ZEROFOX® Intelligence

Brief

Detecting and Countering Synthetic Media

B-2025-09-15a

Classification: TLP:CLEAR

Criticality: Low

Intelligence Requirements: Social Engineering, Fraud, Artificial Intelligence

September 15, 2025

Scope Note

ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were identified prior to 9:00 AM (EDT) on September 15, 2025; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

| Brief | Detecting and Countering Synthetic Media

| Key Points

- Advancements in the quality of synthetic media now available have made it an attractive and powerful tool for threat actors across the cybercrime landscape. By manipulating imagery, video, or audio, attackers can better increase their chances of bypassing traditional security measures and enhance social engineering campaigns.
- In the next 12 months, AI detection tools will very likely remain heavily reliant on forensic analysis of digital artifacts such as pixel-level inconsistencies, metadata anomalies, and signal-based markers introduced during synthetic generation.
- Over the next one to three years, advances in GenAI models will very likely diminish the reliability of current forensic indicators. The convergence of detection with authentication frameworks will very likely shift the burden of proof from detecting fakes to verifying authenticity.

Introduction

The rapid development of artificial intelligence (AI) in recent years has enabled the creation of highly convincing synthetic media that is readily available across the digital landscape, which is likely especially appealing to low-skilled actors. Synthetic media refers to content (audio, video, imagery, or text) generated or modified using AI—particularly machine learning (ML) models, a subset of AI that enables systems to learn from data without being explicitly programmed. Synthetic media has introduced new vectors of attack from threat actors for network, information, and human exploitation; this includes manipulating public opinion and breaching enterprise security via business email compromise (BEC) and various phishing methods. The readily available ability to fabricate audio, video, and images can significantly undermine trust in digital evidence, communication, and identity verification. As model architectures have become increasingly more sophisticated, so has the quality of synthetic media.

How Is Synthetic Media Generated?

Generative models—ML architectures trained to create data that imitates reality—are at the core of synthetic media generation; they can be trained to generate audio, video, image, or multimodal datasets used to produce various types of synthetic content. Below are three of the most common types of generative models:

- **Generative Adversarial Networks (GANs)**. GANs consist of two competing neural networks: a generator (which creates synthetic data) and a discriminator (which evaluates whether the data appears real). GANs are widely used for realistic face generation, high-resolution image synthesis, and identity manipulation (such as face swapping).
- **Autoencoders**. These models learn to compress and reconstruct data and are often used for tasks requiring transformation between two inputs (such as mapping one person's face onto another's). These models form the basis for face swapping, lip-synching, and identity morphing in video content.
- **Diffusion models**. These are trained by progressively adding noise (typically random pixel values) to a data set and then learning to reverse this process in order to generate high-quality data, such as images.

In addition to the core generative models, there is a range of supporting tools and techniques available to further enhance synthetic media creation.

- **Text-to-speech (TTS) and voice cloning models**. Leveraging deep learning software such as Tacotron or WaveNet, users can replicate a target individual's voice with minimal audio samples. This enables realistic impersonation in phone calls, audio messages, or voice-controlled systems.
- **Neural rendering**. Used primarily in video and imagery, neural rendering techniques (such as Nvidia's RTX Kit) enhance the realism of synthetic outputs by adding lighting, shadows, and natural motion patterns that mimic human behavior.
- **Multimodal models**. Some generative models combine text, audio, and visual inputs to create coherent outputs across multiple media types. An example includes generating a video where a person appears to speak scripted text with realistic lip movement and facial expressions.

The accessibility of pre-trained models, cloud-based AI services, and open-source tools has lowered the technical barriers to entry. Threat actors no longer need extensive computing resources or deep technical expertise to produce convincing synthetic media, making it more accessible for less-skilled actors.

How Is Synthetic Media Exploited?

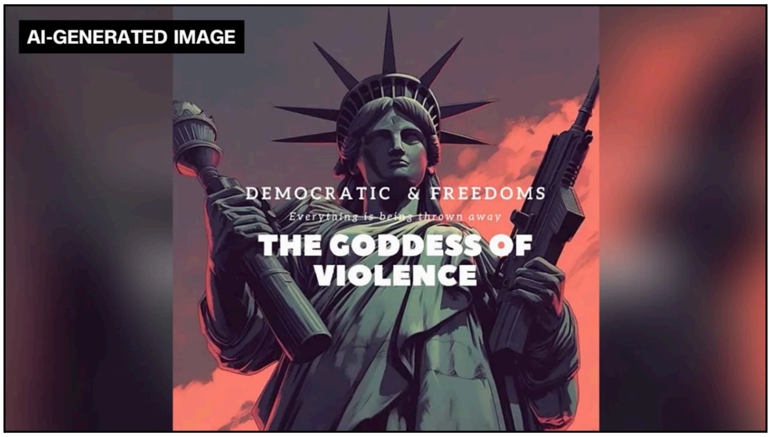
Advancements in the quality of synthetic media now available have made it an attractive and powerful tool for threat actors across the cybercrime landscape. By manipulating imagery, video, or audio, attackers can increase their chances of bypassing traditional security measures and enhance their social engineering campaigns.

Synthetic Imagery

Synthetic imagery encompasses AI-generated or manipulated images of people, documents, products, or environments. These can be indistinguishable from authentic images, allowing attackers to create false visual evidence or fabricate identities to assist in an array of malicious activities. For example, AI-generated photographs of an individual can be used to produce fake identification documents, bypassing security checks and authentication measures.

Furthermore, synthetic images often play a role in social engineering, reinforcing phishing campaigns or fraudulent communications. These lend credibility to requests for sensitive information such as login credentials or personally identifiable information (PII); for example, a synthetic image of a CEO could be used in the sender profile data of a phishing email. Synthetic imagery is also increasingly deployed in disinformation campaigns, wherein manipulated visuals are used to misrepresent events, produce false messaging, erode trust in organizations, undermine the intended target, or manipulate public opinion.

- Since approximately March 2023, suspected Chinese influence operations (IO) assets have reportedly posted on Western social media posing as American citizens and have begun to generate synthetic imagery in likely efforts to spread disinformation.[1]
- Notably, in the image below, there are five fingers and a thumb on the right hand; such mistakes are common in synthetic image generation.

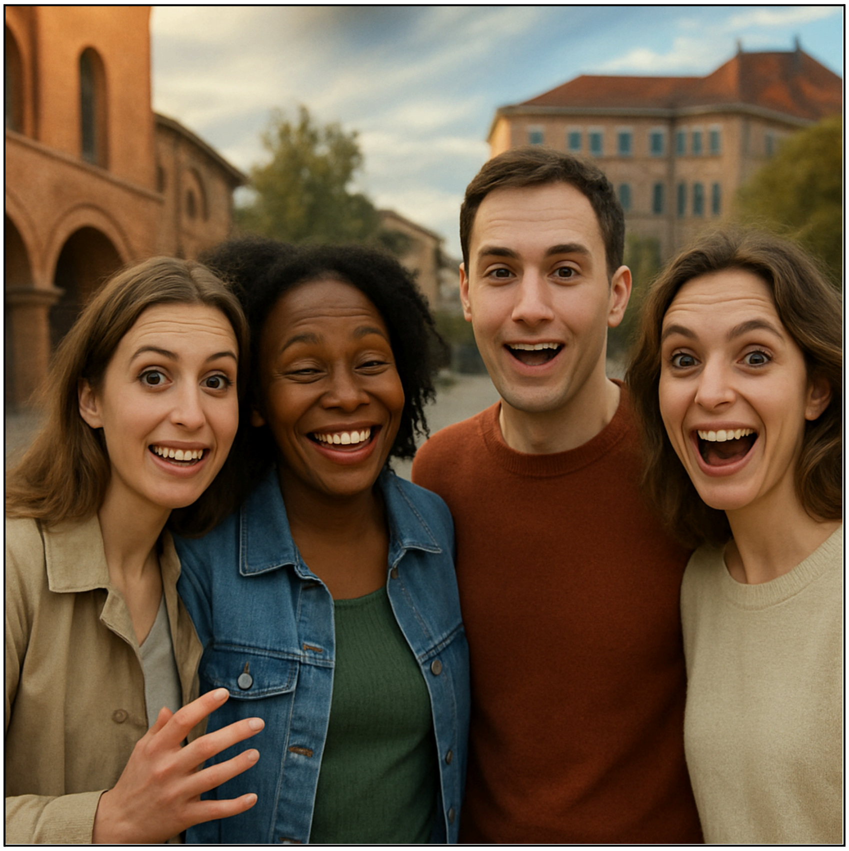


Synthetic image from suspected Chinese IO assets

Source: hXXps://cdn-dynmedia-1.microsoft[.]com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Digital-threats-from-East-Asia-increase-in-breadth-and-effectiveness.pdf

[1] hXXps://cdn-dynmedia-1.microsoft[.]com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Digital-threats-from-East-Asia-increase-in-breadth-and-effectiveness.pdf

- There are several key indicators in the sample image below that would suggest it is AI-generated: distorted fingers, misaligned buttons, warped building structure, and inconsistent lighting between the foreground and the background.



Synthetic image sample

Source: ChatGPT

Resilience: To help reduce the risks, organizations can train employees to identify synthetic imagery features such as visual and lighting inconsistencies, unnatural facial and bodily features, or irregular backgrounds. Technical tools such as SynthID can assist in detection by analyzing pixel patterns and identifying elements in the image that are indicative of generative models. Implementing strict verification procedures—such as cross-checking images against official sources or previously authenticated media—is likely to reduce the chances of successful exploitation.

Synthetic Video

Synthetic video extends the capabilities of AI-generated imagery by introducing motion, speech, and synchronized facial expressions. To produce convincing videos in which real-life victims are targeted to appear to perform actions and their speech is fabricated, attackers can use techniques such as auto encoders, GAN-based face swaps, or multimodal diffusion models. For example, executives or public figures can be impersonated to instruct employees to transfer funds, disclose credentials, or perform other sensitive tasks, significantly increasing the success rate of BEC and other attack campaigns.

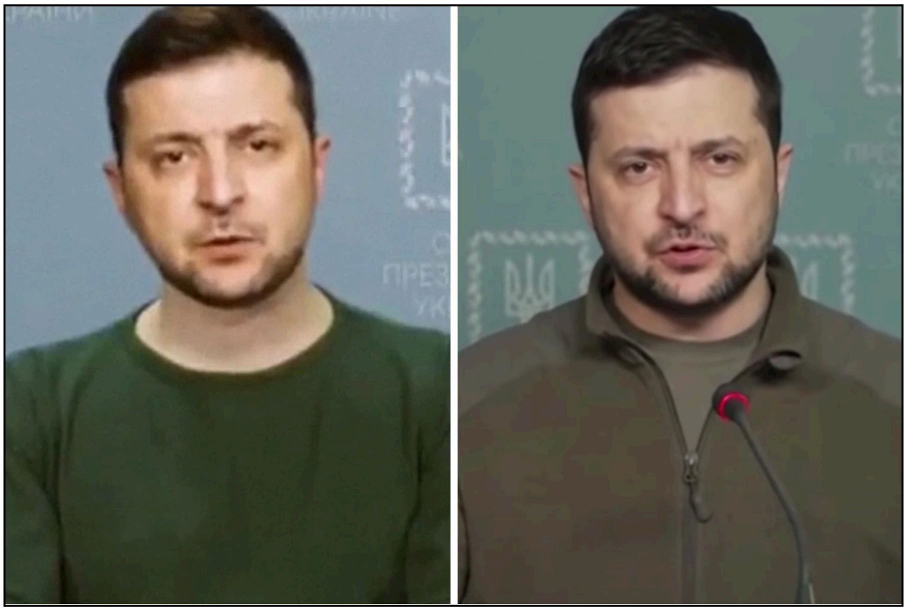
- In February 2024, an employee at a Hong Kong-based multinational company received a phishing email from an individual who appeared to be the Chief Financial Officer (CFO). A video call took place between the employee and what turned out to be a deepfake mimicking the CFO, which convinced the employee to transfer approximately USD 25 million to an account controlled by a threat actor.[2]

Beyond financial exploitation, synthetic video can be used in disinformation campaigns to distort public perception, cause reputational damage, or influence political and corporate environments. Videos placing individuals in fabricated scenarios are also likely used to intimidate, coerce, or manipulate targets, as well as commit malicious activity such as obtaining PII for further social engineering campaigns.

- In March 2022, a deepfake video that appeared to show Ukrainian President Volodymyr Zelensky calling on Ukrainian troops to surrender and lay down their arms circulated on social media. One version of the video was viewed more than 120,000 times on X (formerly, Twitter).[3]

[2] hXXps://arstechnica[.]com/information-technology/2024/02/deepfake-scammer-walks-off-with-25-million-in-first-of-its-kind-ai-heist/

[3] hXXps://www.euronews[.]com/my-europe/2022/03/16/deepfake-zelenskyy-surrender-video-is-the-first-intentionally-used-in-ukraine-war



Images of Zelensky: synthetic (left); authentic (right)

Source: *hXXps://news.virginia[.]edu/content/qa-zelenskyy-surrender-hoax-feared-future-deepfakes-here*

Synthetic Image:

- The skin tone looks flatter and less detailed; it is likely that a smooth filter has been applied.
- The edges around the face and hair seem slightly blurred.
- The lighting on the face appears uniform and unnatural, without the subtle depth and shadowing typical in authentic images.
- The eyes and mouth appear slightly off in alignment and expression, which is a key indicator of deepfakes.

Authentic Image:

- There is more natural skin detail and variation in skin tone.
- The lighting falls more naturally across the face and clothing and feels soft, balanced, and true to how light behaves in the real world, without harsh shadows or unnatural highlights.

- A microphone is present and integrated naturally in the scene.
- The background has clearer, sharper details compared with the blurred or washed-out look in the synthetic image.

Resilience: Detecting synthetic video requires careful human observation, including attention to unnatural facial movements, irregular blinking, inconsistent lip-synching, or lighting irregularities. Forensic tools such as DeepFaceLab, FakeCatcher, and FaceForensics++ enhance detection by analyzing micro-expressions, motion artifacts, and other features that are difficult for generative models to replicate accurately. Verification against trusted sources and corroborating evidence remains a critical step before responding to or acting upon instructions conveyed in video form.

Synthetic Audio

Synthetic audio leverages AI-generated or manipulated voice recordings in order to impersonate real-life victims. With only a few sample recordings—which can be easily attained through online publications—attackers can convincingly replicate a victim's voice through TTS systems or voice-cloning models. This tactic is increasingly being used in telephone-based fraud, extortion campaigns, and social engineering. Automated voice scripts allow these attacks to be conducted at scale with minimal human effort, further amplifying their potential impact.

- In June 2025, a woman in Florida, United States, received a phone call from a scammer using synthetic audio to impersonate her daughter. On the call, the scammer "daughter" stated that she had been in a car accident and needed USD 15,000 for bail, which the mother paid.[4]

[4] hXXps://www.malwarebytes[.]com/blog/news/2025/07/car-crash-victim-calls-mother-for-help-and-15k-bail-money-but-its-an-ai-voice-scammer

Resilience: Defending against synthetic audio requires both human vigilance and technical measures. Employees can be trained to better recognize abnormalities in speech, such as robotic cadences or irregular pauses. Tools such as AI Voice Detector can detect irregularities in voice frequency and patterns that could be indicative of synthetic generation. To prevent unauthorized actions, sensitive communications should be subject to strict verification measures, such as independent, multi-channel identity confirmation and pre-established authentication procedures.

Outlook

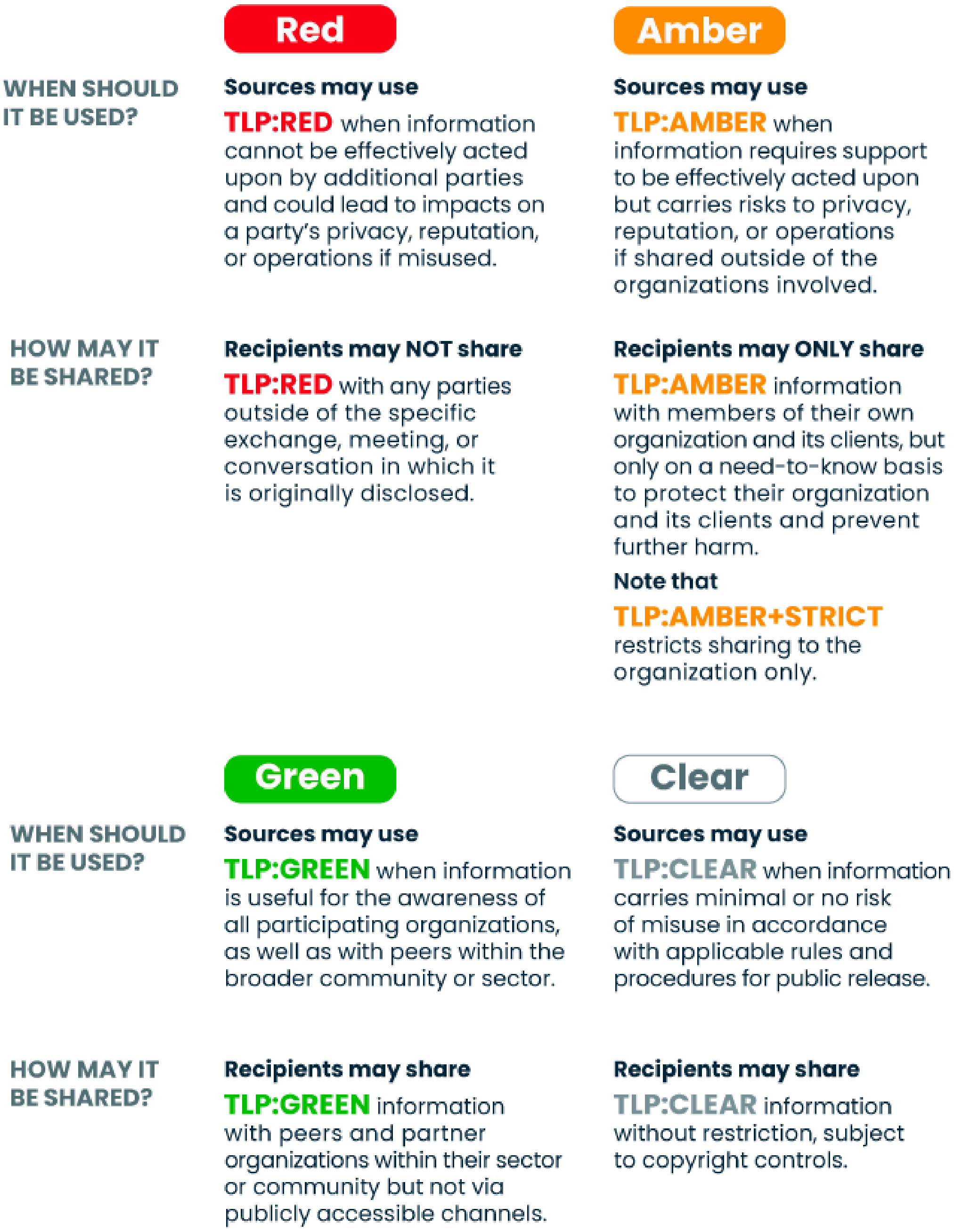
In the next 12 months, AI detection tools will very likely remain heavily reliant on forensic analysis of digital artifacts such as pixel-level inconsistencies, metadata anomalies, and signal-based markers introduced during synthetic generation. Commercial and open-source detection solutions will likely proliferate, but their uneven accuracy will likely limit confidence in operational settings. It is very likely that threat actors will seek to focus on low-effort, high-impact content such as voice cloning for fraud. To mitigate this exploitation, defensive measures will likely be centred around cross-source verification and awareness training.

Over the next one to three years, advances in GenAI models will very likely diminish the reliability of current forensic indicators. The convergence of detection with authentication frameworks will very likely shift the burden of proof from detecting fakes to verifying authenticity. Threat actors will very likely adapt by blending synthetic media with authentic media to better evade automated detection. Greater defensive efforts will almost certainly be required; this includes the integration of multi-source verification techniques and procedures to detect malicious activity. It is almost certain that collaboration between the government, private sectors, and platform providers will be critical to establishing common standards of trust.

Recommendations

- Integrate media authenticity verification tools to confirm the source and integrity of digital content.
- Conduct regular awareness training about synthetic media threats and social engineering tactics that use deepfakes.
- Enhance social media monitoring by using open-source intelligence (OSINT) tools to detect synthetic or impersonated content.
- Limit the availability of publicly accessible voice, video, or image material of key personnel to reduce source data for training synthetic media.
- Tag internal and official media assets with digital watermarks or hashes to support tamper detection and origin training.
- Establish a rapid media validation protocol for verifying suspicious content, especially those involving executives.
- Leverage external intelligence sources to track emerging synthetic media tactics, techniques, and procedures (TTPs) and campaigns relevant to your industry.
- Update incident response plans to include synthetic media scenarios (e.g., impersonation of executives, fake press releases) and ensure legal, public relations, and executive teams are prepared to respond.
- Deploy AI/ML-based detection tools for identifying manipulated media, including frame-level image and video anomaly detection and audio analysis.

Appendix A: Traffic Light Protocol for Information Dissemination

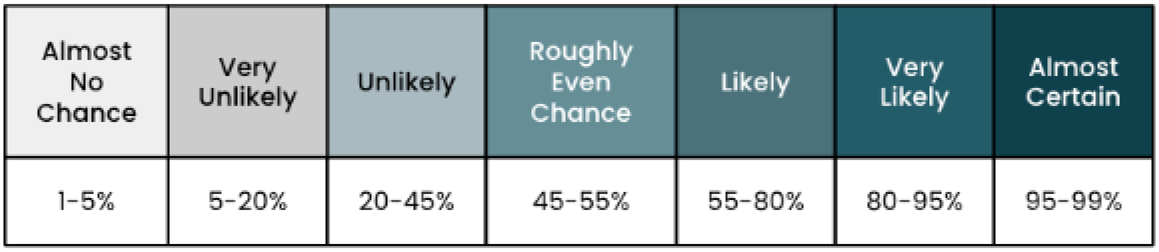


	Red	Amber
WHEN SHOULD IT BE USED?	**Sources may use** **TLP:RED** when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	**Sources may use** **TLP:AMBER** when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	**Recipients may NOT share** **TLP:RED** with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	**Recipients may ONLY share** **TLP:AMBER** information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. **Note that** **TLP:AMBER+STRICT** restricts sharing to the organization only.

	Green	Clear
WHEN SHOULD IT BE USED?	**Sources may use** **TLP:GREEN** when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	**Sources may use** **TLP:CLEAR** when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	**Recipients may share** **TLP:GREEN** information with peers and partner organizations within their sector or community but not via publicly accessible channels.	**Recipients may share** **TLP:CLEAR** information without restriction, subject to copyright controls.

Appendix B: ZeroFox Intelligence Probability Scale

All ZeroFox intelligence products leverage probabilistic assessment language in analytic judgments. Qualitative statements used in these judgments refer to associated probability ranges, which state the likelihood of occurrence of an event or development. Ranges are used to avoid a false impression of accuracy. This scale is a standard that aligns with how readers should interpret such terms.



Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain
1-5%	5-20%	20-45%	45-55%	55-80%	80-95%	95-99%