VirtuousAI

Automatic Content Moderation

Virtuous AI Models

INSIDE

1. Introduction

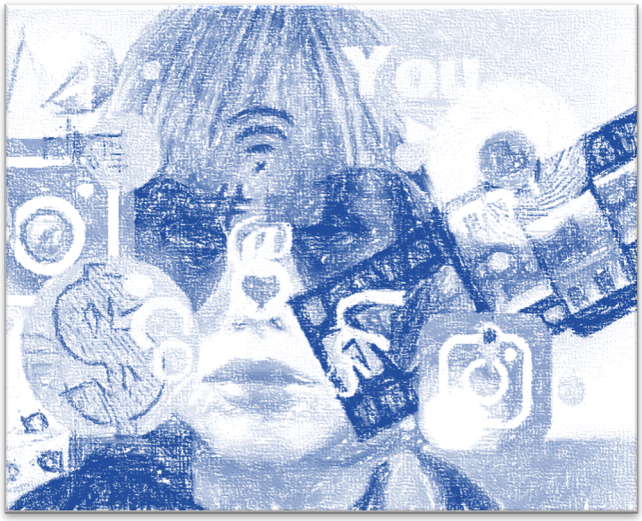
Effective content moderation requires the combined use of various algorithms designed to deal with the different types of content your customers create to engage with each other on your platform. Text messaging is one of the most popular forms of communication, and the use of various types of emojis and gifs help to enhance the experience in the chat.

Users also enjoy adding voice to their text messages, whether short audio clips or full-fledged calls. These give them a more immediate way to engage with each other and with your platform's content. The ability to hear each other and express themselves vocally adds a livelier and more human dimension to the platform experience.

Video adds an even more stimulating layer of interaction to your platform. Because it engages users both visually and aurally, it brings them even further into the experience. Video is a powerful medium because, in

Content Moderation

Models



addition to the content you place on your site, users themselves can create videos for upload and generate responses by other members of the audience. These responses may take the form of text, audio, or video—leading to a further explosion of content. Video chats *between* users is another possible way of of allowing them to engage each other in a social experience that keeps them coming back to your site. You may also host live streams or allow your users to livestream onto your platform.

Rich and engaging content are the holy grail for content hosting platforms, and to maintain it as a positive and prevent it from becoming a liability, automatic moderation is indispensable. Your platform's social networking function can easily be degraded by hate speech, unsavory language, or bullying. Its reputation can become tainted by unwanted violent or pornographic images. The power of video can be hijacked for illicit, violent, or even terrorist purposes.

Often, it is not enough to simply remove bad content after it has been posted, as this gives time for someone to take a screenshot, share a post, or save a video *with your brand's logo or name attached*. Most often, the only way to prevent material like this from permanently damaging the reputation of your platform is to guard pre-emptively against them by employing *automatic* content moderation methods.

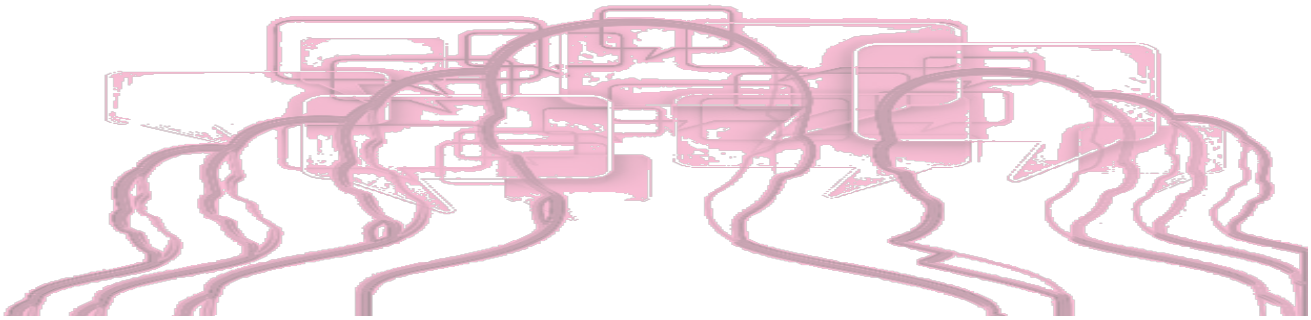
For all these types of content, Virtuous AI provides intelligent, state-of-the-art moderation support to safeguard your users and your brand's reputation. Our machine-learning models continuously work not only to identify and filter out unwanted content, but also to learn the ways in which users might try to "beat" the system. This means that our various models evolve along with your users' methods to keep your moderation processes up to date at all times. Below, we discuss the text, image, video, advertising, and reputation cleansing models we offer.

info@virtuousai.com

2. Text Moderation

Protecting your brand also means moderating the content of the text-based messages that are sent on your platform. Natural language processing and machine learning have made this process easier, and such automated methods are indispensable. In an age where hundreds of millions of persons send billions of text messages in a very short period—and when senders use very sophisticated methods to bypass mechanisms that filter inappropriate materials—your brand can remain truly safe only with the advanced processing power provided through Virtuous AI's ability to handle big data.

Intelligent machine-learning algorithms filter out profane, obscene, offensive, and inappropriate messages before they get posted.



Automatic text moderation assists you in filtering out profane, obscene, offensive, and inappropriate messages *prior* to their being posted. This means your viewing public will neither see nor be affected by any messages sent by malicious users, including trolls, spammers, or unauthorized advertisers. Automatic text moderation benefits your site and its users in a variety of ways, as it can be used to:

- Block inappropriate textual, image, or video content by immediately identifying problematic instances and preventing them from being posted to your site
- Reduce the workload for human moderators by filtering out the obviously bad content, letting through the good content, and sending only questionable material to human moderators for double checking.
- Calculate the likelihood of certain users to post objectionable content based on their past activity.
- Assess the preferences of your users to match them with content created by other members of the platform. You can thus design customized user experiences using filtering methods that learn what users like and feed them more of it while cutting out content they might find offensive or simply dislike.

Below, we consider some examples of problematic types of content that require text moderation.

info@virtuousai.com

Leetspeak

Leetspeak is the act of replacing certain characters in words with alternate characters to reduce its offensiveness or evade filters. The result is that the words look similar to the original ones, but technically they are transformed into a string of characters not considered offensive by filtering mechanisms. Users often use non-alphabetic characters to replace letters of the alphabet in order to trick automatic filters that—traditionally—have worked better with alphabetic characters. For instance, the letter *E* may be replaced with the numeral 3, the letter *A* with the numeral 4 and the letter *I* with the numeral 1.

However, leetspeak can get far more sophisticated, since users also replace letters with non-alphanumeric characters, such as slashes, asterisks, brackets, parentheses, and exotic characters from languages outside their primary one (see the section on "Special Characters" below). Thus, in English, users might interpolate Arabic or Hebrew characters that are similar to English letters but would be overlooked by traditional automatic filters but understood by a human. The following examples will give a good idea of the clever ways in which persons try to evade detection.

Leetspeak	1337$p34k
Come over tomorrow	C0m3 0v3r 2m0rr0w
knives	|<n!\/e$
weapon	\/\/34p0n
Dive	|)!\/e
Huck	]-[(_)© |<
Finn	|=][||||
out there	<>|_|'][' ']['][']-[ëЯë

These examples become increasingly difficult to understand and give an idea of the lengths to which users might go in order to escape detection. Sophisticated filters such as those developed by Virtuous AI go beyond the mere swapping out of letters for numbers (2nd example above) or letters for characters; they catch even those methods of combining and overlapping exotic symbols to form single or multiple characters. Our intelligent models are able to

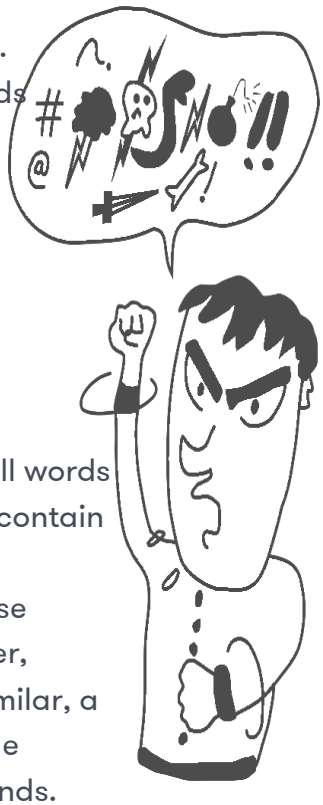
info@virtuousai.com

differentiate between authentic uses of symbols (such as using the dollar sign for actually representing money) and deliberately deceptive uses. When we go beyond mere character replacement, we provide a holistic assessment of user content and offer the most comprehensive text moderation to protect your company or brand's reputation.

The following are some other strategies related to leetspeak, which automatic content moderation is equipped to handle.

Deliberate Misspellings

Related to leetspeak is some users' attempt to use deliberate misspellings to evade detection. By replacing one word with another that sounds similar, such users might succeed in sending offensive messages or even ones that mention illegal contraband. Sophisticated automated moderation tools are also equipped to detect such evasions.



Special Characters

The use of special or foreign characters to spell words is also related to leetspeak. These letters may contain diacritic marks (accents, tildes, umlauts) that technically make the letters different from those usually contained in a given alphabet. However, because the letters themselves are still very similar, a speaker of the original language has no trouble reading and understanding what the user intends.

Profanity & Obscenity

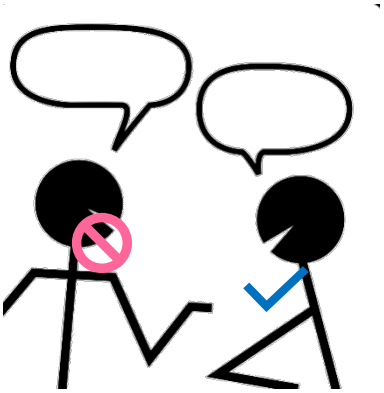
Profanity filtering is indispensable in informal forums such as chat rooms where people generally feel comfortable and free expressing themselves. While that freedom should be encouraged in an effort to get a better understanding of one's audience and to create a welcoming and open atmosphere, care must also be taken to keep that space welcoming for *all* users. That freedom might cause some users to generate lewd or profane content that might be offensive to others.

info@virtuousai.com

Some users may even try to begin with a username your community might find offensive, so content moderation should begin at the very start of the sign-up process. Thus, in addition to clearly providing them with guidelines about how to use the open forum on your platform, it is also important to take comprehensive steps to prevent—or expeditiously remove—profanity and also to deter users from using it. Very often profanity involves leetspeak, since users know certain words are generally not allowed in forums. However, some users will commit outright violations.

Outright Violations

Not all violations are indirect or hidden. Some are blatant violations of your community guidelines. Automatic content moderation allows you to remove all commonly understood profanity or obscenity from your site. Beginning with username-creation monitoring, it can help you establish your brand's standards at the very beginning of your interaction with each user. It also allows you customize your list of approved and prohibited content by adding blacklists and whitelists to your models.



Blacklisted Words

Words that are blacklisted have been added to the list of prohibited content. Your filter will ensure that those words do not get posted in any of the forums on your platform, including live chats, message boards, blogs, comments, or any other area in which users can publish text.

Whitelisted Words

Certain words that are normally filtered by default may not be considered offensive on your platform according to its ethos and as represented by your community guidelines. In that case, you may choose to allow your members to use these words by whitelisting them. Filters will then allow those words to be a part of the conversations that occur on your platform.

info@virtuousai.com

SENSITIVE INFORMATION	
Name (First and Last)	Social Security Number
Mailing Address	National Insurance Number
Zip Code	Credit Card Number
Email address	Image of Photo ID
Phone Number	Banking Information

Personally Identifiable Information
Users place themselves at risk when they post information that could personally identify them, such as their name, address, birthdate, social security number, etc. Revealing this information exposes them to such harms as identity theft or even physical harm by a person who may use that information to find their physical locations. This is an especially likely risk for children, who are often unaware of the harms that can come to them and methods of avoiding them. Automatic content moderation is crucial for preventing your users from inadvertently revealing sensitive information about themselves or others.

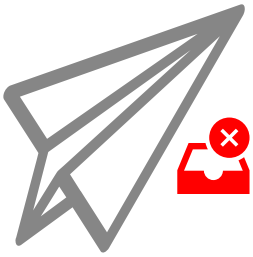
Good filters will do more than simply recognize and delete inappropriate content. They will allow you to employ a gradual system that deters users from posting it in the first place. It will also create an individual profile for each user and monitor created content throughout that user's lifetime on your platform in order to devise predictive moderation measures. Here are some strategies automatic content moderation employs to train users to act as good citizens to your platform:

Think Twice
This takes the form of a message sent to the user after he/she types a message but before it is sent. It causes users to think twice about what they are about to post. Example: "Some users might find this message offensive. Are you sure you want to send it?" Studies show that 93% of young people change their minds about sending a comment if warned it might be perceived as offensive (ReThink). Think twice is, therefore, an effective method.

info@virtuousai.com

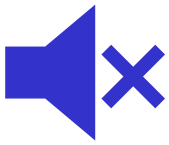
Content Moderation

Models



Faux Send

The faux send allows a user to believe his/her message has been sent inside a chat or forum but shields the other users from its malicious message by hiding it from them. This strategy is most effective in large forums for neutralizing the effect of malicious messages and avoiding any fallout that would have occurred (such as arguments, bad publicity, or user boycotts) as a result.



Mute

Muting prevents an offensive user from sending messages in the forum for a temporary period usually measured in seconds or minutes. Usually, the user is made aware that s/he is being muted, why, and for how long.

Warning

Users that have violated community guidelines while engaging in text communication are issued warnings to let them know that any future violations will result in punitive measures, such as a "kick" (removal from the current forum), suspension, or a permanent ban from the platform if their infractions are serious enough.

Effective text moderation requires comprehensive measures that are carried out before, during, and after users post their content. This can only be fully provided by an automatic content moderation regime that uses the power of artificial intelligence and machine learning to monitor messages and respond decisively to them to preserve the safety of your platform for all users.

Suspend

A user might be suspended after a series of violations tabulated automatically by machine learning algorithms that keep track of users' activities. This serious penalty goes beyond just removing the user's ability to send messages in the forum. Instead, it temporarily removes the user's ability to sign into or use the account.

info@virtuousai.com

3. Image & Video Moderation

Visual images pack a heavy punch when it comes to advertising and brand recognition. They have a significant impact on the public, and this is one of the reasons most effective marketing strategies include images. A company's logo, for instance, is an easily recognizable image. Images are versatile. Video does everything that images to but also engages users aural functions, bringing them more fully into engagement with your content. Together, images and video are capable of persuading an audience on an emotional level, and they can build customer trust and brand credibility by giving valuable technical insight into the way a company's product works. But precisely *because* of their impact, negative images and videos can swiftly have a harmful effect on a brand's image, quickly undoing much of the hard work that has gone into building it.

Monitoring and moderating image/video content involves quickly identifying and removing such files—ones that might be offensive to users or contain explicitly violent or sexual expressions. Effective moderation vastly reduces the likelihood of these objectionable images entering your brand's public image via your social media feed or other online presence. Automatic image and video moderation tools protect your brand and prevent your customers and the public from associating your brand with objectionable content a thoughtless user might randomly generate. The following are various types of image/video moderation tools available in the industry:

Nudity Detection

For most companies, nude images do not complement their ethos or vision, and for brands whose products are family oriented, even partial nudity could cause its customers to think twice about the messages it sends to the public. The nudity detection model Virtuous AI uses in image moderation allows companies to dictate how strictly our algorithms must act when detecting nudity. You may have it detect partial nudity—such as women in bathing suits or lingerie, women with cleavage, small shorts or short skirts, or bare-chested men. Full nudity would detect such images as bare breasts or genitals. This can be done for both live and animated images. Below is a more detailed breakdown of the explicit versus suggestive nudity categories:

info@virtuousai.com

NUDITY	
Explicit Content	**Suggestive Content**
• Bare Breasts on a Female • Bare Female Lower Body (Genital Area showing) • Male Genitals • Sexual Activity • Illustrations of Sexual Activity • Sex Objects or Toys	• Women in Lingerie, Underwear or Swimsuit • Men in Underwear or Swimsuit • Shirtless Males • Cleavage • Women in Small Shorts/Short Skirts • See-through Garments

Facial Detection

Face detection is useful for determining image and video content and gives crucial information such as whether a baby or minor might be displayed exploitatively in an image. It also enables recognition of emotions and allows for comparison between faces. This allows image moderation to provide services to detect vulnerable persons, known criminals or scammers, identity theft or even image spamming. Facial comparison is a related application that enables matching with other images, either of known offenders or of other users, to identify possible profile duplication or identity theft. By performing comparisons between uploaded files and those stored in an image/video database, facial detection and comparison algorithms can identify and compare features with a high degree of accuracy, even after persons have aged.

Minors (babies, children, teens)

All contexts in which images or video footage of babies, children, and minors can be used in online content are regulated by the laws of various countries. Especially when user-generated content has not been vetted by your company's processes, if it contains images of children, every effort must be made to ensure that these do not violate laws or the principles of your online community—and those principles should align with the law. The Virtuous AI model for detection of minors uses facial recognition technology to classify facial features and home in on those closely related to children's features. The model tells you how likely images are to have a child in the picture or video and gives you the option to remove the file, blur the face of the child, or even prevent the content from being uploaded.

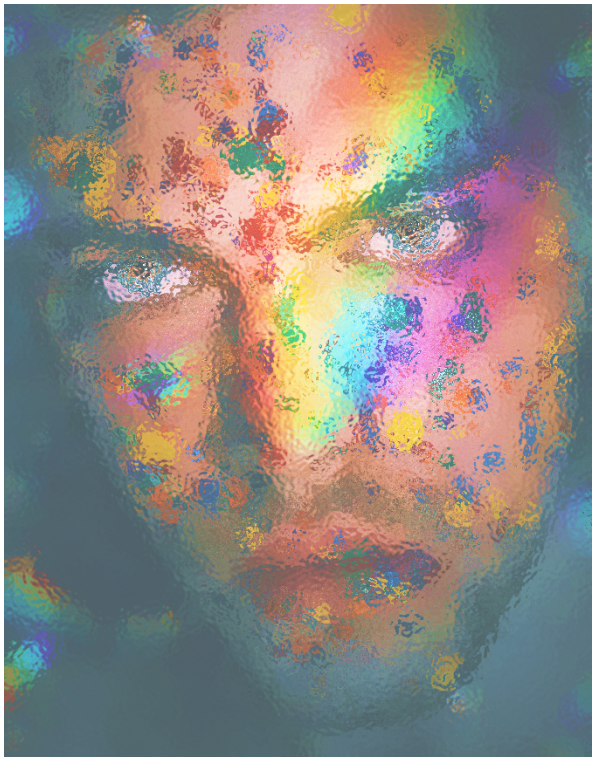
info@virtuousai.com

Face Obscuration Devices (masks, sunglasses)

Images that contain sunglasses or masks may hamper humans and algorithms' ability to estimate the age and identity of the wearer. Image and video moderation also give you the option to detect and remove files in which persons are wearing sunglasses, masks or other objects, in order to prevent users from sneaking inappropriate content onto your platform.

Facial Recognition

Virtuous AI provides you with information about the location of faces within the images and videos uploaded to your platform. This includes the identification of where landmark facial features are situated in order not just to recognize whether the face is human but also predict with a reported level of accuracy (confidence) its correspondence to another facial image. It is also possible to analyze video /images for emotions based on these landmark features—position and orientation of the mouth, eyes, cheeks—to tell, for instance, how likely it is that the person in the footage appears happy or sad.



Faces from both images and videos can be saved and kept in a compendium for future reference. For instance, if an account known to be conducting fraud uses images and video, once those files have been scanned, the images can be saved and used to detect any other instances in which similar images are used. Such instances can then be flagged as having a high potential for fraud or other malicious content and the users investigated. This can occur before the content is even posted. Our service allows you to create one of these collections of images and then search through them using processing specifically for image or for video. This works asynchronously and allows you to compare the stored images with those located within all the videos uploaded to your platform at any time. Videos may also be used for authentication purposes if, for instance, the collection of images holds the faces of individuals authorized to access certain files, areas, or databases on your platform.

info@virtuousai.com

Content Moderation

Models

PROHIBITED IMAGES
Weapons
Pistols
Revolvers
Machine guns
Knives
Cleavers
Chainsaws
Razor blades
Swords
Scimitars/Machetes
Substances
Pills
Powder
Cigarettes
Joints
Vape
Beer Glasses
Cocktails
Wine
Alcohol Bottles

Violent Content and Illegal Substances

Depending on your brand, certain types of violent content may or may not be suitable for your audience. Content (such as terrorist images or video) are not even permitted by law for reasons related to the security of citizens in various countries. Similarly, alcoholic beverages, tobacco, and other substances such as prescription and illegal drugs are highly regulated. Any posting of such images (or video containing these substances) may send an inappropriate message to vulnerable users. It may also cause your brand to breach regulations and open you up to lawsuits or other legal action by regulatory institutions. Automatic content moderation allows you to quickly identify drugs, alcohol, or weapons in images and video and quickly flag them for immediate takedown or review. This allows you to keep your online forum a safe space for users to engage with each other and with your brand.

Hate Images and Gestures

Rather than bring users together, hate images, emojis, and gestures divide a community, and any brand that allows user-generated content to freely contain these types of images will lose credibility with the public and risk legal action being taken against them. Image and video content moderation allow you to be vigilant about detecting and removing hate images even before they get posted to your online forums.

The model detects images such as offensive flags, fascist or white supremacist symbols, the middle finger, or other generally offensive insignia. Algorithms contain an extensive list of signs that are offensive around the world so that you can select the ones most suitable for detection and removal from your site.

info@virtuousai.com

Content Moderation

Models

Frauds and Scams

A major challenge site owners face when creating an online community for its users is the creation of fake profiles and identities. Malicious persons aiming to defraud users of their money and assets prowl the internet looking for social sites where they can carry out their scams. Using such parameters as keywords, posting frequency, location discrepancies, and image recognition, moderation techniques can flag suspicious behavior as fraudulent and put a halt on any user-interaction around such posts before any damage is done. This kind of detection is good for marketplaces but also in phishing situations where scammers attempt to get personal information from other users in order to carry out their scams.

Personal Images and Video Content

Owners of web-based platforms have a responsibility to monitor personal images and video content uploaded to forums, particularly when those files are posted by minors. Personal images must be approved by parents or guardians of children under 13 years, according to the laws of the US, UK, EU, and other countries.

info@virtuousai.com

Content Moderation

Models

Artificial Images

Customers do not want to believe that a brand might be posting artificial images because that behavior appears deceptive. If an online user posts artificial images in a forum hosted on your company's website or app, its viewers may attribute the post to the company itself and lose confidence in the brand. Our moderation tool that identifies image type enables detection of artificial images that have been air brushed or altered so that it appears to contain something that wasn't in the original image. The tool also detects when objects might have been removed from images. Such images can be identified and removed or flagged for review.

Image & Video Quality

Because of the impact images and videos can have, brands always want to put their best ones forward and use every opportunity to create a positive impression with clarity and incisiveness. An image or video may have perfectly appropriate content but still be "bad" because it is difficult to see. Automatic moderation will reduce (or eliminate) the number of low-quality files that surface to users. You can identify the threshold of image or video quality that you want all content to surpass in order to be uploaded to your online community. Parameters for detecting image quality include sharpness, contrast, brightness, and pixel dimensions. For video, quality is measured by the level of definition: high definition begins as 720p. Quality moderation and control allows your brand to make the most of every marketing opportunity.

info@virtuousai.com

4. Ad Moderation

Since your website is the online representation of your company, it is imperative that things run smoothly whenever users visit. It publicizes your brand's values and broadcasts all the information you would like users to know about your company. However, many obstacles can prevent your users from actually getting the experience you'd like them to have on your online platform. This might include spammers—and it might also include the very ads that you use to monetize your website's content. Since you most likely do not produce the ads you allow on your website, in order to maintain control over your user experience, it is important that you have an ability to moderate the activity of the advertisements that are on your site.

Advertisements are a very important source of revenue for many website owners, and it is important to create an optimal balance between the ads that bring in that revenue and the user experience that keeps your audience returning. While 83% of users would prefer being able to block advertisements, this is not necessarily because they hate ads. Rather, they dislike the disruption many ads bring to the experience of browsing a website. In fact, pop-up ads on websites are the ones most hated by users. Disruptive ads degrade user loyalty and defeat the web platform's purpose, which is to retain customers. Producers and marketers are not always mindful of user experience when creating or placing their advertisements, so even after you have subscribed to an ad revenue service, it is still important to monitor the behavior of the ads they send to your website. This includes their content as well as the way they interact with users to enhance or detract from their experience.

info@virtuousai.com

Malware

Many ads pop onto the page as users are browsing. Unfortunately, many malware events are also marked by pop-up activity, and it is important that a distinction be made between these two types of experiences before it gets to the user. Ad moderation prevents the malware event from getting through by assessing pop-ups before they appear on the screen and flagging suspicious content. If a user got

attacked by malware while on your platform, such an experience would be so negative that the user's loyalty may be effectively nullified. If it occurs with several users, the platform might fail to recover from it. Ad moderation software will provide the ability to test for the presence of unusual activity that might indicate the presence of malicious content in the ad. It will then provide a confidence score that allows you to make a decision about whether to allow future ads from that origin on your platform.

info@virtuousai.com

Content Blocking Popups

Some ads are so obtrusive that they block the content you would like to present to your customers, often covering the entire page and preventing your customers from consuming your content. Pop-up ads are the main culprit, and not only do these ads obscure content but they also often disguise the method of closing the ad in such a way that users inadvertently open the ad when trying to escape it. This takes the user off your site without their consent. Automatic content moderation can act with the speed required to assess the ad before it comes through to your user and allow you to table the ad for review. You may then use your review decision to gauge your site's response to similar ads in the future.

Autoplaying Ads with Video & Sound

Ads that play video automatically do more than just annoy users. They may cause the user's online experience to disrupt others around them, thereby embarrassing or even causing trouble for them. Even if the content itself is wholesome, loud auto-playing ads may qualify as NSFW—not safe for work. Many times, these ads are hidden far down the page, so that it is difficult for the user to locate and close the ad. Unlike quick-to-play, quick-to-load ads, these ads are usually very long, and if the user is on a mobile phone, the video may rapidly consume data. Automatic content moderation is able to detect such as and gives you a number of options to stop, mute, or curtail their access to bandwidth if, for instance, the user is one a mobile data plan rather than on wireless.

info@virtuousai.com

Sticky Ads

Some ads take their inspiration from sticky menus, which come in handy as they continue down the screen as the user reads and are always available for navigation when the time comes. The situation changes when it is an ad that becomes sticky. Users are unable to scroll or otherwise navigate your site without the sticky ad pursuing them. The creators of such ads anticipate the user's annoyance and have designed their product to automatically prevent any actions the user might take to improve the experience of your platform. Like a heat-seeking missile, this sticky ad is built to ensure that your users are unable to take evasive action when the ads come on screen. However, on the user-side, this will simply escalate the frustration the ad's initial appearance caused. Some users may decide that the only way to escape the ad is to leave your platform. Software that automatically moderates ads will be able to detect this type of foul-playing advertising strategy and allow you to take steps to reduce your users' discomfort.

Pre- and Postitial Ads

Prestitial and postitial ads are, respectively, the types of ads that appear before and after text, video, audio, or gaming content begins or ends on your platform. The more common of these is prestitial, which appear before content. They sometimes allow users to turn them off, but usually this option does not arrive until after a countdown has taken place. You may want to moderate how difficult it is for users to turn these ads off by designating a maximum countdown length for these ads. Ad moderating software allows you to curate your user-experience in this way to create the best possible environment while maximizing the possibility of ad revenue.

info@virtuousai.com

Postitial ads appear at the end of the user's consumption of content—such as at the end of a video or after scrolling to the end of a blog post. Such ads are perhaps a bit more dangerous to your viewership or audience retention because once viewers have consumed the content they initially requested, the appearance of an ad might simply cause them to close the window on the entire platform. This is especially likely if the ad comes up before the user has even had a chance to see what other content might be upcoming. Therefore, you might want to regulate how postitial ads occur on your site, and ad moderating software can support your decision-making in that area.

Screen-Covering Ads
Some ads cover up the entire screen and make it impossible to view the content for which users came to the page in the first place. Other ads cover only part of the screen but still conceal enough of the content to be obtrusive and make it difficult to actually consume it. If ads cover more than 30% of the screen, the experience will most likely be bad for the users. This is especially true on mobile devices already having a limited display area. Users may simply be unable to view the actual content even when the ad is at its minimum size on the screen. With ad moderation software, such obtrusive ads can be detected and dealt with in the way you choose.

Screen-Covering Ads

Some ads cover up the entire screen and make it impossible for users to view the content for which they came to the page in the first place. Other ads cover only part of the screen but still conceal enough of the content to be obtrusive and make it difficult to actually consume it. If ads cover more than 30% of the screen, the experience will most likely be bad for the users. This is especially true on mobile devices, which already have a limited display area. Users may simply be unable to view the actual content even when the ad is at its minimum size on the screen. Ad moderation software puts you in control by giving you the chance to detect and deal with such ads in the way you see fit.

Content and Brand Ethos

Your brand reputation can suffer severe losses as a result of the advertising content displayed on your website. Since this content is tailored to each user, the type of ad being shown on your site might be different for the entire number of users present at any given time. Since you place a high value on your brand's reputation, you should also exert a certain amount of control over the ads that appear on your platform. Image, text, and video moderation techniques discussed in this document can also be applied to ads as they populate your webpages or media apps.

Good ad moderation selects between the ads that enhance and those that disrupt, providing your user with the type that gives them a good user experience while enabling you to make good ad revenue. By flagging the latter, you may create a database of ineffective ads that can inform your media marketing or ad revenue service in their placement of advertising content on your platform. Ultimately, ad moderation allows you to maximize the appearance effective ads and minimize that of ineffective ads. It does this by creating an overall positive user experience for your audience that optimizes the balance between ads and content in a way that retains your users and maximizes your ad revenue.

The complete removal of all ads from the company's website may not be an option for many—and Virtuous AI makes this extreme case unnecessary by providing solutions that keep both ads and users on your site. While users hate ads that disrupt their online experience, many of them actually find some ads useful. Ads that are well curated and matched to the type of user browsing your site will enhance your site's overall user experience.

info@virtuousai.com

5. Models: Reputation Cleansing

As many as 86% of all persons reconsider their decision to buy a product from a company whose reputation is tainted by negative reviews. Any brand that shows any hint of a negative blemish to its reputation will risk losing its potential customers or business connections forever.

A reputation is perhaps the most valuable asset a person or company can have. Whether you're trying to cultivate a personal or company brand, the way you are perceived by the public is crucial to how well your brand will perform. Reputation cleansing is the perfect solution for any enterprise or any professional going into the business or public world. If your business has had a public relations setback or your personal brand has been smeared in the media, your chances of acquiring new clients or retaining your old ones diminish significantly. The reason for this is that one of the first things people interested in your brand or product will do is search online *specifically to gauge your reputation*. Customers and other businesses do online reputation searches in order to decide whether to associate themselves with a brand. A positive online reputation can lead a company or brand to acquire more customers, followers, and patrons. A negative reputation, on the other hand, has the effect of deterring prospective customers and even discouraging your hard-won leads just as they might be on the brink of making the decision to become a client or customer.

info@virtuousai.com

Content Moderation

Models

If your brand's reputation has a blemish, how can you turn this around in the age of information where that knowledge is right at the fingertips of anyone willing to do a simple online search? Brands have been employing reputation management as a method of (1) crafting the type of brand image they desire and (2) mitigating the effects that negative publicity has had on their image in the past. The most successful reputation management services are able to use a variety of methods to organically improve the face your brand presents to the public on the internet. Such services are also capable, if necessary, of taking drastic measures to remove or significantly reduce the impact of any negative information that exists online about your brand.

Virtuous AI offers a wide range of services for individuals, small businesses, large enterprises, including services that monitor your brand's publicity and even pre-alerts you to incidents that have the potential to cause harm to the brand's reputation.

For Individuals

Individuals may encounter attacks to their reputations in many ways online, and since each person's case is different, reputation cleansing requires a customized approach to assessing and mitigating the damage. In some cases, removing damaging content may require appeals to the various websites or platforms carrying or hosting the data. In more stubborn cases, it might require appeals to certain political rights granted by the various jurisdictions governing online activity. Reputation cleansing exhausts all possibilities—including several ingenious and intelligent methods that together work synergistically to enhance the positive influence and respectability of your image and personal brand.

For Businesses

Companies experience a lot of unfair feedback from disgruntled former employees or even as a result of stealthy defamation projects by unscrupulous competitors. Long after your brand has taken legal action against such bad actors and even won the case, the effects of their actions may remain on the internet, ranking high in search results and continuing its damage to your reputation. Cleansing your brand's reputation in such cases require the use of intelligent systems that are able to counteract the actions of search engines as well as to use the search engines' own methods to your benefit by taking steps to make the negative information less rank-worthy. The result is a much more favorable brand image when your name, brand, or logo is searched on any platform.

info@virtuousai.com

Search Engines

Reputation cleansing works on a wide selection of search engines. It goes far beyond just Google, Bing, or Yahoo to include YouTube, Facebook, Twitter, Instagram, and a host of other very influential platforms on which information might be disseminated about your brand. It is a comprehensive service that identifies your complete online image—even beyond what you have been able to discover for yourself—and then takes the necessary steps to recraft that image in a way that drastically improves the public display of your brand.

Pre-emptive Strike

Powerful proprietary algorithms that mimic SEO ranking spider bots work 24-hrs a day searching for associations to your brand. They track their movement up the ranks of the search engines long before they can become discovered by users. This means that any negative associations in the process of being created can be identified and suppressed before they have a chance to do any harm to your brand or personal image.

info@virtuousai.com

Negative Reviews

Has your brand been hurt by negative reviews from persons who simply did not understand your product? Is your business perpetually suffering because of the actions of a bad employee who has long since been fired? Significantly improving the way your company appears to other customers will efficiently catapult your business toward a higher plane of success.

New customers rely heavily on the experiences of previous customers to help them predict how well your services will satisfy them. When they see negative reviews—even if the number of them isn't overwhelming—they tend to select one of your competitors instead. This means that negative reviews for you are similar to positive reviews for your competition. Our service helps you change the narrative surrounding your company, your brand, and/or your product by streamlining the flow of satisfied customers and empowering them to write more and better reviews.

The Right to Be Forgotten

Capable image cleansing teams will explore the ins-and-outs of a concept gaining momentum in the wake of the lasting effect unwise choices or malicious actions can have on a person or brand's reputation. This concept is known as *the right to be forgotten*—which in certain situations allow you to request the removal of unflattering information from search engine results. Wherever privacy statutes provide for the takedown, re-shuffling, or favorable rearrangement of the information about you found online, a reputation management service may undertake all the procedural requirements to uphold your right to be forgotten. However, if this right is denied, you will still be able to secure an improvement in your online reputation by hiring Virtuous AI's reputation cleansing service.

info@virtuousai.com

Content Moderation

Models

Social Media Clean Up

With smart technology and powerful algorithms, you can remove content that is damaging to your brand's reputation from a variety of social media sites. Knowing where and how to remove content directly will get you a good distance, and even the sticky hard-to-erase content over which you have no control can be re-contextualized to improve the way it comes off to the public. Our algorithms can also quickly and safely remove social media ties with connections who taint your reputation by association. The negative effects of certain associations may not be obvious to you, but convoluted search engine algorithms can trace these connections and use them to associate you with situations or activities that can really harm your brand. Our algorithms can model search engines, identifying and severing the dangerous connections before they have a chance to do further damage to your reputation or that of your brand.

Negative Images

We will populate the internet with personal and brand images that report positively on you and your company's work. These will overwhelmingly dominate the search engines, crowding out the infamy of any other photos. It is proven that 92% of persons remain on the first page of the search engine results and 99% never get beyond the second page. Our algorithms are able to push negative images beyond the threshold that most users' are willing to scroll.

info@virtuousai.com

Negative Name Association

It might be the case that your name or the name of your company becomes associated with an unpleasant incident that places you in a bad light. If your brand is prominent, this information might have been shared widely or discussed heavily in social media—or even online versions of traditional media. Reputation cleansing allows you to dissociate your name or that of your brand from these negative instances by creating positive alternatives. Plus, proprietary algorithms have the power to emulate search engine algorithms and build on them to ensure that the positive associations rank far above the negatives.

info@virtuousai.com