The Mis/Dis-information Problem is *Hard* to Solve

Gregory Hagen[1], Reihaneh Safavi-Naini[1], and Moti Yung[2]

[1] University of Calgary, Calgary, Canada
[2] Google LLC and Columbia University, New York, USA

Abstract. Securing information communication dates back thousands of years ago. The meaning of *information security*, however, has evolved over time and today covers a very wide variety of goals, including identifying the source of information, the reliability of information, and ultimately whether the information is trustworthy.
In this paper, we will look at the evolution of the information security problem and the approaches that have been developed for providing information protection. We argue that the more recent problem of misinformation and disinformation has shifted the content integrity problem from the protection of message syntax to the protection of message semantics. This shift, in the age of advanced AI systems, a technology that can be used to mimic human-generated content as well as to create bots that mimic human behaviour on the Internet, poses fundamental technological challenges that evade existing technologies. It leaves social elements, including public education and a suitable legal framework, as increasingly the main pillars of effective protection, at least in the short run. It also poses an intriguing challenge to the scientific community: to design effective solutions that employ cryptography and AI, together with incentivization to engage the global community, to ensure the safety of the information ecosystem.

Keywords: Information security, misinformation, disinformation.

1 Introduction

Securing information dates back thousands of years ago, primarily in the form of *secret communication*. In the time of Romans, Greeks, and Persians, generals needed to send secret messages to their allies and subordinates during wars and the message had to be hidden from the prying eyes of the enemies [1].

Steganography, the art of hiding the "existence of the message," was born. Steganography conceals a message within another cover message or object to hide its presence from anyone except the intended recipient. One of the most famous early examples of steganography dates from 500 B.C. Following the conquest of Miletus by Darius, King of Persia, Histiaeus, the tyrant of Miletus, found himself imprisoned. To communicate with his son-in-law, Aristagoras, and incite him to rebellion, Histiaeus selected his most loyal slave, shaved his head, and inscribed

a message on his scalp urging Aristagoras to revolt against Darius [2]. As the slave's hair grew back, he was sent to Aristagoras. Upon reaching Aristagoras, the slave's head was once again shaved, revealing the hidden message that ultimately led to the Ionian Revolt against Persian rule.

More recent examples of clandestine communication methods include the use of lemon juice, onion juice, or more sophisticated chemical compounds to conceal messages. Such techniques have been used by spies and agents throughout history. During World War II, agents often employed invisible ink made from various substances, such as milk, urine, or specially formulated chemicals that when applied to paper, remained invisible until activated by heat, ultraviolet light, or another chemical agent [3].

Over time, and especially with the introduction of the Internet and the World Wide Web, the security of information went beyond the need of generals and statesmen and became increasingly relevant and, in fact, crucial for everyone. Electronic commerce, heightened connectivity, and increasing integration of the physical and virtual worlds have elevated information security to a central role in the information exchange infrastructure of society. Two broad categories of techniques to protect information have emerged: cryptographic techniques and techniques that rely on data analysis and machine learning. These technologies have complementary roles: cryptography is primarily preventative and ensures that the security goals of an information system are met. Machine-learning and data analysis techniques, on the other hand, contribute primarily to the detection of misbehaviours. They may look for pre-defined combination of features or anomalous patterns of behaviour, using large data sets that are collected from the system of interest, and use them to detect signs of a breach. The two in combination aim to provide a *safe information ecosystem*: a digital environment where information is trustworthy and accessible to authorized users, while protecting individuals' privacy rights and "digital rights" that would be required in an information-driven world.

In the following, we review important elements of these two approaches as they have evolved over time, and then discuss the problem of trustworthiness of digital content, which, in a very broad sense, encompasses content integrity, veracity, and the ability to trace it to its source. We then outline the limitations of existing tools and techniques to cope with the more recent emerging problem of misinformation and disinformation. Our main observations are the following.

- Protection against mis/dis-information broadens the scope of *information security* to include protection of the *meaning of a message* in addition to the protection of its *representation* (semantics vs. syntax).
- Advancement of computing and, in particular, AI has provided tremendous power to generate content and disseminate it through the Internet, mimicking human language and interaction on the Internet.
- Protecting the semantics of messages is closely linked to identifying the *message's source*, which, in the age of generative AI, is increasingly challenging due to the difficulty of distinguishing AI-generated content from human-

generated content. Technologies aiming to make this distinction face a steep uphill battle.
- Social measures, in particular, public education that emphasizes critical thinking and the evaluation of everyday information streams, as well as carefully constructed laws that define the boundaries of social interaction, form the main pillars of protection, at least in the short run.
- In the long term, combining security technologies with economic incentives could provide a promising direction for developing future technologies.

2 Technologies for Securing Information

Securing information communication systems uses a wide range of hardware and software systems. Algorithmic approaches to security can be broadly divided into cryptographic ones and data analysis/machine learning ones.

2.1 Cryptography

Cryptography started with encryption, whose goal is *hiding the message content*, but not its existence. The Caesar cipher is an elementary yet historically significant encryption algorithm that dates back to ancient Rome [1]. Named after Julius Caesar, who purportedly used it to encode his private communications, the cipher algorithm works by shifting the alphabet. Each letter in the plaintext is substituted with a letter that is a fixed number of positions down the alphabet. The Caesar cipher is an example of a *substitution cipher* which uses a permutation of the alphabet to replace each letter with another one. Another early cipher was the *transposition cipher*, which rearranges the order of letters in the plain-text to create the ciphertext. One historical example of a transposition cipher is the Scytale, an ancient Spartan cryptographic tool [1]. The Scytale involved wrapping a strip of parchment around a rod of a particular diameter and then writing the message lengthwise. When unwrapped, the message would appear scrambled and indecipherable unless the recipient possessed a rod of the same diameter, allowing for the message's reconstruction.

As cryptography evolved, cryptographers introduced a wide range of algorithms which in many cases consisted of variations and compositions of the above two approaches to increase security, while being able to analyse and predict the behaviour of the algorithm. One of the prominent examples of combining substitution and transposition ciphers to construct a powerful encryption system is in the *Enigma machine*, an intricate encryption device used extensively by the German military during World War II [4]. The Enigma machine employed a series of rotors and electrical pathways to perform letter substitutions and transpositions. The resulting ciphertext produced by the Enigma machine was highly secure, owing to the combination of substitution and transposition techniques, along with the machine's mechanical complexity and the frequent changing of rotor settings.

Cryptography and Cryptanalysis. Cryptanalysis was pioneered by ancient cryptanalysts as the art of deciphering encrypted messages without access to the encryption key. As cryptographers developed increasingly complex ciphers to protect communication, cryptanalysts devised new methods to break the ciphers. Al-Kindi, a ninth-century Arab mathematician, wrote one of the earliest known treatises on cryptanalysis [5]. Al-Kindi's treatise not only outlined methods for breaking simple substitution ciphers but also delved into the principles of frequency analysis and linguistic patterns, demonstrating a sophisticated understanding of cryptologic concepts centuries ahead of its time. Cryptography and cryptanalysis are complementary and together form the art and science of cryptology. The interplay between the two disciplines of cryptography and cryptanalysis resulted in the creation of stronger ciphers and the refinement and enhancement of cryptanalytic methods.

The *science of cryptography started with secrecy systems.* In 1949, Shannon laid the foundation of scientific cryptography by formalizing perfect secrecy and showing a system that achieves it [6]. Shannon proved that *One-Time-Pad (OTP*), an encryption system that was invented tens of years earlier [7] and was used during World War I, achieved perfect secrecy. OTP, however, requires a random key that is as long as the message to be transmitted and must be used only once. The key must be securely shared between the two communicating parties before the system is used. OTP found limited application, mainly in highly secure military settings, such as the Moscow-Washington "hot line" during the 1940s and 50s. During this period the main application of cryptography was the secrecy of communication.

Today, standardized encryption algorithms like AES (Advanced Encryption Standard) represent the pinnacle of cryptographic resilience. The AES algorithm has been examined and analyzed by experts around the world for over two decades, without finding any real weakness in the design. It has been adopted by governments and industries and has demonstrated remarkable resistance against advanced cryptanalytic attacks [8]. Additionally, rigorous selection processes that include scrutiny and evaluation by the worldwide community of cryptography and cryptanalysts, followed by the widespread successful deployments in real-life systems, have fostered the required trust and acceptance that are essential for modern security systems.

The authentication problem: protecting against active attacks. In the 1970s, the growth of the Internet brought forward two new challenges beyond hiding the message content (e.g., to provide confidentiality or privacy): the need to detect tampering with the content and the ability to trust the claimed identities of entities from whom information is received. As the digital world expanded, individuals found themselves navigating a realm where verifying the authenticity of source and content became increasingly complex. This challenge was best encapsulated in a widely circulated 1993 New Yorker cartoon featuring a dog sitting at a computer, declaring to a fellow dog: "On the Internet, no one knows you are a dog." The image perfectly captures the uncertainty surrounding identity verification in an expanding virtual world [9]. Indeed there were direct adver-

sarial and financial gains from the mis-representation of identities and content. Malicious actors used strategies such as claiming a false identity and tampering with information for financial fraud, or impersonation of reputable organizations to deceive users into divulging their sensitive information or transfer funds.

The ground breaking invention of public key cryptography and digital signatures by Diffie and Hellman promised a conceptual solution to the problem of trusting identities and content origin in a virtual world [10]. In this paradigm, Alice has a unique pair of keys: a secret key (sk_A) known only to her and a corresponding public key (pk_A) that is public and associated with Alice's identity. Alice can securely link her identity to digital messages and documents by digitally signing them, using her secret key sk_A. The resulting signature can then be verified by anyone possessing her public key pk_A, thereby linking the document to its originator. Ensuring the integrity of messages between two trusting parties, however, uses *Message Authentication Codes (MAC)* that enable two parties who share a secret key to efficiently detect tampering of messages that are communicated between them. *"Codes which detect deception"* and can be used to authenticate messages in the presence of an adversary with unlimited computation were introduced in the pioneering work of Gilbert, MacWilliams and Sloane [11], and later formalised by Simmons [12]. It appeared that a safe information ecosystem was finally on the horizon.

The use of the term "authentication" in cryptography has many subtleties, and depends on the underlying cryptographic primitive, and possibly the context of use. A digitally signed document by Alice guarantees that the document has been generated by Alice when she is (securely) associated with the corresponding public verification key and tampering by *any other party* is detectable. A message with an attached MAC, however, can detect tampering by a *third party* who is distinct from the two trusting parties who have a shared secret key that is used for constructing the MAC. This means that either of the two trusting parties can undetectably change a document that is protected by a MAC. There are even more subtle nuances in using digital signatures to uniquely identify the sender of a message. In [13], it is argued that identification of the source of a message could be done for the purpose of assigning "responsibility" to the source or giving "credit" to them. For example, when Bob receives a signed message M from Alice to delete a file, then Bob can delete the file and gives the responsibility for deleting the file to Alice. However, when Bob receives a signed message M from Alice that shows the two previously unknown prime factors of a large integer, Bob gives "credit" to Alice for finding the primes. It is argued in [13] that the two are very different and a protocol that may work securely for one purpose, can be insecure for the other purpose. In the following we do not consider these subtleties and consider a digital signature as a cryptographic primitive that can uniquely identify the source of a message.

Cryptography everywhere. Fast forward almost 50 years from Diffie and Hellman's groundbreaking paper and cryptography has become ubiquitous as the primary tool for ensuring security in the digital age. Today, its influence permeates every facet of our lives, from the simplest devices like garage door openers to the vast

array of internet-connected devices in our homes, that we carry with us in our daily lives, and that we use in workplaces and industrial systems. Cryptography has revolutionized the way that we communicate, access services, work and interact remotely, and has become an integral component of secure functionalities across various domains. It has also played a pivotal role in innovations such as blockchain technology and other decentralized systems that have transformed traditional paradigms of trust and accountability.

2.2 Machine Learning For Security

The Prehistory of AI. The field of artificial intelligence (AI) developed around creating artificial agents that can mimic cognitive functions that are typically associated with humans, such as learning, reasoning, problem-solving, perception, and decision-making. In Greek mythology, Talos was a giant bronze automaton created by Hephaestus, the god of fire and craftsmanship, tasked with guarding the island of Crete [14]. Talos shares several characteristics of modern AI systems, including *autonomous decision-making*, *pattern recognition*, and *goal-oriented behavior*.

Early philosophers, notably Aristotle, sought to formalize deductive reasoning through symbol manipulation. Aristotle defined deduction as a discourse (logos) in which, given certain premises, a conclusion distinct from these premises logically follows because of the inherent relationship between them [15]. In his syllogistic theory, knowledge is encapsulated in the premises, and new knowledge can be deduced through logical inference. Subsequent formalizations of logic by Boole, Frege, Russell, Hilbert, and others ultimately led to the development of first-order logic [16], a cornerstone of knowledge representation and reasoning in AI [17].

The development of systems of formal reasoning was accompanied by the creation of mechanical systems for the automation of reasoning. The invention of mechanical calculators, such as Leibniz's calculating machine and Pascal's Pascaline in the 17th century, and Babbage's Analytical Engine in 19th century, were significant steps towards the development of machines capable of performing complex calculations [18]. One can also recognize the continued desire to simulate the world by building mechanical automata. These artifacts, which gradually became more sophisticated, are exemplified by Vaucanson's creations in the 18th century, including Digesting Duck [19], which, as the name suggests, was a mechanical duck that simulated the complete digestive system of a duck.

These developments, while not directly related to contemporary AI, represent important milestones in the evolution of ideas and technologies that ultimately contributed to the emergence of artificial intelligence. Mechanical calculators are precursors of todays' computers that use symbolic representation of knowledge and logical rules of deduction to infer new knowledge and mechanical self-governing automata can be seen as mechanical precursors of "bots" that we see on the Internet today. They reflect a growing interest in understanding the human mind and the potential for machines to simulate or even surpass human

capabilities. However, it wasn't until the 20th century that the foundation for modern AI was truly laid.

Birth of AI. The term "artificial intelligence" was coined by McCarthy and first used in the proposal for the Dartmouth Summer Research Project on Artificial Intelligence (1956) [20]. McCarthy considered the field of AI to have its roots in Alan Turing's article "Computing Machines and Intelligence" (1950) and Shannon's paper "Programming a Computer for Playing Chess" (1950) [21]. These works, respectively, laid the groundwork for AI by proposing a criterion for machine intelligence (the Turing Test) [22] and exploring the practical application of algorithms to tackle complex problems, such as developing strategies for playing chess [23]. McCarthy's LISP programming language contributed to knowledge representation, reasoning, and common-sense reasoning that are used in many AI systems today [24].

Other AI pioneers include Minsky and Newell, who played pivotal roles in defining the field's goals and research directions [20]. Minsky, a co-founder of the MIT AI Laboratory, made significant contributions to the development of artificial neural networks (ANNs), robotics, and created the theory of the *society of mind*, according to which intelligence emerges from the interaction of many simple agents from which a mind is built [25]. All of these contributions continue to influence AI today. Newell and Simon were pioneers in cognitive science and developed ideas about the relationship between symbolic systems and intelligent action [26]. They also developed the *Logic Theorist,* the first AI program capable of proving mathematical theorems, and the *General Problem Solver* (GPS), a program designed to solve a wide range of problems using heuristic search [27].

McCulloch and Pitts' 1943 paper "A Logical Calculus of the Ideas Immanent in Nervous Activity," sparked the study of neural networks (NNs) inspired by the brain's biological structure as a system of connected neurons [17]. Basic NNs consisted of input, hidden, and output layers, and could be trained on labeled datasets (e.g., dog and cat pictures) to be able to classify future new and unseen data [17]. Despite an early promising start, a combination of unmet expectations, funding cuts, and technical limitations led to periods of stagnation known as "AI winters" in the 1970s and 1980s [17]. The late 20th and early 21st centuries witnessed a resurgence in AI, driven by advancements in machine learning and the increasing availability of data. Successes in game-playing AI, such as IBM's Deep Blue defeating the world chess champion using a brute-force approach, and DeepMind's AlphaGo, which combined deep learning with search algorithms to defeat a world champion Go player, demonstrated AI's growing ability to tackle complex problems once thought to be the exclusive domain of human intelligence [17].

Conversational agents, designed to simulate human conversation, began in 1966 with ELIZA [28], which used pattern matching and substitution to engage with users in text conversations. They evolved into more capable agents by incorporating advancements in natural language processing and machine learning. Chatbots like A.L.I.C.E. (1995) and SmarterChild (2001) in the 1990s and 2000s, respectively, offered increasingly sophisticated responses, often serving as

virtual assistants or information providers [29]. Voice assistants like Siri (2011) and Alexa (2014) further blurred the lines between human and machine interaction, as these agents could understand and respond to spoken language [29]. The release of ChatGPT 3.5 in 2022 introduced an exceptionally more powerful assistant with unprecedented text-generation capabilities. Generative AI refers to AI systems that can produce new data, such as images, text, or audio, that are generated from examples it has been trained on. Today's generative AI systems, often based on deep learning techniques, generate "synthetic" content that is derived from captured knowledge and appears remarkably original and realistic. Deep generative AI models like OpenAI's ChatGPT, DALL-E, Google's PaLM, and Meta's Llama 2 are notable for their ability to generate content in audio, visual, and text formats that are increasingly difficult to distinguish from human-created content.

AI and Machine Learning in Security. The use of AI and machine learning for securing information communication dates back to the late 1980s and early 1990s. Initially, rule-based systems were employed for anomaly detection, where predefined rules were used to identify deviations from normal behavior. These early systems were limited in their ability to adapt and learn from new threats [30]. Building upon earlier work in the field of intrusion detection, the seminal work of Dorothy Denning introduced the concept of using statistical anomaly detection to identify intrusions in computer systems [31]. This laid the groundwork for the development of Intrusion Detection Expert Systems (IDES), which initially relied on rule-based systems but later incorporated machine learning techniques, such as decision trees and neural networks, to enhance their detection capabilities. Denning's work stands out as a major milestone due to its rigorous formulation of a statistical anomaly detection model and its explicit connection to intrusion detection.

The wider development of intrusion detection systems began in the early 1990s, with companies like Haystack Labs and SRI International [32] developing systems that utilized statistical anomaly detection. Today machine learning algorithms use a wide array of data including network traffic, system, device and user data, to identify unusual patterns or unexpected changes in each or combination of these data to signal a possible cyber-attack or data breach [33]. Deep learning models, like convolutional neural networks (CNNs), excel at analyzing vast and intricate datasets, enabling more accurate identification of anomalies in network traffic and other collected data that may indicate malicious activities [34].

AI-based approaches, however, face two main challenges. First, how to reliably translate a similarity score that is calculated for a potential breach and using many features, into a concrete decision, "attack" or "no attack", possibly with some indication of severity level, and balance the system's false positive and false negative. Second, how to detect new zero-day attacks for which prior information does not exist in the system. These challenges have motivated new approaches, such as unsupervised learning, to address these limitations.

In practice cryptographic and machine-learning based approaches are used in concert to improve security.

3 Identifying the "Source" of a Digital Object and the Integrity of the Object

Interest in identifying the "source" of digital content, in terms of both the originator and owner, as well as verifying its authenticity and intactness, grew with the rise of electronic commerce and the distribution of digital content over the Internet.

In the early 2000s, music sharing experienced a boom, with music and media sold in smaller units, such as individual songs, and packaged in various appealing forms. Digital signatures could securely link content to its originator and be used for signing contracts. However, digital signatures could easily be stripped from the content, allowing it to be copied and redistributed. The surge in copyrighted digital content distribution over the Internet posed a new challenge: preventing illegal copying and redistribution through peer-to-peer file sharing and other forms of unauthorized reproduction or communication of works.

The new emerging technological challenge was how to authenticate the source and integrity of a digital object and trace it to its origin or owner once cryptographic protections are removed. The challenge stems from the fact that one can simply pay the required price (or fee) to access a digital object and then, since the object is in digital form, easily and perfectly copy and redistribute it, ignoring the copyrights of the content owner. New cryptographic and noncryptographic solutions in the form of innovative *fingerprinting* and *watermarking* methods were introduced to trace digital objects to their originators and/or owners and to verify the origin and/or owner of the content, thereby facilitating copyright infringement actions.

3.1 Watermarking

A watermarking system embeds imperceptible or barely perceptible data into digital objects, such as images, audio, or video, to signify properties like authenticity, integrity, or ownership, as needed. The *robustness* of watermarking systems refers to the watermark's ability to withstand various forms of manipulation, distortion, or attacks while remaining detectable [35]. A robust watermark ensures the embedded information is reliably retrievable even after the content undergoes alterations or attempts to remove or tamper with the watermark itself.

Robustness is essential for ensuring the integrity of content and authenticity of origin, but hard to achieve because of the array of tools and techniques that can be used to remove the watermark by modifying the content or creating a new copy using a different device. Techniques such as image cropping, content scaling, and lossy compression algorithms have emerged as particularly effective tools for removing or altering watermarks to make them undetectable. Creating a new recording of a song or taking a new picture of an image can heavily reduce

the detectability of a watermark at the cost of reducing the quality of the digital object. Thus, watermark robustness, although essential for ensuring integrity and traceability (to the source) of the content, is hard to achieve because of the array of tools and techniques that can be used to manipulate the content and remove the watermark.

3.2 Digital Rights Management

Digital Rights Management (DRM) technology promised to securely manage access to, and the copying of, copyrighted content in accordance with copyright licenses [36]. DRM systems used watermarking and fingerprinting technologies combined with cryptography and tamperproof hardware to achieve security but introduced a plethora of new technological and legal challenges, as well as factors to be taken into account related to social dimensions.

Content origin, ownership, and tracing in practice. Today, watermarking, together with wide monitoring of Internet traffic, is used to detect copyright breaches across digital platforms. One concrete example of such a monitoring service is the Content ID system developed by YouTube [37]. Content ID employs algorithms to scan and analyze uploaded videos, comparing them against a vast database of copyrighted content provided by rights holders. When a match is detected, rights holders have the option to block the video, track its viewership metrics, or monetize it by running ads alongside the content.

A notable example of the Content ID system in action occurred when a user uploaded a music video featuring Ed Sheeran's hit song Shape of You without the necessary authorization from the copyright owner. The rights holder, alerted by Content ID, was able to identify the infringement and take appropriate action [38]. This example highlights the effectiveness of Internet traffic monitoring services in detecting and addressing copyright breaches in real-time.

Societal challenges. DRM gives rise to a myriad of citizens' rights issues, including how to implement copyright's balance in information technology. The fine line between, on the one hand, infringement and, on the other, fair use or fair dealing (or some other non-infringing use) is not easy to inscribe in code. As a result, the US Digital Millennium Copyright Act [39] and its analogues in other jurisdictions created the possibility that the copyright balance in technology and in law does not match. For instance, in Canada, the Copyright Act does not permit circumvention of an access control measure for the purposes of fair dealing, such as extracting a short excerpt of a music video for educational purposes and posting it online [40]. Arguably, forbidding such circumvention violates the right to freedom of expression as guaranteed by, e.g., the Canadian Charter of Rights and Freedoms [41]. A general conclusion that can be drawn is that regulating information technology systems runs the risk of violating the right to freedom of expression.

3.3 Fake Content

Fake content refers to any digital material that has been deliberately created, altered, manipulated, or falsified to convey misleading, deceptive, or false information. It encompasses a wide range of media, including images, videos, audio recordings, and written text, that have been tampered with to distort their original meaning or context. In the context of images and videos, fake content may use editing software to alter visual elements, such as adding or removing objects or modifying colors. Fake content includes counterfeit websites that mimic legitimate ones, with the intent to deceive users. These fake websites may closely replicate the design, layout, and branding of authentic sites, making them appear genuine at first glance.

Detecting fake content is a multifaceted exercise that may use a wide range of technologies, including forensic techniques and machine learning approaches. While detecting clumsy manipulation of media objects, for example an image, is possible for skilled users, more sophisticated cases that misrepresent the truth can become hard to detect, if possible at all. A widely discussed recent example of a clumsily modified image is a depiction of a street scene in Cuba that included a distant image of a man walking near some steps with a post coming out of his right leg [42]. After this discovery, several other images of the photographer were discovered to have been altered. The photographer admitted to using photoshop, saying that he was no longer a photo-journalist but, rather, a "visual storyteller" [43].

There are also examples of fake images that end up being used as evidence in a claim versus counter claim scenario. For example, in the "Obama skeet shooting" controversy, following debates over gun control legislation in the United States, the New Republic magazine tweeted a photograph that purported to show former President Barack Obama skeet shooting at Camp David, as he had claimed to have done [44]. The photograph was fake, however, having been inadvertently copied from a parody of *Whitehouse.gov* rather than the real website. To provide evidence for Obama's assertion, the White House later revealed another photograph that also purported to show Obama skeet shooting at Camp David, but it was derided by several conservative commentators as fake.
In the end, to be taken as evidence, there was a need to trust that the White House provided a genuine photograph of Obama skeet shooting and that it had not been tampered with.

4 The Problem of Fake Content in the Age of Generative AI

The problem of fake content in today's societies has two components: generation of fake content and its wide distribution. Advances in AI technologies have provided extremely powerful technologies for both.

Generative AI can quickly produce synthetic text, images, and audio that are increasingly difficult to distinguish from content generated through natural

or human-driven processes. Although Chatbots like ChatGPT are programmed with safety guidelines to avoid generating harmful or biased content [45], they can be "tricked" to produce misleading and fake output. Determined users can exploit vulnerabilities in the system by carefully crafting prompts (queries to the chatbot) or by directly modifying small parts of the generated content. Despite efforts to employ watermarking techniques to detect and trace AI-generated content, the race between possible manipulations of AI-generated content, with the goal of removing the watermark, and fortification of watermarking techniques to be robust against possible manipulations, will be an uphill battle. Without robust automated watermarking embedding and detection mechanisms, proliferation of AI-generated fake content will be unavoidable.

The distribution of fake content can be powered by AI-enabled agents (bots) that mimic user interaction on the web and social media. Using fake or stolen credentials, bots can be registered on various platforms and programmed to interact and participate in online networks to distribute fake content as part of well-orchestrated campaigns.

Disinformation and Misinformation are both used to describe false or misleading information, but they differ in their meaning. *Misinformation*, refers to false or inaccurate information shared without the intent to deceive, while *disinformation* involves the deliberate spread of false or misleading information with the intent to deceive or manipulate others [46]. Both can have significant implications for public discourse, decision-making, and social trust.

AI significantly amplifies the ability to generate and disseminate misinformation and disinformation. AI can be used to create digitally altered or synthesized media that convincingly depict events or individuals that never occurred or existed. (When such content is generated by deep neural networks, it is referred to as a "deepfake".) The outputs of generative AI systems are becoming increasingly indistinguishable from genuine content, blurring the lines between reality and fiction. Even reputable sources, including media organizations, are susceptible to being fooled by AI-generated content. For instance, users have submitted AI-generated images purportedly depicting events such as the Israel-Gaza crisis to stock image marketplaces like Adobe Stock [47].
AI-powered bots can be used to mimic human behavior on social media platforms, creating the illusion of widespread support for, or opposition to, specific ideas, candidates, or causes. By flooding online spaces with orchestrated messages, AI-generated misinformation can distort public discourse and amplify certain narratives while suppressing others. In [48], authors have described how the convergence of technologies, including AI, social media, bots and big data analysis have created an "epistemic crisis" that endangers democracy in the US. The United Nations Policy Brief on Information Integrity on Digital Platforms considers the spread of disinformation that undermines established scientific facts to be "an existential risk to humanity" [49].

Generative AI algorithms, fueled by deep learning techniques, have democratized the creation of synthetic media, making it strikingly easy for individuals without extensive expertise, advanced technical skills or specialized knowledge,

to generate highly realistic content. This accessibility has lowered the barrier to entry for malicious actors seeking to create and disseminate deceptive or misleading information.

4.1 The Challenge of Creating A Safe Information Ecosystem

Because of the ease of generating content by generative AI, a key challenge to the establishment of a safe information ecosystem is the detection of synthetic content, with the primary technique being the use of watermarking technologies. As discussed earlier, embedding robust watermarks in perceptual data is a formidable task. The task is even harder in text-based content that can be re-written in a myriad of ways. This challenge is well-recognized by experts and technology companies. It even led OpenAI, which had previously announced plans for a synthetic text detector, to retract its plan due to low accuracy [50]. Currently, there is no reliable method to discern whether text-based content was generated by AI or a human [51]. Additionally, there are many open-source generative models that can be run on personal computers and/or modified and extended to remove in-built protections and so there is no easy way to ensure that all synthetic contents will be watermarked.

Laws and regulations. Even if synthetic content could be reliably detected, it does not guarantee that its generation or transmission could be prevented, or that it should be prevented, given our right to freedom of expression. A safe information ecosystem does not require that all misinformation and disinformation be banished. As the Special Rapporteur on Human Rights has said: "The right to freedom of opinion and expression is not part of the problem, it is the objective and the means for combating disinformation."[52]. Freedom of expression does not imply that there can be no legal restrictions on the development or use of AI systems. The communication of some deepfakes or misinformation may constitute *private law* causes of action, such as defamation, deceit and misappropriation of personality as well as a criminal offence. For an example of the latter, the recently enacted UK Online Safety Act criminalizes the transmission of deepfake pornography [53]. Similar laws exist or are proposed in the US, Canada, Australia and other countries, in all cases, dealing with harms that have occurred. In addition, the European Union regulates the development, sale and use of AI under the Artificial Intelligence Act [54], and some countries, such as Canada [55], are seeking to create similar legislation.

The Global Declaration on Information Integrity Online says that a safe information ecosystem requires information integrity [56]. The declaration defines "information integrity" as "an information ecosystem that produces accurate, trustworthy, and reliable information, meaning that people can rely on the accuracy of the information they access while being exposed to a variety of ideas" [56]. This concept of "information integrity" goes well beyond the integrity of the "presentation of information" itself that is the goal of cryptography, watermarking techniques and similar technologies. Such an ecosystem will of necessity include humans who have developed the epistemic virtues (e.g. to properly rea-

son, doubt, and interpret) necessary to determine the trustworthiness of the source and veracity of the information provided.

4.2 What Is Possible

While efforts around the world aim to get a handle on AI safety, the focus is mostly on the safety and security of AI systems and how they are used in practice, rather than increasing public understanding of their operation and effects.

Given the limitations of both information security technology and current regulations to detect and control the spread of fake content and misinformation and to ensure the integrity of information, public education is the first and the most important step in curbing the problem of AI-generated fake content (in addition to continued efforts to find more effective technical solutions). Regulators and AI model developers need to prioritize informing the public about the capabilities, limitations, and effects of current AI models. Educational initiatives can begin by transparently demonstrating the capabilities of AI models to the public, accompanied by explanations of the underlying mechanisms and limitations of existing tracing and detection techniques, including watermarking.

Public education must also emphasize the necessity of learning how to scrutinize audio, visual, and text-based content, in order to evaluate its authenticity and the reliability of their sources. Developing these critical skills requires individuals to overcome inherent trust biases and navigate the complexities of media consumption in an age of social media and fragmented news sources.

Given the rise of generative models capable of creating synthetic media, fostering epistemic vigilance is crucial for individuals to navigate a world where differentiating between authentic and manipulated content is increasingly difficult. Automated tools that provide judgments and alerts can be helpful, and intensifying efforts to develop technical solutions is a must.

5 Concluding Remarks

The need to secure information and provide a safe information ecosystem is intertwined with our presence in the digital world. Protection against mis/disinformation requires a solution to the problem of ensuring the integrity of the *meaning* of messages, something which goes beyond traditional security technologies that largely aim to protect the integrity of *message representation* as a string of symbols. While cryptography and machine learning approaches have provided effective methods of protection for bit representations (syntax) of digital objects, they become blunt tools in ensuring the veracity of information that they carry. The protection against mis/disinformation in the age of generative AI is entangled with distinguishing between human-generated and AI-generated content. Generative AI poses a significant threat by blurring the line between authentic and fake content and interaction on the Internet as a whole. AI-powered bots can be employed as an organized army of agents in the service of a defined goal, weaponizing misinformation and disinformation to manipulate public

discourse and to achieve political objectives. In the absence of effective technological countermeasures, public education and carefully constructed laws become increasingly important in restoring safety to our today's information ecosystem.

Looking ahead, while no immediate, purely technological solution is in sight, innovative solutions combining technical approaches like cryptography, watermarking, and AI (i.e., "fighting fire with fire") with incentivization mechanisms offer intriguing opportunities for researchers and technology developers.

In recent years the combination of cryptography and economic incentives has been one of the key innovations behind the success of Bitcoin and blockchain technology. Smart contract-based systems like the MakerDAO protocol on the Ethereum blockchain have further extended the use of financial incentives and *penalties* to ensure the stability and security of the system. Such combined technological solutions, when supported by a well-crafted legal framework, may prove effective in moving towards a safer information ecosystem.

References

1. S. Singh, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*. New York: Anchor Books, 1999.
2. Herodotus, *The Histories*. Penguin Classics, 1996.
3. K. Macrakis, *Prisoners, Lovers, and Spies: The Story of Invisible Ink from Herodotus to al-Qaeda*. New Haven, CT: Yale University Press, 2014.
4. J. Greenberg, "The Enigma Machine," in *The Turing Guide*. Oxford University Press, 2017. [Online]. Available: https://doi.org/10.1093/oso/9780198747826.003.0018
5. al Kindi, "Al-Kindi (841) Manuscript on Deciphering Cryptographic Messages," King Faisal Center for Research of Islamic Studies, Riyadh, Saudi Arabia, 841, manuscript.
6. C. E. Shannon, "Communication Theory of Secrecy Systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.
7. S. M. Bellovin, "Frank Miller: Inventor of the One-Time Pad," *Cryptologia*, vol. 35, no. 3, pp. 203–222, 2011.
8. Canadian Centre for Cybersecurity, "Cryptographic algorithms for UNCLASSIFIED, PROTECTED A, and PROTECTED B information," Canadian Centre for Cyber Security, ITSP.40.111, 2023. [Online]. Available: https://www.cyber.gc.ca/sites/default/files/itsp.40.111_1-e.pdf
9. G. Fleishman, "Cartoon Captures Spirit of the Internet," *The New York Times*, December 14, 2000.
10. W. Diffie and M. E. Hellman, "New Directions in Cryptography," *IEEE Transactions on Information Theory*, vol. IT-22, no. 6, pp. 644–654, Nov. 1976.
11. E. N. Gilbert, F. J. MacWilliams, and N. J. A. Sloane, "Codes which detect deception," *Bell System Technical Journal*, vol. 53, no. 3, pp. 405–424, 1974.
12. G. J. Simmons, "Authentication theory/coding theory," in *Advances in Cryptology: Proceedings of Crypto' 84*. Berlin: Springer, 1985, pp. 411–432.
13. M. Abadi, "Two Facets of Authentication," *IEEE Symposium on Security and Privacy*, 1998.
14. A. Shashkevich, "Stanford researcher examines earliest concepts of artificial intelligence, robots in ancient myths," *Stanford Report*, 2019. [Online].

Available: https://news.stanford.edu/stories/2019/02/ancient-myths-reveal-early-fantasies-artificial-life/
15. R. Smith, "Aristotle's Logic," *Stanford Encyclopedia of Philosophy (Winter 2022 Edition)*, 2000. [Online]. Available: https://plato.stanford.edu/ENTRIES/aristotle-logic/
16. W. Ewald, "The Emergence of First-Order Logic," *Stanford Encyclopedia of Philosophy (Spring 2019 Edition),*, 2018. [Online]. Available: https://plato.stanford.edu/archives/spr2019/entries/logic-firstorder-emergence/
17. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2021.
18. H. H. Goldstine, *The Computer from Pascal to von Neumann.* Princeton, N.J.: Princeton University Press, 1972.
19. J. Riskin, "The defecating duck, or, the ambiguous origins of artificial life," *Critical Inquiry*, vol. 29, no. 4, pp. 599–633, 2003.
20. J. Moor, "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years," *AI Magazine*, vol. 27, no. 4, p. 87, Dec 2006.
21. J. McCarthy and P. J. Hayes, "Some philosophical problems from the standpoint of artificial intelligence," in *Readings in Artificial Intelligence*, B. L. Webber and N. J. Nilsson, Eds. Burlington, Mass: Morgan Kaufmann, 1981, pp. 431–450.
22. A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, October 1950.
23. C. E. Shannon, "Programming a computer for playing chess," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 41, no. 314, pp. 256–275, 1950. [Online]. Available: https://doi.org/10.1080/14786445008521796
24. S. Valencia, "The Lisp approach to AI (part 1," Medium, 2017. [Online]. Available: https://medium.com/ai-society/the-lisp-approach-to-ai-part-1-a48c7385a913
25. P. H. Winston, "Marvin L. Minsky (1927-2016)," *Nature*, vol. 530, p. 282, 2016.
26. A. Newell and H. A. Simon, "Computer Science as Empirical Inquiry: Symbols and Search," in *ACM Turing Award Lectures.* New York, NY, USA: ACM, 1976, vol. 19, no. 3, pp. 113–126, originally published as the 1975 ACM Turing Award Lecture.
27. J. E. Laird and P. S. Rosenbloom, "In Pursuit of Mind: The Research of Allen Newell," *AI Magazine*, vol. 13, no. 4, pp. 17–65, Dec. 1992. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1019
28. J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
29. I. Maglogiannis, L. Iliadis, and E. Pimenidis, "An Overview of Chatbot Technology," *Informatics*, vol. 7, no. 3, p. 37, 2020.
30. J. R. Yost, "The march of IDES: early history of intrusion-detection expert Systems," *IEEE Annals of the History of Computing*, vol. 38, no. 4, pp. 42–54, Oct.-Dec. 2016.
31. D. Denning, "An intrusion-detection model," in *Proc. IEEE Symp. Security and Privacy*, 1986, pp. 118–133.
32. G. Bruneau, "The history and evolution of intrusion detection," 2021. [Online]. Available: https://sansorg.egnyte.com/dl/TmT2wf11v7
33. A. Halimaa and K. Sundarakantham, "Machine learning based intrusion detection system," *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 197–205, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204229558

34. J. Lansky, S. Ali, M. Mohammadi, M. K. Majeed *et al.*, "Deep learning-based intrusion detection systems: A systematic review," *IEEE Access*, 2021.
35. M. van der Veen, A. Lemma, M. Celik, and S. Katzenbeisser, "Forensic Watermarking in Digital Rights Management," in *Security, Privacy, and Trust in Modern Data Management. Data-Centric Systems and Applications*, M. Petković and W. Jonker, Eds. Berlin, Heidelberg: Springer, 2007.
36. F. Dingledy and A. Berrio Matamoros, "What is Digital Rights Management?" may 2019.
37. Google, "Copyright and Rights Management," 2024, accessed on May 2, 2024. [Online]. Available: https://support.google.com/youtube/topic/2676339
38. "Ed Sheeran wins copyright case over 'Shape of You'." [Online]. Available: https://www.bbc.com/news/entertainment-arts-61006984
39. "Digital millennium copyright act," Pub. L. No. 105-304, 112 Stat. 2860 (1998) (codified in scattered sections of 17 U.S.C.), 1998.
40. "Copyright Act (R.S.C., 1985, c. C-42)," 1985, s. 41. [Online]. Available: https://laws-lois.justice.gc.ca/eng/acts/C-42/
41. G. Reynolds, "Step in the Wrong Direction: The Impact of the Legislative Protection of Technological Protection Measures on Fair Dealing and Freedom of Expression," *CJLT*, vol. 5, no. 3, 2006. [Online]. Available: https://digitalcommons.schulichlaw.dal.ca/cjlt/vol5/iss3/4/
42. D. Cade, "Botched Steve McCurry Print Leads to Photoshop Scandal," *Petapixel*, May 2016. [Online]. Available: https://petapixel.com/2016/05/06/botched-steve-mccurry-print-leads-photoshop-scandal/
43. L. Sanders IV, "'Ethical Lapse': Photoshop Scandal Catches up with Iconic Photojournalist Steve McCurry," *DW*, May 2016. [Online]. Available: https://www.dw.com/en/ethical-lapse-photoshop-scandal-catches-up-with-iconic-photojournalist-steve-mccurry/a-19296237
44. A. Tartar, "The Totally Serious Guide to Obama Skeet Shooting Photo Conspiracy Theories," 2013. [Online]. Available: https://nymag.com/intelligencer/2013/02/obama-skeet-shooting-photo-conspiracy-theories.html
45. OpenAI, "Our approach to AI safety." [Online]. Available: https://openai.com/index/our-approach-to-ai-safety/
46. A. M. Guess and B. A. Lyons, "Misinformation, disinformation, and online propaganda," in *Social Media and Democracy: The State of the Field, Prospects for Reform*, N. Persily and J. A. Tucker, Eds. Cambridge: Cambridge University Press, August 2020.
47. C. Wilson, "Israel-Gaza: Adobe accused of selling fake AI images," 2023. [Online]. Available: https://www.crikey.com.au/2023/11/01/israel-gaza-adobe-artificial-intelligence-images-fake-news/
48. Y. Benkler, R. Faris, and H. Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, 2018.
49. United Nations, "Information Integrity on Digital Platforms," June 2023. [Online]. Available: https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf
50. E. David, "OpenAI can't tell if something was written by AI after all," *The Verge*, 2023. [Online]. Available: https://www.theverge.com/2023/7/25/23807487/openai-ai-generated-low-accuracy
51. M. Heikkilä, " Why detecting AI-generated text is so difficult (and what to do about it) Plus: AI models generate copyrighted images

and photos of real people," *Technology Review*, 2023. [Online]. Available: https://www.technologyreview.com/2023/02/07/1067928/why-detecting-ai-generated-text-is-so-difficult-and-what-to-do-about-it/

52. I. Khan, "Disinformation and freedom of opinion and expression, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression," 2021. [Online]. Available: https://digitallibrary.un.org/record/3925306?ln=en&v=pdf
53. B. Edwards, "UK Seeks to Criminalize Creation of Sexually Explicit AI Deepfake Images Without Consent," *Ars Technica*, April 2024. [Online]. Available: https://arstechnica.com/information-technology/2024/04/uk-seeks-to-criminalize-creation-of-sexually-explicit-ai-deepfake-images-without-consent/
54. "Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828," 2023.
55. "Bill C-27, An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts (Canada)," (second reading 24 April 2023).
56. "Global Declaration On Information Integrity Online (Canada)," 2023, accessed: 2024-05-12. [Online]. Available: https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/information_integrity-integrite_information.aspx?lang=eng