provide a range of community functions including chat groups, technical support, and image sharing. Since the bot first launched on July 11th 2019, there has been a significant increase in the number of members across the ecosystem. This growth is best captured by the increase in members subscribed to the bot's 'central hub' channel.

Number of bot ecosystem "central hub" channel members



User demographics

A poll on the geographic location of over 7,200 of the bot's users indicates that 70% are from Russia or countries formerly of the USSR. Most other global regions are represented by users, although at significantly lower proportions.

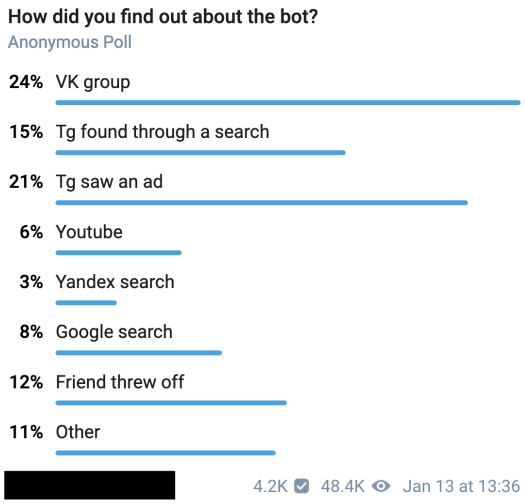


The poll featured both English and Russian (translated to English in the above image) options. The poll was originally posted on July 23rd 2019 in the bot's 'central hub' channel, but remained open for users to complete at the time of our investigation.

VK advertisement

Aside from Telegram, the bot was found to have a presence on VK, the largest social media platform in Russia. This takes the form of a page providing direct links to the bot and its affiliated channels, while also posting screenshots demonstrating its ease of access and use. Several VK accounts posting adverts for the bot were also identified.

The poll below illustrates that VK has played a significant role in attracting new users. It also indicates that cross-promotion and advertising of the bot with other Telegram channels has directly contributed to its growing user base.



The poll (translated from Russian to English in the above image) was originally posted on January 13th 2020 in the bot's 'central hub' channel, but remained open for users to complete at the time of our investigation.

Further investigation found that the bot fits into a broader pattern of deepfake activity on VK. Over 380 VK pages were found to be dedicated to the creation and sharing of explicit deepfake images, with many claiming to offer similar automated bots to the one found on Telegram. Most of these bots had identical user interfaces and payment schemes, with no free access granted.

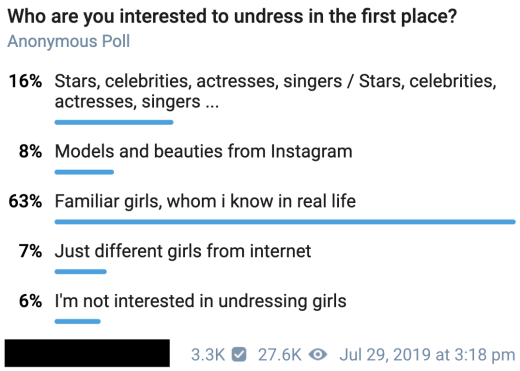
DeepNude

On June 23rd 2019, DeepNude launched as a website offering Windows and Linux apps that used deep learning to "strip" images of women. Overwhelmed by over 500,000 visitors and 95,000 downloads, the website was taken offline by the owners on June 28th 2019. On July 19th 2019, the creators sold the DeepNude licence on an online marketplace to an anonymous buyer for $30,000. The software has since been reverse engineered and can be found in enhanced forms on open source repositories and torrenting websites.

Targets

Demographics of targets

At the time of reporting, the bot could only successfully strip images of women. A user poll posted on the ecosystem's 'central hub' channel explicitly indicated that the majority (70%) of users' motivation for using the bot was to target private individuals. In sharp contrast to previous Sensity research on deepfakes (see Deepfake videos databox below on page 8), only 16% of users indicated that they were using the bot to target celebrities.



The poll featured both English and Russian (translated to English in the above image) options. The poll was originally posted on July 23rd 2019 in the bot's 'central hub' channel, but remained open for users to complete at the time of our investigation.

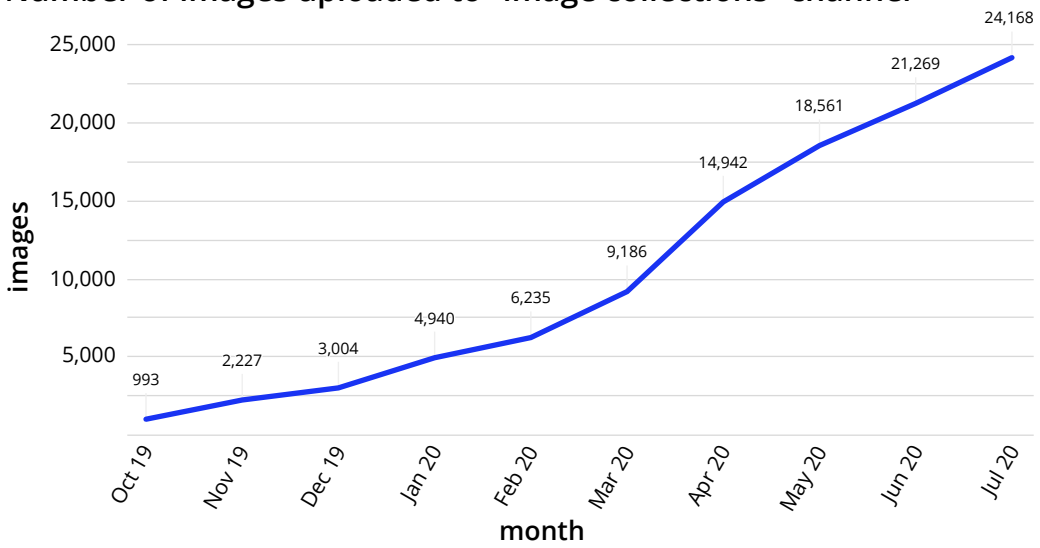
These results were largely confirmed through analysis of the images shared on the affiliated image sharing channels. Most of the original images appeared to be taken from social media pages or directly from private communication, with the individuals likely unaware that they had been targeted. While the majority of these targets were private individuals, we additionally identified a significant number of social media influencers, gaming streamers, and high profile celebrities in the entertainment industry. A limited number of images also appeared to feature underage targets, suggesting that some users were primarily using the bot to generate and share pedophilic content.

Further analysis of these images identified targets from a wide range of countries including Argentina, Italy, Russia, and the US, despite the bot's majority Russian user base.

The number of targeted individuals

The two most active channels affiliated with the bot focus on the public sharing of "stripped" images. A total of 104,852 of these photos had been publicly shared across these channels at the end of July 2020; the graph below illustrates the significant growth in the number of images shared on the "image collections" channel since its creation. This provides an estimate of the total number of individuals targeted since the creation of this community. However, the actual number is likely much higher, given that the proportion of user-generated images that have not been publicly shared is unknown.

Number of images uploaded to "image collections" channel



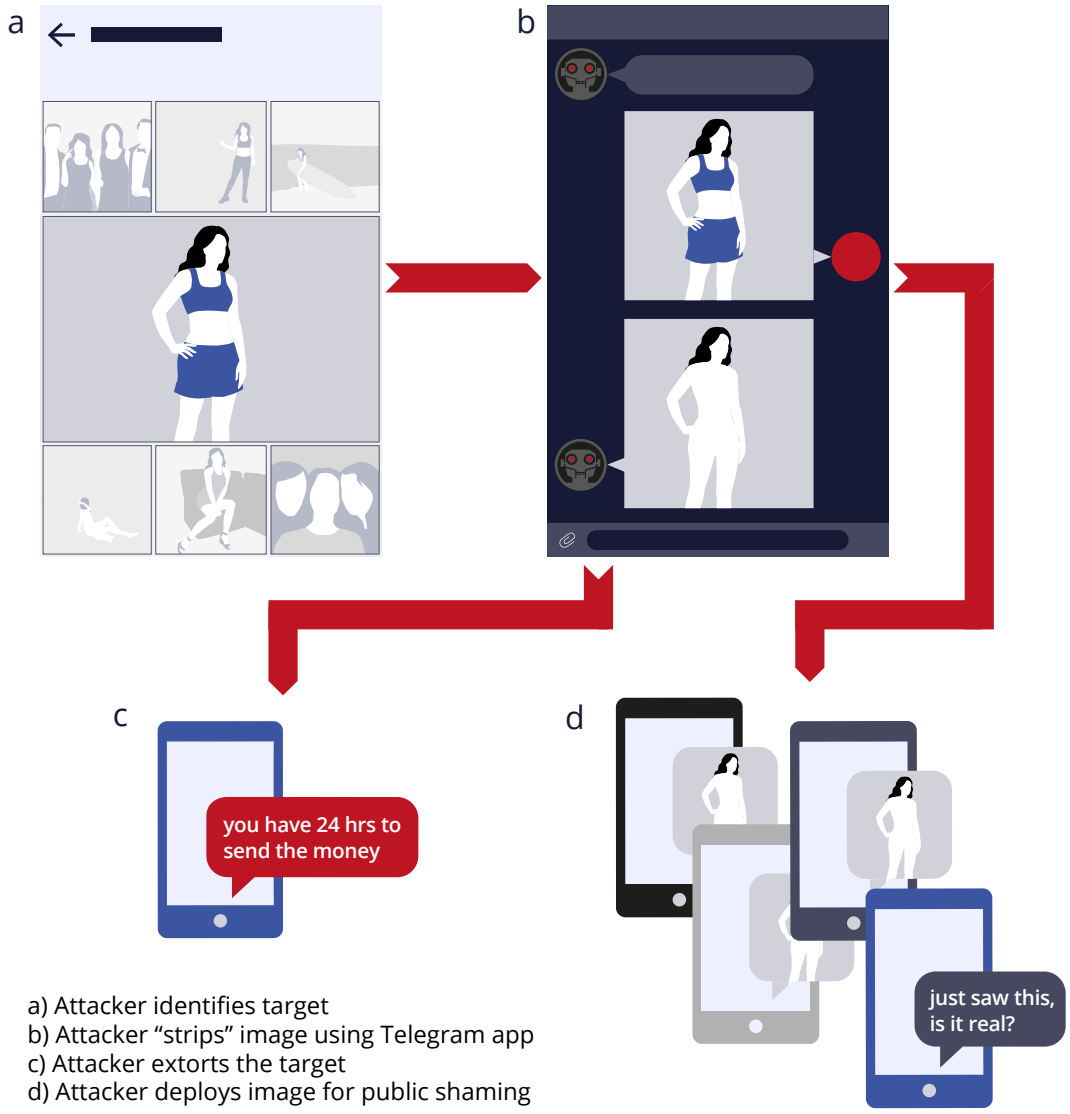
Beyond the bot: public shaming & extortion

It is reasonable to assume that many of the bot's users are primarily interested in consuming deepfake pornography, which is known from prior Sensity research (see Deepfake videos databox below) to be the most popular use for deepfake videos. However, bot-generated images can also be explicitly weaponized for the purposes of public shaming or extortion-based attacks. Based on analysis of the bot's affiliated channels, it is unclear to what extent bot-generated images are being used in this way on platforms external to Telegram.

Deepfake videos

Our 2019 *State of Deepfakes* report found that 96% of deepfake videos online are pornographic, and that the vast majority of these videos targeted high profile celebrities. The tools used to create these "faceswap" deepfake videos still require moderate programming skills and a GPU to operate, making them significantly less accessible than the bot identified in this report.

The diagram below outlines how these attacks could be orchestrated. Each scenario involves an attacker obtaining images of a target from public social media galleries, or private photos exchanged on messaging platforms or dating apps.



a) Attacker identifies target
b) Attacker "strips" image using Telegram app
c) Attacker extorts the target
d) Attacker deploys image for public shaming

After "stripping" the target's image using the bot, the attacker deploys it to publicly shame the target, sharing the image openly on social media, or sending it privately to the target's relatives, friends, and acquaintances. Alternatively, the attacker extorts the target by threatening the publication of the "stripped" image online unless a sum of money is transferred by a certain time. The personal information and images of targets could also be sold to other malicious actors on underground forums and marketplaces.

@sensityai
/sensityai
/sensityai
info@sensity.ai
www.sensity.ai