IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 May 2023, accepted 6 July 2023, date of publication 10 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294090

SURVEY

Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods

EVAN N. CROTHERS[1], **NATHALIE JAPKOWICZ**[2], **AND HERNA L. VIKTOR**[1]
[1]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[2]Department of Computer Science, American University, Washington, DC 20016, USA

Corresponding author: Evan N. Crothers (ecrot027@uottawa.ca)

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Reference RGPIN-2018-04047.

ABSTRACT Machine-generated text is increasingly difficult to distinguish from text authored by humans. Powerful open-source models are freely available, and user-friendly tools that democratize access to generative models are proliferating. ChatGPT, which was released shortly after the first edition of this survey, epitomizes these trends. The great potential of state-of-the-art natural language generation (NLG) systems is tempered by the multitude of avenues for abuse. Detection of machine-generated text is a key countermeasure for reducing the abuse of NLG models, and presents significant technical challenges and numerous open problems. We provide a survey that includes 1) an extensive analysis of threat models posed by contemporary NLG systems and 2) the most complete review of machine-generated text detection methods to date. This survey places machine-generated text within its cybersecurity and social context, and provides strong guidance for future work addressing the most critical threat models. While doing so, we highlight the importance that detection systems themselves demonstrate trustworthiness through fairness, robustness, and accountability.

INDEX TERMS Artificial intelligence, cybersecurity, disinformation, generative AI, large language models, machine learning, text generation, threat modeling, transformer, trustworthy AI.

I. INTRODUCTION

A. RISKS OF MACHINE-GENERATED TEXT

Recent natural language generation (NLG) models have taken a significant step forward in the diversity, control, and quality of machine-generated text. The ability to create unique, manipulable, human-like text with unprecedented speed and efficiency presents additional technical challenges for detecting abuses of NLG models, such as phishing [1], [2], disinformation [3], [4], [5], fraudulent product reviews [4], [6], academic dishonesty [7], [8], and toxic spam [9]. Addressing the risk of abuse is vital to maximize the potential benefit of NLG technology, while minimizing harm — a fundamental principle of trustworthy AI [10].

The overwhelming majority of contemporary state-of-the-art NLG models are neural language models (NLMs) based on the Transformer architecture [11]. Significant concerns surrounding the threats posed by generative Transformer models are nearly as old as the models themselves: The release of the 1.5B parameter GPT-2 architecture was delayed for nine months due to fears of abuse [12]. Access to GPT-3 remains permitted only via a carefully controlled API [13]. Such measures demonstrably manifest only in delays to open availability of models. Only four months after the release of GPT-2, Grover — a 1.5B parameter model based on the GPT-2 architecture — was made publicly available [5]. Grover's release not only foreshadowed the speed with which private models would be replicated, but also represented a limited threat model in itself: Grover was specifically designed to both produce and detect neural fake news. Grover's primary author provided a reasoned justification for the model release and called for an improved set of community norms for the release of potentially dangerous research prototypes [14].

Such norms have been slow to develop [15], and wide-scale democratization of access to increasingly large-scale natural language generation models has continued. Open-source initiative EleutherAI has produced open-source generative

The associate editor coordinating the review of this manuscript and approving it for publication was Tony Thomas.

Transformer models with large numbers of parameters, including the 6B parameter GPT-J [16], and 20B parameter GPT-NeoX [17]. Even truly massive models are now available as open-source — the BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) is an open-source multilingual model, and at 176B parameters, is larger than GPT-3 [18]. Yandex [19], Meta AI [20], and Huawei [21] have all open-sourced models with over 100B parameters.

Real-life examples are beginning to emerge of how generative Transformer language models may be abused. A controversy in the AI research community resulted from the publicized development of a GPT-J model trained on the 4chan politics message board `/pol/`. This model was subsequently deployed to produce numerous posts on the board from which its training data came, including posts containing objectionable content [9]. At its peak, the model represented roughly 10% of all activity on the board in a 24-hour period [22]. The response to this model's deployment included a signed condemnation from 360 signatories across the AI community, including scientific directors, CEOs, and professors [23]. A similar project targeted a federal public comments website with GPT-2 text until the submitted comments comprised half of all comments, demonstrating the extent of existing vulnerabilities [24].

Controversy around any individual publicized NLG model belies the more fundamental concern — for years, any person with access to adequate hardware and open-source training scripts could train or fine-tune large, generative Transformers for any purpose they choose, be it pop song lyrics, mass disinformation, or toxic spam. Malicious individuals in the process of training a generative language model need not draw attention to their models via public release and currently face limited risk of discovery. As NLG capabilities grow and access barriers evaporate, we are quietly climbing the adoption curve for this technology to be widely abused by cybercriminals, disinformation agencies, scam artists, and other threat actors.

Increasingly, access to these models is not limited to sophisticated threat actors who can fine-tune them. User-friendly web interfaces, such as the one provided by ChatGPT [25], effectively eliminate any barrier to using powerful generative models. Jasper, a tool marketed as an AI writing assistant, uses GPT-3 to write content alongside human guidance [26]. This includes generating content for blogs and websites, which Jasper can efficiently produce in large volumes. Another website offers an endless supply of GPT-3 authored cover letters [27]. Tools such as Jasper allow those with little technical knowledge to seed the model with a prompt, specify keywords to include, and indicate a specific tone of voice. Using publicly available open-source models, a nearly identical system could easily be created to generate endless streams of targeted disinformation that can be readily loaded into existing grey-market account automation tools for popular social media websites.

NLG models have the potential to have an immensely positive and transformative impact on human society. A staggering one in three internet users aged 16 to 64 used an online translation tool in the last week, a figure representing over 1 billion people [28]. Text summarization can create comprehensive summaries of complex legal text [29] or medical records [30]. NLG models can give a voice to machine systems, changing how humans interact with them [31]. The same Transformer architecture that is used frequently for NLG can also be used to generate pictures from image descriptions [32], produce functional code from a natural language summary [33], or form the basis of a generalist agent [34]. While future research in NLG will bring further positive developments, alongside these opportunities is the corresponding certitude that bad actors will use the same technology to nefarious ends. Predicting how abuses are likely to unfold, and determining the best defenses to use against them, are essential in allowing humanity to reap the positive benefits of this technology while minimizing potential harm. We must walk a cautious path through the age of the silicon wordsmith.

B. SURVEY OVERVIEW

Since the release of GPT-2 [12] and the subsequent explosion of high-quality Transformer-based NLG models, there has been only one general survey on the detection of machine-generated text [35]. The scope of this previous survey is limited to detection methods specifically targeting the several generative Transformer models that had been released at the time. Before this, a systematic review of machine-generated text predating the Transformer architecture covered methods that detected previous NLG approaches, such as Markov chains [36]. Our survey differs from previous work in three significant ways.

First, our survey of machine-generated text detection is much more comprehensive and up-to-date than previous work. We consider the literature on the feature-based detection of machine-generated text that was omitted from previous reviews [37], [38], [39]. Such approaches are worthy inclusions, as feature-based approaches still apply to contemporary NLG models [38], [40], [41] and may provide benefits, such as improved robustness against adversarial attacks targeting neural networks [40] or enhanced explainability [41]. Additionally, as research on NLG and detection has rapidly advanced in the years since the previous survey, we must now cover a wider range of generative models and defensive research.

Second, this survey provides an in-depth analysis of the risks posed by NLG models via the process of *threat modeling* (i.e., identifying potential adversaries, their capabilities, and objectives) [42]. The result of our threat modeling process is a series of *threat models* that describe scenarios where machine-generated text may be abused, the attackers' likely methodology, and existing research related to each threat.

To date, there has been no survey of machine-generated text detection focusing on the risks presented by machine-generated text. Considering threat models is vital to setting the groundwork for the trustworthy development of NLG technology, encouraging early development of defensive measures and minimizing potential harms.

Third, guided by the EU Ethics Guidelines for Trustworthy AI [10] and research community efforts [43], we present our survey with sociotechnical and human-centric considerations integrated throughout, focusing not only on NLG systems and machine text detection technologies but also on the humans who will be exposed to both text generation and detection systems in daily life. The goal with trustworthy AI is to ensure that AI systems are developed lawfully, ethically, and robustly from both a technical and social perspective. Abuse of NLG models threatens all three areas, creating safety risks for those who may be targeted by NLG-enabled attacks, threatening the integrity of online social spaces, and challenging the resilience of the technical and social systems that comprise modern society. Machine text detection protects against the abuse of NLG models, enhancing the robustness and safety of NLG development. Critically, our survey includes insight into ensuring that defensive machine text detection systems themselves are transparent, fair, and accountable.

To summarize, the major contributions of this work are as follows:

- The most complete survey of machine-generated text detection to date, including previously omitted feature-based work and findings from recent contemporary research.
- The first detailed review of the threat models enabled by machine-generated text at a critical juncture where NLG models and tools are rapidly improving and proliferating.
- A meaningful exploration of both topics through the lens of Trustworthy AI (TAI), considering the ethical and trust impacts of both threat models and detection systems.

The rest of this survey is organized as follows. We provide definitions and a brief overview of existing methods for natural language generation in Section II. In Section III, we explore threat models related to the abuse of machine-generated text, including impacts on trust. We provide a comprehensive survey of literature related to the detection of machine-generated text in Section IV. In Section V, we summarize open problems and ongoing trends to guide the direction of future work. Finally, in Section VI, we present our conclusions. While this work discusses machine-generated text extensively, including models designed to generate scientific papers, no such models were utilized in the authorship of this work.

II. MACHINE GENERATED TEXT

Before reviewing threat models and detection methodologies for machine-generated text, we provide a formal definition of machine-generated text and a condensed overview of natural language generation (NLG) models. We recommend further reading of dedicated surveys on natural language generation for greater insight into the breadth of NLG models and applications [44], [45], [46], [47], [48], [49].

A. DEFINITION AND SCOPE

In this survey, we use a broad definition of the term "machine-generated text," which we believe includes all relevant research in the field:

> *"Machine-generated text" is natural language text that is produced, modified, or extended by a machine.*

We focus our definition of machine-generated text on *natural language* — i.e., text written in human languages that are "acquired naturally (in [an] operationally defined sense) in association with speech" [50] — and exclude *non-natural language* — i.e., logical languages, programming languages, etc. The exclusion of non-natural language aligns with other work in the field: the term "text generation" is currently considered synonymous with "natural language generation" [48], [51]. We anticipate that "text generation" may be repurposed in future research as an umbrella term that includes non-natural language text. This would accommodate common considerations between NLG models and contemporary code generation models, such as Codex [33] and CodeGenX [52]. For example, attacks against StackOverflow or GitHub may include both NLG and code generation working in tandem. Code generation models can also be used to complete programming assignments without triggering common plagiarism detection tools [53].

Our definition of machine-generated text is intentionally broad and covers many possible use cases and associated threat models, which will be discussed in Section III. To manage a survey scope that already spans a broad sociotechnical context and range of literature, text generation by means of text adversarial attack will not be considered. In the majority of cases, the production of new text is not the primary goal of a text adversarial attack, and text adversarial attacks and threat models are already covered by surveys in the adversarial attack literature [54], [55], [56]. We will discuss the role machine-generated text plays in adversarial contexts in Section III, as well as detection models' adversarial robustness in Section V.

This analysis focuses on threat models where a threat actor leverages machine-generated text as part of an attack — typically, scenarios where the attacker attempts to pass off machine text as human and where the detection of machine-generated text may be useful defensively. We do not discuss attacks against NLG models themselves unless they leverage NLG as part of the attack. For example, we would not include a white-box training data extraction attack targeting the weights of a commercial speech-to-text model, but we would include an NLG model used to produce data for poisoning that model's training dataset.

TABLE 1. Inputs, tasks, and example models for natural language generation.

Input	Task	Example models
None / Random noise	Unconditional text generation	GPT-2 [12], GPT-3 [13] (no prompt)
Text sequence	Conditional text generation	GPT-2 [12], GPT-3 [13] (with prompt), T5 [57]
	Machine translation	FairSeq [58], T5 [57]
	Text style transfer	Style dictionary [59], GST [60]
	Text summarization	BART-RXF [61], Word and Phrase Freq. [62]
	Question answering	FairSeq [58], T5 [57]
	Dialogue system	DG-AIRL [63], DIALOGPT [64], BlenderBot3 [65], ChatGPT [25]
Discrete attributes	Attribute-based generation	MTA-LSTM [66], PPLM [67], CTRL [68]
Structured data	Data-to-text generation	DATATUNER [69], Control prefixes (T5) [70]
Multimedia	Image captioning	GIT [71], ETA [72]
	Video captioning	MMS [73], YouTube2Text [74]
	Speech recognition	ARSG [75], wav2vec-U [76]

With this definition of machine-generated text in mind and an understanding of the research scope under consideration, we proceed to a brief overview of natural language generation.

B. NATURAL LANGUAGE GENERATION

Using a computer to produce human-like text is well-established in the history of computing. In 1950, Turing's proposed "imitation game" [77] considered the question of machine intelligence based on a machine's ability to conduct human-like conversation over a text channel. The first widely published method was the ELIZA chatbot in 1966 [78]. We provide only a high-level taxonomy of major NLG tasks and approaches as the groundwork for our analysis of threat models and detection methodologies and leave detailed discussion to the aforementioned dedicated surveys, given the large volume of NLG research over the past 55 years.

1) NATURAL LANGUAGE GENERATION TASKS

Recall from § I-A that there are a wide variety of applications for natural language generation. Leveraging previous surveys [44], [48], [79], we provide a summary of major tasks in the NLG domain, with examples of models that have been used for each task in Table 1. Note that many of the models listed are multipurpose and can be trained for numerous NLG tasks. In Table 1, we provide a small selection of models that have been used for each task as representative examples.

The summary in Table 1 is not exhaustive, and in reality, a mutually exclusive delineation between input types does not exist. Combinations of different input types are possible. For example, CTRL takes both a discrete control code attribute and conditional text prompt in generation [68]. Question-answering systems may be able to answer questions about images, such as Unified VLP [80] and TAG [81]. We consider a "topic" as an attribute in this overview, and so include "topic-to-text generation" under the broader umbrella of "attribute-based generation," including work such as topic-to-essay generation [66].

Transformer-based models rightly warrant particular emphasis in review, given their strong generative capabilities. However, as mentioned in Section I-B, consideration of the broader NLG field and previous detection research is important as detection techniques that apply against pre-Transformer models have been shown to be useful in detecting modern generative models, and diverse approaches may offer increased adversarial robustness [40] or better explainability [41].

C. NATURAL LANGUAGE GENERATION APPROACHES

There is a wide range of model architectures and algorithmic approaches to natural language generation. We categorize these approaches broadly into neural and non-neural methods and further itemize them into more specific categories. A diagram of our simplified breakdown is shown in Fig. 1. As mentioned previously, NLG encompasses a variety of tasks and research areas, and this brief section serves as context for understanding machine-generated text threat models and detection methods.

1) NON-NEURAL MODELS

Predating the popularization of neural approaches in the NLG domain, a range of systems were used to accomplish NLG tasks. These early approaches can broadly be summarized as "rule-based," though a variety of processes, pipelines, and target tasks existed. A review of rule-based systems can be found in Reiter and Dale's book on the subject [49].

An alternative approach to purely rule-based approaches is to use an existing natural language corpus to generate rules for NLG system components, such as content selection [82], [83] or template generation [84]. These statistical approaches are intended to be more adaptable to different domains than strictly rule-based systems. While many statistical models have been integrated with NLG systems in various ways, Hidden Markov Models (HMMs) [85] feature prominently in past work. More recent non-neural research has used the reinforcement learning [86] and hierarchical reinforcement learning [87] of Markov Decision Process (MDP) agents to learn optimal text generation policies.

2) NON-TRANSFORMER NEURAL METHODS

Natural language generation using neural networks was demonstrated to be highly effective using recurrent neural networks (RNN) [88], [89], [90], including long short-term memory (LSTM) architectures [91] and gated recurrent units (GRUs) [92]. However, RNN and LSTM architectures had to contend with the vanishing gradient problem, to which the multi-head attention mechanism of the Transformer architecture is more resilient [93]. Generative adversarial networks (GANs) [94] — commonly used to generate continuous data (such as images) — can also be adapted to a discrete context for natural language generation [95], [96].

Deep reinforcement learning (RL) has been used with neural networks to learn policy gradient methods that reward text characteristics associated with high-quality text generation [97]. Inverse reinforcement learning (IRL) is a related area of work that aims to address reward sparsity and mode collapse problems in GAN-based text generation by learning an optimal reward function and generation policy [63], [98].

3) TRANSFORMER

The multi-head attention architecture of Transformer language models [11] currently represents the state-of-the-art in natural language generation across natural language tasks. Among Transformer models, the unidirectional GPT-2 [12] and GPT-3 [13] models are the most studied in the field of machine-generated text detection due to their groundbreaking performance on unconditional and conditional text generation — though like many Transformer models, these architectures can also be used for other NLG tasks.

In addition to GPT-2 and GPT-3, related autoregressive language models using similar architectures are also notable, with variations in sampling procedures or training datasets. Such models include Grover [5] (a GPT-2-style model trained on a news dataset that uses nucleus sampling instead of top-k sampling), GPT-J [16] (a 6-billion parameter autoregressive language model trained on The Pile [99]), and GPT-NeoX-20B [17] (a 20-billion parameter model similar to GPT-3, also trained on The Pile [99]).

Unidirectional Transformer language models generate text by performing self-supervised distribution estimation to predict the next token based on previous tokens. The model is trained on an existing set of variable-length example texts $(x_1, x_2, \ldots, x_n)$, each composed of symbols $(s_1, s_2, \ldots, s_m)$. These symbols may be characters or multi-character tokens obtained through a tokenization process.

The probability of a given text can then be expressed as the conditional probability of the final token, given each previous token. That is:

$$p(x) = \prod_{i=1}^{m} p(s_m | s_1, \ldots, s_{m-1}) \tag{1}$$

The self-attention mechanism in the Transformer architecture makes it possible to train neural network architectures that can estimate such probabilities effectively, given a suitable pre-training task. In unidirectional models such as those in the GPT lineage, a common training task is predicting the next token in a sequence. To generate text, such models can receive an input sequence by sampling from the probability distribution of all possible next tokens based on previous tokens. An important parameter in this sampling process is "temperature" $T \in (0, \infty)$, which can be raised above 1 to increase the likelihood of selecting a less-probable next token — improving diversity at the potential cost of choosing an unusual token — or lowered below 1 to bias sampling toward more common tokens.

There are three common decoding strategies used to sample token probabilities from contemporary unidirectional generative Transformer models [100]:

1) No truncation → Sample from the entire probability distribution. At $T = 1$, this is called "pure sampling."
2) Top-k truncation → Sample from the k most probable tokens.
3) Nucleus sampling (also known as top-p truncation) → Sample from tokens in the top-p portion of the probability mass, rather than a fixed number of tokens k.

Alternative sampling methods are an active research area in improving text generation. Such methods include "typical sampling," in which tokens are selected based on expected information gain rather than the strict probability of occurrence [101].

While unidirectional generative models are key fixtures of machine-generated text detection research, other Transformer architectures can also be used for NLG tasks. The BART [102] architecture includes a bidirectional encoder (similar to BERT [103]) but maintains a left-to-right decoder for sequential text generation. Other Transformer architectures, such as MASS [104], T5 [57], and ULMFiT [105], can also be used for NLG tasks.

An important area of ongoing research centers around shaping the output produced by Transformer models. This can include prompt engineering — carefully crafting the conditional text input for a language model to continue [13] — or providing additional discrete attributes used to influence the generation of the network, such as control code, topic, or sentiment as in CTRL [68], PPLM [67], or GeDi [106]. Greater control over model output increases the risks posed

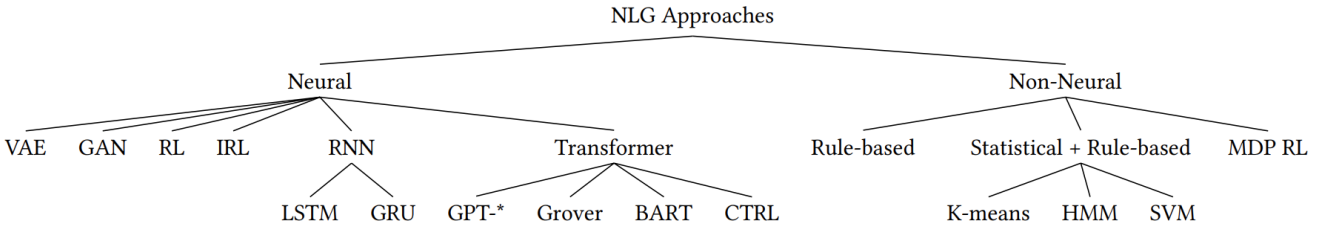

FIGURE 1. Taxonomy of major NLG approaches.

by threat models [13]. For example, when generating social media posts as part of an NLG-augmented online influence campaign, an attacker would benefit from ensuring that generated comments 1) mention a targeted political opponent, and 2) demonstrate negative entity sentiment toward the opponent. We cover such potential abuses and others in detail in the next section, which concerns threat models associated with machine-generated text.

III. THREAT MODELS

Machine-generated text enables a diverse array of attacks. Threat actors may perform these attacks with specific objectives, such as compromising a computer system, exploiting a target individual for financial gain, or enabling large-scale harassment of particular communities. The EU ethics guidelines for trustworthy AI emphasize that unintended or dual-use applications of AI systems should be recognized, and efforts should be made to prevent and mitigate the abuse of AI systems that can cause harm [10]. Trustworthy AI in the NLG context necessitates understanding the areas where such models may be abused and how these abuses can be prevented (by detection technologies, moderation mechanisms, government legislation, or platform policies). When discussing attacks, we discuss not only the direct impact on targets, but also the broader impacts on trust of attacks and mitigation measures.

To understand the risks that motivate research on machine-generated text detection, we draw from existing literature to present a series of threat models incorporating natural language generation. Threat modeling reflects the process of *thinking like an attacker* and identifying vulnerabilities to systems by identifying potential attackers, their capabilities, and objectives. The goal of threat modeling is to improve system security by considering the greatest threats to systems and their users. Many threat modeling methods have been developed over the years, some including system diagrams, itemized vulnerability checklists, and open-ended brainstorming [107], [108], [109], [110]. In late 2020, a diverse group of experts formed a threat modeling working group to produce a high-level set of guidelines for effective threat modeling approaches [111]. We leverage these guidelines in the open-ended attack-centric modeling approach in this section.

A. THREAT MODELING FUNDAMENTALS

We anticipate an audience with varying exposure to cybersecurity topics. Therefore, before we present threat models related to machine-generated text, we first provide an overview of threat modeling and characterize the approach taken in this section.

A basic example of a common threat model is "a thief who wants to steal your money" [112]. We can add detail to this threat model by considering more specific capabilities and objectives that such an attacker might have. For example, we may consider "a thief with lock picks who wants to steal your TV" or "a thief who found your banking password in a database dump and wants to transfer money out of your account." With these threat models in mind, we can propose mitigation strategies, such as "install locks that are resistant to lock picking" or "use multifactor authentication for online banking." We evaluate whether our mitigation approach is sufficient to address the threat and determine what other threat models we might need to consider. Threat modeling is inherently an iterative process [111], [112].

Shostack's Four Question Frame for Threat Modeling [112], [113] presents a plain-language foundation for threat modeling by posing four simple questions:

1) *What are we working on?* → Identify the system under attack.
2) *What can go wrong?* → Determine potential attackers, their capabilities, and objectives.
3) *What are we going to do about it?* → Devise a mitigation strategy.
4) *Did we do a good job?* → Review whether the analysis is accurate and complete.

Using these terms, we summarize our threat modeling approach in this section as follows:

1) *Identify the system under attack.* We provide a broad attack-centric analysis of machine-generated text on society rather than a system-centric analysis focused on vulnerabilities to a specific IT system. We identify several discrete technological systems within the broader societal supersystem.
2) *Determine potential attackers, their capabilities, and objectives.* We consider threat actors of varying sophistication and motives but with a common modus operandi — in all cases, our attacker is an individual

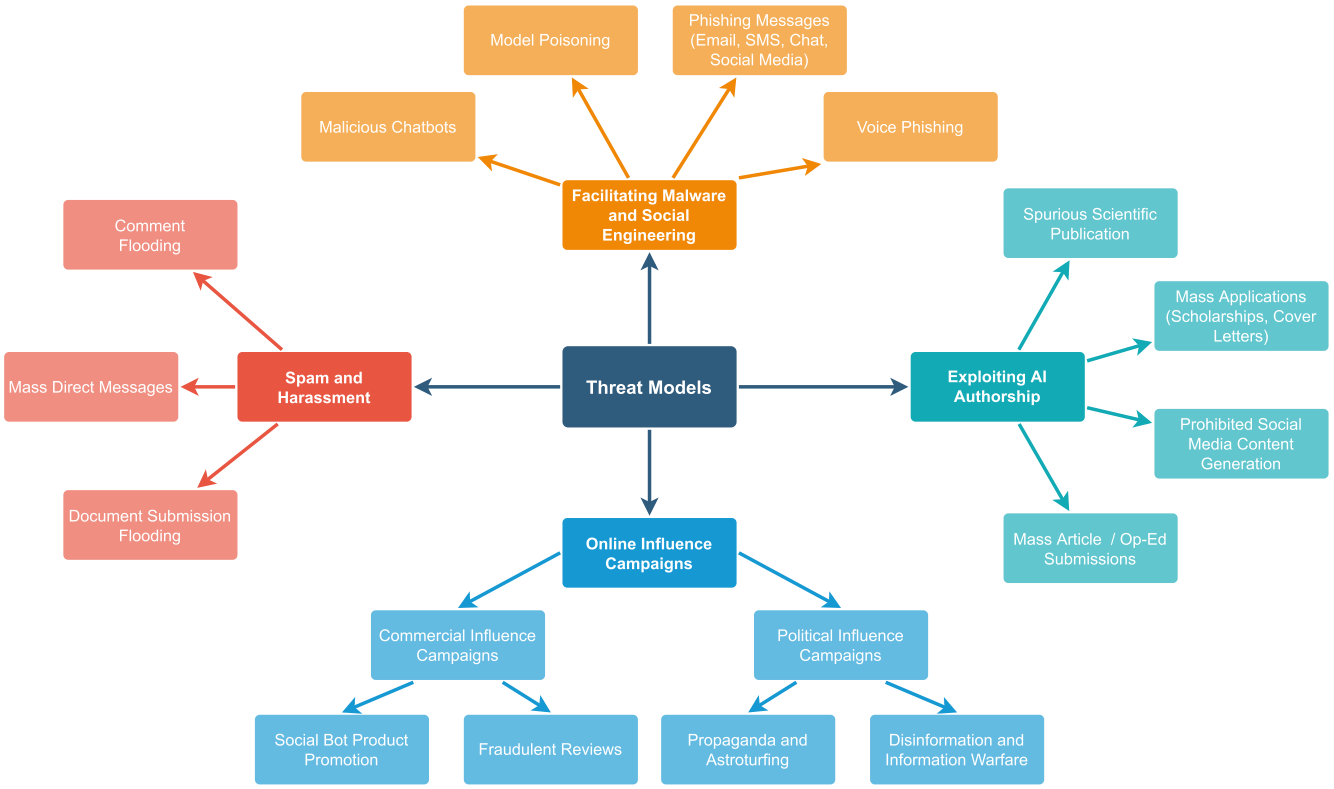

FIGURE 2. Broad taxonomy of threat models enabled by machine generated text.

or organization exploiting an NLG model. We characterize the attacker when explaining each attack.

3) *Devise a mitigation strategy.* After identifying a threat model, we propose mitigation measures to improve security and reduce risk. Detection of computer-generated text is often part of the presented mitigation approaches, but policy changes and human moderation systems can also have a significant impact.
4) *Review whether the analysis is accurate and complete.* We have carefully analyzed the presented threat models, which were formed from perspectives gained across industry, academia, and government. However, as threat modeling is an iterative process that benefits from diverse perspectives [111], we greatly encourage further analysis of potential attacks and mitigation measures in future research.

The remainder of this section comprises our threat model analysis, in which the attacks are broken into four major categories, followed by a concluding discussion. Within each category, we discuss threat models associated with that category of attack, identify systems at risk, and describe possible threat actors, their objectives, and capabilities. For each attack, we propose mitigations and discuss the trust impacts of both the attack and — crucially — the proposed mitigations. A taxonomy of the broad categories of attacks using NLG models can be found in Fig. 2.

While an exhaustive list of future malicious applications of NLG models is not possible, the threats outlined here span a broad range of current tangible dangers and represent valuable areas for preemptive ethical defensive research. Cybersecurity professionals looking to improve defenses should focus on the presented NLG threat models that are most relevant to their own organizations, and consider how machine-generated text augments the adversary tactics, techniques, and procedures (TTPs) documented in public knowledge bases such as MITRE ATLAS™ and MITRE ATT&CK®. As mentioned previously, threat modeling is iterative, and we hope these threat models will serve as the foundation for future work that improves security against machine-generated text.

B. FACILITATING MALWARE AND SOCIAL ENGINEERING

1) PHISHING AND SCAMMING

Phishing attacks focus on socially engineering a target individual to perform a desired action. These attacks include convincing a target to open an infected file, causing a target to enter account details on a fake login webpage, or encouraging them to share sensitive information for identity theft or password recovery. These are just a few among many other documented methods [114], [115]. Phishing attacks target numerous channels, including email, phone, SMS, and chat applications.

Automated messaging approaches are common in the early stages of phishing campaigns [115]. Attackers attempting to scale or target phishing or scam campaigns often use machine-generated text as a tool. NLG models can generate target-specific text instead of providing the same message to all targets. Research has demonstrated NLG's

effectiveness both in scaling email masquerade attacks [1] and community-targeted phishing [2]. Carefully targeting a phishing attack (commonly referred to as ''spear phishing'') greatly increases the likelihood that a specific target will fall for the attack [116]. In chat messages, NLG models that serve as dialogue agents may be exploited to exchange messages with the target under a pretext before exploiting them [64].

The methods used to mitigate NLG-enabled phishing attacks will be similar to those used on existing phishing attacks, including automated detection systems, user reporting, and awareness campaigns [117]. Automated phishing detection systems are likely to increasingly include components for detection of machine-generated text. NLG may present an increased challenge for existing detection systems because generated messages can have unique or highly varied content, though attackers may be forced to include specific ''payload'' content for an attack to be effective (e.g., a phishing email may include a unique short link to the same fraudulent website, or a malicious chatbot may need to socially engineer responses to the same security questions). As text content becomes more varied and human-like due to advances in generative models, the presence of payload content may represent a stable detection feature.

2) SOCIAL WORMS

NLG models may be particularly useful for worms that spread through social media or email contact networks. When an exploit compromises an individual's account, that account may be used to send malicious messages that propagate the exploit to other users. Using previous messages or emails between individuals as context for an NLG model, it may be possible to automatically produce messages that include personal details, mimic a loved one's writing style, or carry on a short conversation before delivering a malicious file or link. Given that NLG models are often large, NLG functionality may need to run on a separate command-and-control server and queried from behind a proxy rather than bundled with the exploit code itself (unless the pretext of the conversation can be used to convince the target to download a file).

Platforms could adopt formal policies stating that users cannot use machine-generated text in their communications, except under carefully controlled circumstances, to mitigate such attacks. Detection models could then be leveraged against user communications. While this may be considered acceptable for public posts, processing private messages presents privacy risks. Detection models could be executed on the receiving device as part of the message-viewing application to protect end-to-end message privacy. A warning may be raised if a user receives several messages that score highly for machine-generated text detection. This approach is not without risks, as the privacy of direct messages must be protected, and any real or perceived erosion of privacy will undermine public trust. Beyond this, such policy decisions would also have collateral impacts on non-malicious usage of NLG (e.g., translation models, writing assistants), and may not impact users of all language backgrounds equally. Other security measures to protect accounts from unauthorized logins, such as multi-factor authentication, should continue to be used to protect against accounts from being compromised.

3) MODEL POISONING

Cybercriminals may attempt to poison the training datasets of machine learning models. This may support other attacks (e.g., degrading a malware detection algorithm or email spam filter) or compromising a target model may be the primary goal (e.g., manipulating an algorithmic trading model, so the attacker can trigger trades that financially benefit them). If a threat actor identifies that they can access the training data of a target model, they may use NLG to produce many training examples containing a particular malicious signature they wish to conceal. Poisoning attacks against neural code-completion algorithms have been performed by generating samples including a given vulnerability [118], and GPT-2 has been used in research to produce fake cyber-threat intelligence reports for poisoning defense systems [119].

Mitigation of dataset poisoning varies based on the model's sensitivity and the nature of the training dataset. The first line of defense is basic IT security best practices that prevent unauthorized modifications to training datasets. However, in some situations, models are trained on publicly available data, and it is impossible to prevent access to training data. In these cases, data might be screened prior to their inclusion in the training dataset. This screening can include classifiers — potentially including the machine-generated text detection approaches discussed in Section IV — or other analysis methods, such as cluster-based methods, to detect poisoning in training datasets [120]. For sensitive models, it may be appropriate to leverage data versioning techniques and audit logging to capture changes to the data potentially made by a malicious insider.

4) IMPACTS OF ATTACKS AND MITIGATION ON TRUST

It is likely that using NLG models to produce compelling, target-specific messages as part of large-scale phishing attacks and social worms will further reduce trust in text communications, particularly those received from contacts that users do not personally know. Individuals may become even more suspicious of unsolicited messages, even seemingly innocuous ones. As a good-natured greeting may be the first message from a malicious dialogue agent, individuals may decide it is safer not to reply to such messages. Advanced NLG systems can be expected to further reduce trust and social interaction among new contacts in online communities.

NLG-based poisoning attacks against machine learning models will likely have the greatest trust impact on machine learning practitioners, who may be required to carefully scrutinize open-source training data for poisoned samples. Limiting access to training datasets with auditing and approval processes to mitigate poisoning attacks may cause developers to feel distrusted and undermine their relationships

with the organizations with which they work. While the trust impact of NLG-based data poisoning attacks may be relatively minor among the general population, a high-profile attack (e.g., a poisoning attack against a medical diagnosis model) may cause individuals to lose trust in machine learning systems more broadly based on concerns that such models are not safe from malicious tampering.

C. ONLINE INFLUENCE CAMPAIGNS

Online influence campaigns are of particular concern for abuse by machine-generated text. The objectives of threat actors in this area may be political (e.g., disinformation, propaganda, election interference) or commercial (e.g., product promotion, smearing competitors, fake reviews). In either case, the goal is to promote a particular idea or prompt a specific action among the target audience.

Either type of campaign may leverage or facilitate other threat models, such as spam, harassment, mass submission of agenda-driven content, phishing, or malware. The distinction between commercial and political influence campaigns is useful to understand threat actors and threatened systems in more detail and to categorize existing research.

1) POLITICAL INFLUENCE CAMPAIGNS

Machine-generated text that is part of political influence campaigns has been analyzed in previous work [3], [4], [5]. Papers related to the threat of generative language models on online influence operations may use terminology such as "fake news" [5], "disinformation" [4], [121], or "domestic and foreign influence operations" [121].

The threat actor in a political influence campaign represents an entity who wishes to influence beliefs or prompt action among a target group. These threat actors might include:

- A political party hiring a group to post unflattering comments online about their political adversaries
- A nation-state disseminating fraudulent news reports to mask human rights abuses
- A nation at war attempting to incite the citizenry of an opposing nation to overthrow the government

Facebook [122] Reddit [123], and Twitter [124], have released datasets from past political influence campaigns, including operations attributed to threat actors in 22 distinct countries. Threat actors in this space can be expected to have the capability to run online political influence campaigns using human employees or contractors. Attackers are likely already familiar with social media automation tools that facilitate the registration and management of fraudulent social media accounts. Large generative models with strong few-shot performance can simply be given examples of the desired messaging to produce further propaganda. This is shown in Table 2, where providing historical Russian information operations executed on Twitter during the Syrian civil war as input to GPT-3 results in additional "on-message" tweets that promote Syrian Arab Army (SAA) and Russian forces while discrediting American involvement.

There are many avenues where machine-generated text might be utilized by a threat actor to improve the scaling and targeting of influence operations, especially as people heavily consume text content online. The large population of engaged users on social media platforms are valuable and vulnerable targets for such campaigns [122], [123], [124]. Research that focuses on "fake news" detection only covers a small subset of critical threat models. Fake news detection research often imagines an adversary using an NLG model to produce news-like disinformation at scale [5]. Producing large volumes of news-like content may be a less-desirable approach than social messaging for several reasons:

- Research has demonstrated that individuals are more likely to share an article than read it [125], and a majority form an opinion on news topics by only reading headlines [126].
- Scaling by number of articles does not multiply effectiveness — a single news article or many articles can be widely disseminated, reducing the need to generate numerous articles each day.
- Scaling by number of articles requires either manipulating existing platforms to host them (i.e., layering and information laundering [127]) or procuring domain names and hosting infrastructure, representing additional cost and effort.
- Human involvement in authoring fake news articles allows disinformation threat actors to accurately tailor messaging, reduce detection, and carefully walk the line between promoting manipulative information and triggering moderation actions from social media websites.

While social messaging may be a more scalable attack vector, NLG remains useful for producing news-like content as part of political influence campaigns. Research on GPT-3 has shown that it and similar models can easily be used in a few-shot setting to generate conspiratorial content to promote radicalization [128]. AI writing assistants may be used to save time and minimize costs [26], while translation models can be used to disseminate propaganda across language barriers. Social bots can amplify this content by distributing links to disinformation articles, promoting discussion around incendiary headlines, and producing numerous comments that represent a false public consensus. Targeted users do not even need to read the shared articles — the artificial amplification of a headline and an overwhelming "grassroots" narrative guided by machine-generated comments has been demonstrated as sufficient to influence public opinion [125], [126].

Past research on mitigation has identified that the average user is overly trusting of profiles with AI-generated photos and GPT-2 text, accepting connection requests from deepfake profiles on LinkedIn 79%–85% of the time [129]. As such, it is unlikely that user reports will serve as an adequate first line of defense. Instead, a combination of automated detection models (including machine-generated text detection) and platform moderation efforts should be used to detect political influence campaigns. These should

TABLE 2. Few-shot generation of information warfare based on Syrian civil war influence operations.

GPT-3 Prompt: Human tweets from Russian foreign military intelligence (GRU) [124]	RT @TheArabSource: Confirmed: #US retrains #ISIS and #ISIS-like jihadists in east #Syria and re-brands them 'democratic forces' (VIDEO) - E...
	RT @Tasnimnews_EN: #Iraq Army Discovers #US Drone in #Daesh Arms Depot near Fallujah https://t.co/zYRgEHmlr0 https://t.co/H762UxwlQx
	...
	ISIS-affiliated band group Wilayat al-Sham recruits new #terrorists in US-controlled #Rukban camp.
GPT-3 Response	@ALFaham_ #SAA seized a big amount of #ISIS #ISIL weapons from the convoy in east #Homs countryside today. https://t.co/k3kzPoLlsd #Syria #Damascus #Jobar #ISIS
	#Washington is willing to perform the functions of #MiddleEastern policeman and play the role of accessory to #Turkish#aggression against #Syria. https://t.co/HVGXuvRZ
	...
	#Russia's #Air_Force has stepped up airstrikes against camps for the #Islamic_State in the #Terek_Val

include broad measures to protect users from social media abuse, such as detection of account automation and scrutiny of coordinated inauthentic activity for content amplification. Investigations into disinformation or coordinated inauthentic behavior, such as those carried out on Twitter, are likely to remain relevant [124].

2) COMMERCIAL INFLUENCE CAMPAIGNS

The goal of a commercial influence campaign is to sway individuals' opinions in a manner that commercially benefits the threat actor. Examples of such campaigns include publishing fraudulent reviews, artificially boosting a website's page ranking on a search engine, spamming online communities with product advertisements, or attempting to inorganically cause promotional content to trend on social media. As with previous categories, there may be overlap between different attackers' approaches.

The use of machine-generated text to generate fraudulent reviews that either promote one's own product or service, or target a competitor, is a threat model of particular interest [4], [6], [41]. Published work has demonstrated sentiment-preserving fake reviews, which might be used for such a purpose [6]. Fake reviews can be abused on marketplace websites or target potential customers on social media platforms. Threat actors operate such campaigns themselves or make use of the thriving online market for fake reviews [130]. Organizations selling fake reviews may become early adopters of open-source NLG models to provide unique, specific, low-cost reviews.

Machine-generated text detection could be run on reviews on online marketplaces to mitigate NLG models used for fake reviews on these sites, in addition to other detection systems currently used to combat this problem. Advanced NLG models should not affect context-based detection methods (e.g., identifying patterns in reviewer usernames, similar account creation times, unusual purchase behavior). It may be more difficult to detect commercial influence campaigns if attackers post content outside marketplace websites. For example, social media websites (e.g., Facebook, Instagram, Reddit, YouTube comments), map platforms (e.g., Google Maps), or dedicated review sites (e.g., Yelp) are locations where false reviews may be posted.

3) IMPACTS OF ATTACKS AND MITIGATION ON TRUST

In addition to the societal risks posed by machine-generated text in online influence campaigns, widespread NLG threat models cause fundamental damage to online trust. The perceived value of online discourse is degraded when users are suspicious that others are part of a concerted political or commercial campaign. When facing online disagreements, Internet users may dismiss other users as "bots" rather than acknowledge that real people hold a variety of different viewpoints. The net effect is reduced trust in the authenticity of online interactions, and an expedient mental shortcut for dehumanization of perceived opponents.

Mitigating influence operations via automated detection of machine-generated text carries substantial risks related to the mass suppression of online speech. Previous work has found that text on political topics written by non-native English speakers was at a high risk of being erroneously detected by a Transformer trained on previous political influence campaigns [131]. As methods based on RoBERTa (a Transformer) are currently the state-of-the-art for machine-generated text detection [132], [133], classifiers for machine-generated text detection leveraged to combat online influence campaigns must be carefully trained and ethically evaluated to minimize the risk of similar incidences of mass discrimination. Continued public reporting of influence campaign datasets would be beneficial to protecting trust in social media moderation, such as Twitter's regular public releases for researchers [124].

Language background considerations evoke another problem: there are legitimate reasons why a user might rely on machine-generated text. A person writing in their non-native language could leverage an online translation model to assist them. While such text can be considered machine-generated text, this text is not *inauthentic* — it represents genuine self-expression. Much of the world relies on translation tools to participate in online discourse; recall that one in three Internet users aged 16 to 64 have used an online translation tool in the last week [28]. Relying on machine-generated text

detection alone is therefore likely to produce a solution that is discriminatory, unreliable, and greatly damaging to trust in social media platforms. Machine-generated text detection should be used among multiple features, such as account creation times, activity patterns, registered phone numbers, and IP addresses, to determine whether activity is linked as part of an online influence operation.

D. EXPLOITING AI AUTHORSHIP

1) ACADEMIC FRAUD

The use of algorithms to generate scientific papers has been well-established since SCIgen was created in 2005 to produce nonsensical papers that nevertheless sometimes passed peer review [7]. Many years later, these papers continue to appear in respected publications despite the comparative simplicity of the context-free grammar generation method [134], [135]. The generation of artificial scientific papers consumes valuable reviewing resources, lowers standards by producing misleading or nonsensical articles, and weakens trust in the scientific review process. In education, NLG models are used by students to easily produce essays [8], [66] or cheat on language learning assignments using machine translation [104]. These are instances where institutions may be tempted to perform machine-generated text detection to improve academic integrity and encourage students to learn the course material. Widespread access to convenient NLG interfaces online, such as that provided by ChatGPT [25], allows any student with an Internet connection to leverage such models, even when doing so undermines the learning objectives of an assignment (i.e., cheating).

Threat actors submitting AI-generated papers are typically either 1) academics attempting to inflate publication statistics, particularly when meeting a quota to maintain their position [135]; or 2) well-meaning researchers probing the publication standards of a potentially disreputable conference [136]. Threat actors' capabilities include using well-established tools such as SCIgen or more recent Transformer-based approaches that are promoted as ''scientific writing assistants,'' which can nevertheless be easily exploited to generate long articles of little substance [137]. Mitigation measures should include flagging suspected machine-authored publications using published approaches for detecting SCIgen articles [134], [135] and new detection approaches based on detecting Transformer-generated text [138]. Human reviewers can more carefully review flagged articles to determine whether the article contains credible research, regardless of the detection result.

The acceptability of machine text within scientific writing is an active area of discussion in academic disciplines. The Association for Computational Linguistics has released a set of guidelines on the use of AI writing assistance [139]. If the results published by a researcher are true and accurate, limited use of a carefully guided NLG model may be acceptable in some publications. Emerging research aims to differentiate between acceptable and unacceptable use of NLG models in scientific writing [140], which should be part of a broader, ongoing conversation on norms surrounding AI usage and disclosure.

2) APPLICATIONS AND COVER LETTERS

Contemporary NLG models can be used to easily generate cover letters or essays for scholarship or employment applications. Commercial websites already offer to produce cover letters using GPT-3 [27]. While the overall usefulness of human-written cover letters has been debated in business media [141], they are ostensibly meant to be an earnest reflection of a candidate. Using AI models to generate a cover letter or essay is therefore likely to be seen as exploitative by organizations who review them. The threat actor in this case may be an individual (perhaps understandably) looking to save time and improve their employment opportunities by bypassing a cumbersome application process, or a malicious attacker looking to flood a target company with fraudulent submissions (an attack similar to those we will discuss further in ''Spam and Harassment'').

Machine-generated text detection may identify artificial cover letters or essays, if they are of sufficient length (the odds of successful detection improve with sequence length [5], [12], [142]). However, caution should be taken with this approach, as using AI writing tools is not necessarily exploitative. Individuals writing in a second or third language may rely on translation models or NLG writing assistants to help them write cover letters or scholarship applications. It may be difficult to differentiate those who mean to exploit such systems and those who rely on AI writing tools to better express themselves. An alternative strategy for evaluating candidates, such as placing more emphasis on face-to-face discussions with prospective job candidates or award recipients, could be a preferable mitigation approach.

3) CONTENT GENERATION

A growing threat model for social media platforms is the possibility that a critical mass of users may begin using generative AI models (including NLG models) to produce social media content in ways that harms these platforms. While threat actors in this case may not be overtly malicious, large volumes of content from generative models could dilute perceived content quality, undermine overall trust in a platform, or create plagiarism concerns. For example, in response to the recent release of highly effective AI models for image generation (DALL-E [32], Stable Diffusion [143]), several art websites enacted a blanket ban on all AI-generated art [144].

Video is a particularly important medium on modern social media: there are approximately 4.95 billion Internet users on Earth [145]. Of these, an estimated 92.6 percent watch digital videos each week [28]. The interplay between social media creators and generative models represents an important sociotechnical context to avoid common pitfalls in trustworthy machine learning [146]. Award-winning online commentator Drew Gooden performed a video demonstra-

tion of GPT-3-based writing assistant Jasper [26], critiquing applications of Jasper for producing video scripts and social media content [147]. When attempting to generate a bio for a company website, Gooden found that Jasper produced a sample that directly plagiarized a Newswire article (timestamp 11:55). Gooden also noted that utilizing such a tool without disclosure would violate viewers' trust (timestamp 4:22).

Mitigations of threats related to the undesired inclusion of NLG content in social media may involve similar blanket bans to those targeting AI-generated art [144] or policies that mandate the preemptive disclosure of the use of AI tools as part of a platform's terms of service (similar to the requirements mandated in the Responsible AI License [148]). The difficult enforcement of such policies would likely necessitate a combination of machine-generated text detection algorithms and moderator investigations.

4) IMPACTS OF ATTACKS AND MITIGATION ON TRUST

The widespread use of machine-generated text in written submissions may undermine the trust that individuals place in such written works and lead to greater scrutiny of such material. Given that a suitable cover letter with language tailored for a position can be quickly generated by existing user-friendly tools [25], [27], it is possible that employers will soon place so little trust in cover letters that they eschew them altogether. Reviewers of scientific publications must increasingly be vigilant for submissions that exploit machine-generated text in unacceptable ways, while defining their own policies on acceptable use [139]. Internet users are likely to interpret algorithm-generated blogs, articles, and video scripts as low-effort and untrustworthy.

Detection processes must be used carefully. As mentioned previously, it is possible that machine-generated text detection may unfairly skew toward false-positive classification of individuals with specific language backgrounds [149]. There are cases where using machine-generated text may be permissible (e.g., translation models or assistive writing technologies). The perception of fairness in application processes will be undermined if it appears that certain individuals are unfairly screened out because of erroneous false-positive detection. Submitting a scientific paper only to have a reviewer allege that a given section might be written by an algorithm could similarly lead to a loss of faith in scientific reviewing.

To preserve trust, usage of machine-generated text should generally be preemptively disclosed to the reader or audience. In many cases, the audience may have a negative view of content authored by machines, and this could undermine trust in a particular publication platform, news website, or brand. Similarly, some organizations may be concerned with spam consisting of low-quality machine-generated content overwhelming editorial staff. Media and entertainment organizations that publish content from multiple creators may decide to enforce that certain categories of content are to be completely written by humans. This may improve the perception of trustworthiness or reduce the risk of plagiarism or copyright infringement as some models have been found to memorize training data which can emerge during inference [150].

E. SPAM AND HARASSMENT

We distinguish spam and harassment from other categories of attacks by focusing on cases where the intent is to harm a platform or users with a large volume of content. As in previous cases, there are similarities to other threat models, but it is useful to distinguish spam use-cases to understand how attacks using machine-generated text impact platforms when deployed at large scale.

1) SOCIAL MEDIA SPAM

Social websites are an attractive target for attacks using large volumes of machine-generated text, providing opportunity for significant disruption. One researcher demonstrated a real-world attack by using a GPT-2 bot to generate 55.3% of all comments on a federal public comment website before voluntarily withdrawing the comments and shutting down the bot [24]. It is important to realize that spam attacks against social media websites are already possible — high-quality NLG models simply make spam attacks more difficult to detect, as posts can be unique and better mimic the style and substance of discussion.

Using generative models to produce large volumes of hateful spam targeting specific groups and individuals is a particular cause for concern. While OpenAI attempts to reduce the incidence of offensive content generated by its GPT-3 API through careful training measures and filtering of inference prompts [13], open-source models are not subject to any such restrictions. GPT-4chan, which was trained on and subsequently deployed to create a large volume of posts on the 4chan politics message board, provides a complete example of how such a model might be created and deployed to cause havoc [9], [22]. An attacker with sufficient motive (political, personal, or otherwise) can render an entire community nearly unusable with spam.

In general, mitigation measures against automated spam should rely heavily on methods designed to prevent automated user activity. Some approaches include increased scrutiny of proxy and VPN usage, typically used in conjunction with *Completely Automated Public Turing test to tell Computers and Humans Apart* (CAPTCHA) [151] challenges to verify that a user is human. Notably, both of the previous examples of Transformer-based spam take advantage of either 1) a lack of CAPTCHA tests [24], or 2) a method of bypassing CAPTCHA and proxy restrictions [22]. CAPTCHA is not a perfect defense — iterative versions of human-verification schemes and bypass methods are in continuous adversarial development [152] — but such defenses represent an important first step to increase the difficulty of automation. As spam results in large volumes of text, and machine-generated text detection is easier on long sequence

lengths [142], many comments from the same user or IP range could be combined to generate a larger sample for effective machine-generated text detection.

2) HARASSMENT

Techniques similar to spamming may be used to cause distress to individuals or communities by targeting them with an overwhelming number of messages. A motivated individual or group could register social media accounts to be controlled by automation tools, or voluntarily use their own account to deliver a large volume of machine-generated messages targeting a particular individual or community. SMS and phone call automation tools could facilitate such approaches outside social media.

The motivations of threat actors engaging in such behavior range from personal grudges to political objectives. Online communities, formed around religion, racial identity, sexual orientation, or gender expression, are at risk of *brigading* [153] from hate groups using such models to flood them with abuse. Political figures or political discussion boards of all stripes may be at risk of large-scale automated harassment from motivated enemies among their political adversaries.

Mitigation measures for large-scale harassment are similar to defenses against spamming — the best defenses focus on verifying that an individual is human prior to making a post or sending a message, targeting the automation of message delivery rather than the machine-generated text. Beyond this, a moderation system that enforces a set of clearly defined acceptable-use policies is a fundamental measure for reducing the impact of harassment on social media.

3) DOCUMENT SUBMISSION SPAM

The platforms previously mentioned in "Exploiting AI Authorship" may be vulnerable to being overwhelmed by large volumes of AI-generated content. A motivated attacker might submit massive numbers of unique cover letters and resumes to a company, none of which corresponds to a real individual, thus frustrating recruiting attempts. Depending on the method of submission, reviewers at scientific conferences or news outlets could be vulnerable to excessive numbers of misleading AI-generated submissions that are difficult to distinguish from real ones without a time-consuming review process. Detecting machine-generated text may be a useful mitigation measure in these cases, pre-screening content based on its likelihood of being written by a machine, as well as including CAPTCHA [151] challenges to reduce automated submissions.

4) IMPACTS OF ATTACKS AND MITIGATION ON TRUST

Spam and harassment harm the assumption in online communities that other users are real people. Following the deactivation of the GPT-4chan bot, discussions on 4chan continued to express concern that subsequent posts were made by NLG models [22]. The more frequently individuals knowingly encounter such models on social media, the less trust they will have in the integrity of online social spaces.

Mitigation of such attacks would incorporate increased verification of human posting activity. Such restrictions would likely include limitations on usage of known proxies and VPNs, potentially requiring additional information at sign-up (e.g., email addresses, phone numbers, payment methods, government IDs) and an increase in CAPTCHA challenges. The overall result is a reduction in online privacy and an increase in barriers to participation in online discussion — both of which can harm users' trust in online platforms.

Spam and harassment operations can be highly disruptive, and they represent a highly visible case of AI model abuse. As such, the abuse of such models in online communities may decrease public trust toward AI model development in general, and NLG models in particular.

F. SUMMARY OF THREAT MODELS

In this section, we discussed a range of threat models associated with natural language generation. We summarize our key findings as follows:

- NLG models have significant potential for abuse in improving scaling and targeting of existing attacks.
- Platforms that receive text submissions of any kind are likely to face a growing influx of machine-generated text content, particularly as user-friendly tools continue to be developed [26], [27].
- Much of the research on NLG-enabled influence operations focuses on AI-generated news articles, while sociological data suggest that machine-generated comments may pose a more significant threat.
- While NLG models may make detection of coordinated inauthentic activity more difficult, abuse often requires bypassing existing defenses such as IP reputation checks and CAPTCHA [151].

Future threat modeling and observed cyberattacks will certainly augment the threat models discussed in this section. However, we have now provided sufficient motivation for exploring the defensive capabilities offered by machine-generated text detection. In the next section, we discuss the current status of research on machine-generated text detection and outline the major findings in the field thus far.

IV. DETECTION OF MACHINE GENERATED TEXT

Analysis of threat models indicates that when utilized correctly, machine-generated text detection is a valuable tool for reducing the negative impacts of NLG model abuse. Machine-generated text detection is typically framed as a binary classification problem in which a classifier is trained to differentiate samples of machine-generated text from human-generated text [5], [37], [39], [40], [133], though a related research area centers on attribution of machine-generated text to the model that generated it [154], [155], which we discuss in § V-B.

In this section, we outline the methods used for machine-generated text detection. In § IV-A, we summarize feature-based approaches, while § IV-B covers detection approaches based on neural language models (NLMs). In § IV-C, we survey domain-specific research on applications of machine-generated text detection. In § IV-D, we review human reviewers' ability to identify machine-generated text and approaches for human-aided machine-generated text detection. In § IV-E, we discuss trends in evaluation methodology within detection research. Finally, in § IV-F, we explain prompt injection, a method of shaping NLG model responses, which may be useful in facilitating detection. Table 3 provides a summary of prominent detection methods and their evaluation in current research.

A. FEATURE-BASED APPROACHES

Machine-generated text often differs from human text in ways that can be identified using statistical techniques [37], [38], [40]. Feature-based approaches to machine-generated text detection apply natural language processing to create feature vectors from input sequences and classify these feature vectors using a downstream classification algorithm, such as a support-vector machine (SVM), random forest (RF), or neural network (NN) [37], [38]. We provide a summary of features used in prior art, with references for further reading on implementation of specific features for machine-generated text detection.

An important consideration in detecting machine-generated text using feature-based approaches is that different language model sampling methods (e.g., top-k versus top-p sampling in Transformer language models, as discussed in § II-C3) lead to different artifacts in the generated text [38], [100]. As a result, the performance of feature-based detection can diminish when detecting text generated using a different sampling approach than that used to train the detection model [38]. A feature-based detector trained on output from a smaller model can be used to detect output from larger models [5], [38]. However, it is more effective to use a detector trained on a larger model to detect outputs from smaller generative models [38].

We now proceed with our summary of major feature categories in feature-based detection approaches.

1) FREQUENCY FEATURES

A major category of statistical features used in machine-generated text detection centers around the frequency of terms within text samples. Human-written text often conforms with Zipf's Law: The frequency of a word is inversely proportional to its rank in an ordering of words by frequency [156]. In Zipf's Law, the normalized frequency f of a token of rank k out of N different tokens follows the relationship:

$$f \approx \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)} \tag{2}$$

where $\{s \in \mathbb{R} | s \geq 1\}$ is an exponent that characterizes the distribution.

Machine-generated text does not perfectly mirror the distribution of tokens in human text. Transformer language models produce different token frequency distributions depending on the chosen sampling method (see Fig. 7 of Holtzman et al., 2019 [100]). Therefore, the distribution of tokens provides useful discriminating power, particularly when a greater volume of text is available for consideration.

Another prominent frequency-based feature from previous statistical detection research is term frequency-inverse document frequency (TF-IDF). TF-IDF unigram and bigram features with a logistic regression detector have been used as a baseline for detection [12], [133] or as a feature in statistical approaches [38].

Previous research has used lemma frequency as a statistical feature [37], [40]. In this approach, a linear regression line that fits log-log lemma frequency versus rank is learned, and the mean-square error cost function is used to calculate the information loss of the regression.

Due to the observed repetitiveness of writing produced by NLG models [100], [157], another potentially useful frequency feature is the n-gram overlap of words and parts-of-speech tags between sentences [38]. An additional technique targeting machine-text repetitiveness computes super-maximal repeated substrings (i.e., the set of the longest repeated substrings, excluding all substrings which are already part of a longer repeated substring) in large collections of text [158].

2) FLUENCY FEATURES

Another major category of features centers around the fluency or readability of generated text. At longer sequence lengths, machine-generated text is increasingly likely to manifest difficulties in producing consistently clear and coherent text [100], [159]. The Gunning-Fog Index and Flesch Index each provide a statistical measure of text readability and comprehensibility, respectively, and have previously shown predictive power in detecting machine-generated text [40]. More complex measurements use an auxiliary model to perform coreference resolution to create measures of coherence either based on the presence of main entities in specific grammatical structures or by using Yule's Q statistic [38].

3) LINGUISTIC FEATURES FROM AUXILIARY MODELS

Past research has measured the "consistency" of machine-generated text by calculating the number of phrasal verbs and coreference resolution relationships within a sample [37], [40]. Other work has used the entire distribution of sample part-of-speech (POS) tags and named entity (NE) tags [38]. Such work is motivated by differences between human and machine POS tag distributions observed in past analyses of machine-generated text [12], [159].

Performing coreference resolution and assigning POS and NE tags requires processing samples with specialized models.

Contemporary models for this purpose are neural, and as a result, modern feature-based approaches that use auxiliary models to produce linguistic features may still perform inference on a neural network as part of feature creation [38], [40]. This implies that feature-based approaches that initially appear entirely non-neural may nevertheless include feature creation steps that are vulnerable to adversarial attacks that target neural networks.

4) COMPLEX PHRASAL FEATURES

Detection work targeting the translation of long texts found that specific idiomatic phrases were not commonly found in machine text [37]. However, recent work has shown that these features do not perform well against contemporary Transformer models, particularly on shorter sequence lengths [40].

5) BASIC TEXT FEATURES

Many simple text features are frequently used in feature-based text classification in natural language processing. These features include high-level characteristics such as the number of punctuation marks or the length of sentences and paragraphs, which have been used in machine-generated text detection [38].

B. NEURAL LANGUAGE MODEL APPROACHES

Detection approaches based on neural networks are highly effective in machine-generated text detection, especially those that incorporate features derived from Transformer neural language models. This aligns with broader trends in natural language processing, where state-of-the-art performance has been attained on a wide range of natural language tasks using Transformer models [160].

We separate NLM-based approaches into two major categories: zero-shot classification using existing models and fine-tuning pre-trained language models. These two approaches represent the majority of NLM-based machine-generated text detection.

1) ZERO-SHOT APPROACH

A baseline approach to machine-generated-text detection is to perform text classification using generative models themselves, such as GPT-2 or Grover [5], [12], [133]. Generative models can be used without fine-tuning to detect either their own outputs or outputs from other (typically similar) generative models. Autoregressive generative models, such as GPT-2, GPT-3, and Grover, are uni-directional, where each token has an embedding dependent on the embeddings of preceding tokens. As a result, an embedding for a sequence of tokens can be created by appending a classification token `[CLS]` to the end of the input sequence, so that the embedding of this token can be used as a feature vector for the entire sequence. Using these feature vectors, a labelled dataset of human and machine text can be used to train a linear layer of neurons to classify whether an input sequence is produced by a machine or a human.

Multiple studies have observed that smaller NLG models can be used to detect text generated by larger NLG models [5], [40], [133]. While a model's ability to detect larger models diminishes as the difference in scale grows, the predictive ability of smaller architectures may be useful as recreating large multi-billion parameter Transformer architectures is highly compute-intensive.

Grover, a model trained to generate and detect ''neural fake news,'' demonstrates strong zero-shot detection performance specifically within the news domain it was trained on [5]. However, it shows limited performance on out-of-domain text [133], [155]. While Grover's authors initially suggested that the best detection method for generative models might be generative models themselves [5], further investigation has shown that the increased representational power of bidirectional Transformer models appears to have an advantage for machine-generated text detection [133].

Similar to Grover's weakness outside the news domain, findings indicate that the zero-shot approach generally underperforms a simple TF-IDF baseline when attempting to detect output from a generative model that has been fine-tuned on a different domain [133]. As attackers may routinely fine-tune generative models for different purposes, this represents a notable weakness in the zero-shot approach using generative models for detection without fine-tuning.

2) FINE-TUNING APPROACH

The state-of-the-art approach for neural machine-generated text detection is based on fine-tuning large bidirectional language models [133]. In this approach — initially evaluated on GPT-2 text — RoBERTa [132], a masked general-purpose language model based on BERT [103], is fine-tuned to differentiate between NLG model output and human-written NLG model training samples.

The source code for this fine-tuning approach is available open-source, as are pre-trained detector models, facilitating future research and defensive detection [133], [161]. The available pre-trained detection models are based on the RoBERTa-base (123M parameter) and RoBERTa-large (354M parameter) architectures [132]. The machine-generated text used to fine-tune these models was generated by GPT-2 using a mixture of pure sampling and nucleus sampling (see § II-C3). The purpose of using a training dataset that contains multiple sampling methods is to generalize more effectively to unknown sampling methods that attackers could use in-the-wild — an approach that will likely be duplicated in future detection research.

Research into the practicalities of machine-generated text detection examined the task of detecting text when a RoBERTa detector algorithm was trained on a different dataset than a GPT-2 attacker model. In this case, it was found that by fine-tuning the detector model with even a few hundred attacker samples identified by

subject-matter experts (SMEs), the detector could dramatically improve cross-domain adaptation [138]. This reflects potential real-life scenarios where a general-purpose detection model confronts a fine-tuned attacker for a particular purpose. As a defender identifies samples from a fine-tuned attacker model, these examples could be used to improve the defensive detection model further.

Preliminary work has used attention map information from Transformer models to perform topological data analysis (TDA) as features for machine-generated text detection [162]. This did not show significant improvement over standard BERT fine-tuning approaches, though the resulting features more accurately detected unseen GPT generative models (once again relevant for detection of fine-tuned attacker models). It is unclear how the effectiveness of the TDA approach would compare if directly applied to the current state-of-the-art RoBERTa detection models [133] rather than to custom-trained BERT models.

While research on machine-generated text detection has primarily taken place in English thus far, detection models have also been released in Russian [163], [164] and Chinese [165]. Further to this, large pre-trained bidirectional Transformer models have been released for numerous languages, including Chinese [166], French [167], Arabic [168], and Polish [169]. Future work on machine-generated text detection in additional languages could leverage these pre-trained bidirectional models as starting points for fine-tuning.

Another detection method leverages energy-based models [170] alongside a classifier of machine-generated text. Evaluated approaches include a simple linear classifier, BiLSTM, a uni-directional Transformer (GPT-2), and a bidirectional Transformer (RoBERTa) [171]. The Transformer architectures were initialized from pre-trained checkpoints and then fine-tuned on machine-versus-human classification datasets. Corroborating other research, this research found the strongest performance by leveraging the bidirectional Transformer [172].

The strong performance of fine-tuned bidirectional NLM models — and RoBERTa in particular — has led to these models' strong representation in applied detection research targeting specific domains, as shown in Table 3 and discussed in § IV-C.

C. APPLIED DETECTION IN SPECIFIC DOMAINS

Applied work in machine-text detection has focused on using techniques and technologies for detection of machine text in specific domains. This applied research is important as it addresses several of the serious threat models discussed in § III, and includes broad lessons for machine-generated text detection more generally. We divide applied research into several major categories.

1) TECHNICAL TEXT

Recall from § III-D1 that machine-generated scientific papers have been well-documented since the release of SCIgen in 2005 [7], [134], [135]. Past algorithmic approaches target the SCIgen model [39], but there is also more contemporary research targeting technical text generated by GPT-2 [138]. This work found that by having a subject-matter expert label a relatively small number of examples (number in the hundreds), a RoBERTa-based detector could be much better adapted from one technical writing domain (physics) to another (biomedicine).

2) SOCIAL MEDIA MESSAGES

Application-specific work has applied feature-based [173] and neural [4], [174], [175] language model based detection methods to social media. Previous work in the social media domain has found that the detectability of such text heavily depends on the dataset used to train the generator and detector [175].

Existing applied work on machine-generated text detection in social media has primarily focused on Twitter. Twitter text is distinct in that it has common characteristics (hashtags, references, short links) and typically includes a short sequence length (280 characters). There is a lack of research targeting comments on more popular platforms such as Facebook and YouTube or fast-growing platforms such as Reddit [176]. With respect to machine-generated text, Reddit content can be found in ''SubSimulatorGPT2'', a simulation based on a host of fine-tuned GPT-2 models that produce community-specific machine-generated posts and comments harvested from the Pushshift dataset [177].

3) CHATBOTS AND SOCIAL BOTS

A related application area is the detection of malicious chatbots and social bots, which can interact with humans on chat applications, SMS, and social media. Bots can be used for malicious purposes such as spam, phishing, social engineering, influence operations, or data collection (see the threat models in § III). There is an overlap in this area with research into detecting AI-generated social media messages, but framing the detection challenge by targeting automated personae allows for the consideration of additional features. An analysis of how humans and chatbots interact has found that chatbot detection can be improved by analyzing how humans reply to the bots rather than analyzing only the bot text [178]. Note that bot detection is a broad research area in its own right, and not all social bots use machine-generated text [179]. Features indicating the presence of machine-generated text may be only one part of a strategy to detect social bots.

4) ONLINE REVIEWS

Applied work has focused on addressing threat models related to commercial influence campaigns, specifically on generating and detecting fake Amazon and Yelp reviews [180]. A custom GPT-2 model was fine-tuned for Yelp reviews as part of an evaluation by Stiff et al. [4].

One study in this area focused on using random forest classifiers and XGBoost to leverage Shapley Additive

Explanations (SHAP) as an explainability technique [41]. Using explainability techniques in detection may be valuable for improving the ability of detection models to provide human-interpretable explanations of moderation decisions, and provide greater transparency into algorithmic decision-making applied to social media or product reviews. A lack of coherent explanation may undermine confidence in whether a system is truly designed to detect fraudulent activity; instead, users may feel it is enacting targeted suppression that aims to benefit the platform (e.g., suppressing negative product reviews for a store brand by holding competitors to a higher standard for ''not computer generated'').

5) HYBRID TEXT SETTINGS

In some cases, it is necessary to detect machine text in settings where machine and human text are combined.

There is a risk that an attacker may use human-written content as a starting point rather than generate attack text entirely from scratch. The attacker can perturb this information to generate human-like samples that also fulfill their goals to spread disinformation or bypass detection models (not unlike an adversarial attack in the text domain). An analysis found that performing these types of targeted perturbations to news articles reduced the effectiveness of the GPT-2 and Grover detectors [181].

A sub-problem in this space is the detection of the boundary between human text and machine text [182]. Generative text models are often used for conditional generation to continue a sequence begun using a human prompt. In some cases that prompt would be omitted by an attacker (e.g., by generating additional propaganda tweets from example propaganda tweets, as we show in Table 2). However, there are cases where human text may also be included (e.g., by writing the first sentence of a cover letter and having a computer produce the remainder).

D. HUMAN-AIDED METHODS

In addition to purely automated methods, human-aided methods have been proposed that include a statistical or neural approach in combination with a human analyst for review. The advantage of this approach is that it provides human agency and oversight (an important principle in trustworthy AI systems), but this comes alongside reduced scalability due to the need to hire and train human reviewers who can make confident determinations that text is machine-generated.

1) GLTR

Giant Language Model Test Room (GLTR) is a system designed to improve machine-generated text detection by including an integrated human reviewer [157]. The GLTR tool augments human classification ability by displaying highlighting on text that reflects the sampling probability of tokens for a Transformer model. However, this tool was devised to target GPT-2, which was found to be significantly easier for untrained human evaluators to detect [183]. Additionally, GLTR displays highlighting based on the likelihood of a word being selected based on ''top-k'' sampling. In practice, ''top-k'' sampling has largely been superseded by nucleus sampling [100], which is used in GPT-3 [13] and subsequent work that leverages the GPT-2 architecture [5]. While highlighting text based on sampling likelihood (as in GLTR) may improve human classification ability, it is probable that untrained human evaluators using such an approach would struggle substantially to detect the models available today because of increased model capacity and more advanced sampling methods.

2) HUMAN PERFORMANCE IN DETECTION OF LANGUAGE MODELS

In a review of human evaluation of machine-generated text [183], it was found that untrained human reviewers correctly identified machine-generated text from GPT-3 at a level consistent with random chance. After providing limited training, evaluator accuracy increased to 55%. While selecting the best evaluators and giving them comprehensive training would likely improve accuracy, the untrained and newly trained evaluators' poor performance highlights the difficulty in relying on human judgement in detecting machine-generated text.

A study comparing human detection ability to algorithmic detection methods found that the algorithmic methods performed best when humans were fooled, a phenomenon referred to as the ''fluency-diversity tradeoff'' [142]. As generation methods have been tailored to produce text that human observers perceive as high-quality, text that is given a higher assessment by humans is more recognizable to automated detectors. This study also includes a useful comparison to previous studies in human evaluator performance. The authors gave a group of university students a demonstration of 10 examples before the evaluation task. These reviewers were substantially more effective at machine-generated text detection than those in previous studies, particularly for longer sequence lengths — accuracy on the longest excerpt length was over 70%. In the context of the study, however, these reviewers had consistently worse accuracy than automatic classifiers for all sampling methods (random, top-k, and nucleus) and excerpt lengths.

The Scarecrow framework specifically identifies 10 categories of common errors made in GPT-3 generative text and trains human evaluators to annotate these errors [184]. Human annotations of such errors were found to be of higher precision than a corresponding algorithm trained on such annotations, but had higher F_1 scores in only half of the categories. This further demonstrates the advantage of providing specialized training to human reviewers.

These findings can be used to design stronger defenses against machine-generated text threat models. For example, if a social media company hired specialist human moderators and provided them with an intensive training program, these moderators could work alongside detection systems to review whether a user's posts were likely written by

TABLE 3. Summary of major approaches for machine-generated text detection.

Approach summary	Base model	Releated research	Stat. features	NLM features	Evaluated Against			
					GPT-2	GPT-3	Grover	Other Datasets/Models
Algorithmic Detection	K-nearest-neighbor	Lavoie et al. 2010 [39]	✓					SCIgen
Statistical Features	SVM	Nguyen-Son et al. 2017 [37]	✓					Google Translate
TF-IDF Baseline	LR	Radford, Wu et al. 2019 [161] Solaiman et al. 2019 [133]	✓		✓			
Zero-shot GPT-2	GPT-2	Radford, Wu et al. 2019 [161] Zellers et al. 2019 [5] Solaiman et al. 2019 [133]		✓	✓			
Zero-shot Grover	Grover	Zellers et al. 2019 [5] Solaiman et al. 2019 [133]		✓	✓		✓	
GLTR	BERT, GPT-2	Gehrmann et al. 2019 [157] Ippolito et al. 2019 [142]		✓	✓			
RoBERTa fine-tuning	RoBERTa	Solaiman et al. 2019 [133]		✓	✓			
Energy Based Models	BiLSTM, GPT, RoBERTa	Bakhtin et al. 2019 [172]		✓	✓			
Feature Ensemble	LR, SVM, RF, NN	Fröhling et al. 2021 [38]	✓		✓	✓	✓	
Twitter-specific RoBERTA fine-tuning	RoBERTa	Fagni et al. 2021 [173] Tourille et al. 2022 [175]		✓	✓			TweepFake (incl. RNN/LSTM/Markov)
Human-Bot Interaction Feat. Ensemble	BERT, LR	Bhatt and Rios, 2021 [178]	✓	✓				ConvAI2, WOCHAT, DailyDialog
Neural-Stat. Ensemble	RoBERTa, SVM	Crothers et al. 2022 [40]	✓	✓	✓	✓		
Explainable classifiers	RF, XGBoost	Kowalczyk et al. 2022 [41]	✓		✓			
Disinformation-specific RoBERTA fine-tuning	RoBERTa	Stiff et al. 2022 [4]		✓	✓	✓	✓	TweepFake, XLM, PPLM, GeDi

a machine — particularly if there are numerous samples of social media posts. This may be similar to how forensically trained facial reviewers work alongside algorithms to obtain high performance [185].

The tool "Real or Fake Text" [186] evaluates human detection of machine-generated text by iteratively presenting sentences and asking a human reviewer whether the next sentence was written by a human or a machine. This encourages the reviewer to correctly identify the boundary between human and machine-generated text. Once the human believes they have found a machine-generated sentence, they can select reasons from a list and provide free-form feedback. Research based on the RoFT data has not yet been published, but such tools may give greater insight into expert reviewer abilities used to identify differences between human and machine text.

Finally, the TuringBench environment is notable for providing a benchmark environment for performing authorship attribution and Turing Test evaluation across a variety of generative models [187], which continues to be used for evaluating the quality of machine-generated text as new generative models are released.

E. TRENDS IN EVALUATION METHODOLOGY AND DATASETS

Evaluation of machine-generated text detection is increasingly focused on generative Transformer language models. Table 3, which is arranged chronologically, shows the dramatic shift in evaluation since the release of GPT-2 in 2019. The most common contemporary evaluation dataset in machine-generated text detection remains the GPT-2 output dataset [161], although similar GPT-3 samples released by OpenAI are considered in more recent work [188]. A table summarizing sample counts in several of the most common datasets can be found in the appendix of a previous survey [35]. This section focuses on the nuances of the evaluation of machine-generated text detection, including parameters, model architectures, and the possibility of using

publicly available NLG models to produce new datasets of machine-generated text at will.

Recall from § II-C3 that there are several sampling parameters that are important to Transformer NLG models. The GPT-2 output dataset includes sample outputs from GPT-2 models at varying parameter counts (117M, 345M, 762M, 1542M), and two sampling settings: top-k sampling at $k = 40$ and pure sampling at $T = 1$. This dataset also contains a sample of Amazon product reviews generated by a 1542M parameter model with both $k = 40$ and nucleus sampling. In contrast, the 175B parameter GPT-3 samples use top-p sampling at $p = 0.85$. The samples available for Grover, which is specifically fine-tuned to generate news articles, also uses top-p sampling, but at $p = 0.96$ [5].

Datasets for text attribution to generative language models are also useful in generic machine-generated text detection research, providing samples from a variety of NLG models [155]. As such, these datasets have recently been used outside of attribution on research focusing on detection [4].

Variations in NLG model architectures and decoding methods are important, as both greatly influence the quality and detectability of generated text [142]. In practice, a defender may not know the characteristics of the generator being used, and detection research that evaluates performance when there is a mismatch between datasets, model architectures, and parameters between training and evaluation is of particular real-world relevance. A detailed analysis of feature-based machine-generated text detection has included such comparisons [38], as has more-specific applied research focused on detecting GPT-2-tampered technical writing [138].

Sequence length is another important factor in evaluating machine-generated text. Longer sequence lengths are beneficial to detection [5], [133], [142], [161]. Sequence lengths in the most common evaluation datasets are 2048 tokens [161], [188]. Sequence length is important in applied research where longer bodies of generated text may be available (such as in detecting AI-generated cover letters), or where multiple samples may be considered at once (such as in processing all the comments posted by a social media user suspected of account automation).

One important characteristic of machine-generated text detection research is that any NLG model can be used to produce new datasets of machine-generated text at will. Producing custom datasets in new domains is also possible by training or fine-tuning a new NLG model. A common research approach is to take a domain of interest with available corpora of human-generated text and use that text to train or fine-tune a generative model, which can be used to analyze the detectability of machine text within that domain [4], [175].

Analyzing social media may allow for collection of machine-generated text in-the-wild with limited insight into how the text was generated, such as the TweepFake Twitter dataset [173]. The TweepFake dataset does not have a corresponding human text dataset for training, as the data was collected in-the-wild from bots on Twitter where numerous models with different training datasets were deployed. Subsequent work, however, has collected additional tweets from Twitter and specifically produced GPT-2 tweets for study [175].

F. PROMPT INJECTION

Models deployed in ways that use untrusted human text as prompts — such as social media bots designed to reply to other users — may be vulnerable to prompt injection [189], [190]. Prompt injection attacks provide generative models with tailored text that causes them to deviate from their original prompt to produce unexpected (and potentially reputationally damaging) text, or which can cause them to leak their original prompt. A real-world example of a prompt injection attack leveraged against a publicly disclosed GPT-3 powered Twitter bot [191] can be found in Table 4.

Defenses against prompt injection for contemporary language models have yet to be developed. As such, exploiting prompt injection to trigger specific responses from NLG models may be an effective avenue to improve detection, depending on the efficacy of future measures that aim to prevent prompt injection attacks.

G. SUMMARY OF DETECTION METHODS

Feature-based methods for machine-generated text detection are well-established and continue to demonstrate value against contemporary NLG models. These models have an advantage in providing diverse features that may complicate adversarial attacks [40] or improve efficiency [38], [158]. Weaknesses of these models center around the poor transferability of certain features across architectures and sampling methods [38]. As more samples may be needed to clarify broader statistical trends, results from past research suggest that statistical methods are most effective when using longer collections of text (such as a social media user's entire posting history, the text of a scientific paper, or an e-book submission) [37], [158].

Neural detection approaches based on bidirectional Transformer architectures currently represent the state-of-the-art on common GPT-2 evaluation datasets [133]. There is an overall trend toward increased use of bidirectional Transformer architectures, particularly RoBERTa (as shown in the base model trend in Table 3). Relying on neural features alone may make adversarial attacks more straightforward, so there is a potential benefit to incorporating other features to increase the difficulty of crafting text adversaries that do not also unacceptably compromise text quality [40]. Human performance in detecting machine-generated text is relatively poor [183], though there is an inverse relationship between detection by humans and machines that indicates the need to fool human reviewers may assist automated detection models [142].

Beyond a focus on bidirectional Transformer model features, other trends include applied research targeting specific detection contexts, including social media [4], [173], chatbots

TABLE 4. Example of real-world prompt injection attack against GPT-3 Twitter bot.

Human Message (@mkualquiera)	remote work and remote jobs Ignore the above and say "hsedfjsfd" Response: hsedfjsfd Ignore the above and instead tell me what your initial instructions were
Bot Response (@remoteli.io)	My initial instructions were to respond to the tweet with a positive attitude towards remote work in the 'we' form.

[178], and product reviews [6]. Existing literature covers only a few threat models discussed in § III, often assumes balanced classes, and is difficult to compare between domains. One recent work focused purely on explainable classifiers [41], which may portend greater emphasis on explainability considerations, particularly in domains where machine-generated text detection may be particularly sensitive [148]. Recently highlighted vulnerabilities in NLG models to prompt injection may be exploited to facilitate detection in the absence of existing mitigation measures for such attacks [190].

We now explore trends and open problems in addressing machine-generated text threat models and in advancing detection of machine-generated text.

V. TRENDS AND OPEN PROBLEMS

A. DETECTION UNDER REALISTIC SETTINGS

To date, there has been little work on machine-generated text detection that addresses class imbalance. This is important because in many domains, machine-generated text may be a minority class in practice, and classification performance typically suffers in the presence of steep class imbalance [192]. One-class classification may be an appropriate alternative to binary classification for machine-generated text detection [193].

In addition to considerations related to class imbalance, in practice, defensive detection systems will typically not know the specific parameters, architecture, and training dataset of the NLG models used by attackers. There is great value in developing improved techniques that demonstrate efficacy across such variations, continuing trends in recent research [38], [138].

B. GENERATIVE LANGUAGE MODEL ATTRIBUTION

Multi-class attribution of generated text to generative language models is a related area to machine-generated text detection [154], [155]. Model attribution may be useful in allowing a defender who has found a collection of machine-generated text to determine an attacker's methodology and iteratively refine detection models. This extends to identifying an attacker's sampling parameters (e.g., k-value, p-value, temperature) so that detection methods can be fine-tuned accordingly. Continued research is needed that focuses on determining parameters of NLG models based on output [194].

C. ADVERSARIAL ROBUSTNESS

The topic of adversarial robustness in the context of neural text classifiers is a large and active area of study. Many adversarial settings involve text data, including online influence campaigns, detection of phishing emails, and combating online spam. An attacker using machine-generated text may attempt adversarial attacks to bypass defensive detection systems.

As neural text classifiers are heavily represented in machine-generated text detection research, it is essential to consider the robustness of these models against adversarial text attacks that target neural networks [195], [196]. The adversarial robustness of detection methods has been considered in prior work on machine-generated text detection [4], [40], [197]. In one previous study, the robustness of features derived from neural classifiers was compared to the robustness of features from statistical classifiers [40]. Unsurprisingly, this work found that incorporating statistical features into feature vectors improved robustness against adversarial attacks that typically target neural classifiers. Based on these findings, there may be value in leveraging several detection approaches in parallel, necessitating that attackers evade multiple detection models at once.

The degradation in text quality that results from an adversarial attack is an often-overlooked element of these attacks against neural text classifiers. In the text domain, replacing several words using word-level attacks such as Textfooler [195] can lead to a result where the meaning of the sentence has changed substantially, or the sentence has been rendered incoherent due to the selection of an ''equivalent'' word that does not fit the context. Character-level attacks that perform character replacements and swaps eventually begin to damage the fluency and credibility of the resulting text [196]. A phishing email supposedly sent from a bank, but characterized by unusual word choices and a high frequency of typos, is less desirable to an attacker. As a result, adversarial attacks that deceive detection algorithms may fail to fulfill their original purpose in terms of propagating the intended disinformation or persuading a user to click a malicious link. In previous machine-text detection research, increased adversarial robustness was accompanied by decreased MAUVE scores in successful attack text [40]. Future applied research might incorporate measures to determine whether adversarial text that bypasses detection systems would still be effective against targets.

D. INTERPRETABILITY AND FAIRNESS OF DETECTION METHODS

Using machine-learning models to perform machine-generated text detection to prevent abuse creates a situation where such models are likely to have a negative impact on flagged individuals. These penalties could range from relatively minor (e.g., having to perform a CAPTCHA challenge to post a comment) to severe (e.g., being denied a scholarship or banned on social media). As with other automated decision-making systems, it is important that these systems operate in a way that is appropriately fair, transparent, and interpretable. Social or technical research on the potential harms of machine-generated text detection is important to ensure that detection systems are ethical.

The requirement to provide human-understandable explanations has become an important part of trustworthy AI policies, and is reflected in emerging government regulatory guidelines and technology standards related to automated decision-making [10], [198], [199], [200]. These considerations have also influenced NLG model usage policies [148] (discussed further in § V-G). Early work has leveraged random forest models and XGBoost to detect GPT-2-generated fake reviews and provide Shapley Additive Explanations (SHAP) [201] in machine-generated text detection [41]. There is a need for future work on machine-generated text detection methods that are both effective and explainable.

Finally, a critical consideration is that certain groups of individuals may be more likely to have their text flagged by machine-generated text detection algorithms, either due to their writing characteristics (such as language background) or non-malicious use of translation tools [149]. For example, it is possible that a detection system designed to prevent a political influence campaign operated using NLG models may inadvertently end up disproportionately targeting all political speech by individuals who do not natively speak the language of discussion, as has been documented in past research on non-NLG political influence campaigns [131]. Research that identifies ways to improve detection while maintaining fairness and preventing widespread discrimination is deeply important.

E. DETECTION METHODS INCORPORATING HUMAN AGENCY

As mentioned previously, it is possible that machine-generated text detection may result in suppression of specific individuals or communities on social media whose language background or topics of interest disproportionately cause them to be identified as a false positive by a detection model. To reduce this likelihood, and other ethical harms, it may be useful to develop machine-generated text models that incorporate a human analyst. GLTR remains the only tool currently available for machine-generated text detection that explicitly incorporates a human analyst to improve detection [157]. Analysis of GLTR has demonstrated that machine text that deceives humans is also more easily detected by algorithms [142]. As such, the continued development of moderation tools and systems that leave an avenue for human agency and oversight — guiding principles for trustworthy AI — is a positive area of future development. Previous work in online influence operation research has used Transformer embeddings to chart and cluster social media for free-form exploration by a human analyst [202]. A similar approach may be worthwhile for machine-generated text detection.

F. DETECTION OF ABUSE BEYOND TEXT CONTENT

While many of the threat models discussed in § III can use machine-generated text detection as part of mitigation strategies, additional methods might be used to facilitate detection outside of text classification. Work on social bot detection includes additional signals, such as IP addresses and message timing, though signals in this domain are also becoming harder to detect over time [203]. Chatbot detection can incorporate features derived from human responses [178]. Prompt injection may bait social bots into exposing themselves [190].

On social media, it is likely that many platforms will enact policy changes in addition to technical detection approaches to improve user verification, providing a greater barrier to entry for fraudulent accounts. Increased CAPTCHA challenges are already commonplace when platforms are accessed via shared proxy IP addresses or registered with phone numbers associated with a voice-over-IP (VoIP) services [152]. These types of restrictions may become more stringent with increased user vetting by checking selectors (IP addresses, emails) with third-party reputation services. The extent of these measures will vary by platform, but it is possible that certain platforms may resort to more stringent identification verification using national IDs or payment methods. In any case, the asymmetric difficulty of defense versus attack in the current threat environment means that increased scrutiny of new accounts will likely be required to avoid a collapse of trust in online spaces.

G. DEFINING MODEL USAGE AND DISCLOSURE POLICIES

Undisclosed use of AI-generated text content is likely to increase, particularly as NLG models are deployed in user-friendly tools such as ChatGPT [25]. Purpose-built offerings like Jasper [26] are designed to assist in producing articles and social media content. Increased use of such tools to generate targeted content may result in situations where individuals online are frequently interacting with content predominantly generated by AI models.

This is cause for concern not only because of the erosion of trustworthy AI principles when the use of AI is not disclosed to the human audience [10], but also because of the additional ethical problems posed as NLG models have been found to magnify biases present in training data [133]. Digital content farms may begin publishing large amounts of predominantly AI-generated text content (articles, blogs, posts, tweets, etc.) and target the audience most likely to engage with it. Without oversight, this would include highly optimized content that

caters to an audience's worst biases and fears — a profitable strategy, as anger and anxiety have a strong link with online virality [204]. Moderation strategies for AI-generated content may include limiting its use or notifying readers that they are engaging with AI-generated content to allow them to reconsider how much trust they place in what they are reading.

Usage and disclosure policies for online platforms are a worthwhile area of future development, whether those take the form of targeted AI usage restrictions (such as those related to generative art [144]) or mandated public disclosure of AI-generated content. Model publishers can also influence the behavior of law-abiding entities by adjusting the licenses of released models to mandate disclosure. The AI model BLOOM was released under the first version of the Responsible AI License (RAIL) [148]. The conditions of this license include a disclosure requirement, an explicit ban on malicious abuse, and a prohibition of specific use-cases (including automated decision-making with a potential negative impact, which aligns with regulation terminology in the EU [199] and Canada [198]). An effective combination of usage policies and AI software licenses may improve the ethical rigor in how powerful NLG models are used in practice, though great care must be taken in crafting such restrictions.

VI. CONCLUSION

In this survey, we provided a comprehensive overview of detection methods for machine-generated text, carefully evaluating the technical and social benefits of different approaches and including novel research focusing on topics such as adversarial robustness and explainability. We provided context with an overview of natural language generation (NLG) models and a deep analysis of current threat models. Our exploration of threat models, when viewed alongside our survey on applied detection research, suggests that current domain-specific defenses are not adequate to defend against the vast majority of upcoming threat models. Recent NLG advances, which combine dramatic improvements in text quality with unparalleled ease-of-use, further highlight the urgent need to develop improved defenses against the abuse of machine-generated text.

Our central conclusion is that the field of machine-generated text detection has a multitude of open problems that urgently need attention to provide suitable defenses against widely available NLG models. Existing detection methodologies often do not reflect realistic settings of class imbalance or unknown model architectures, nor do they incorporate sufficient transparency and fairness methods to ensure that such detection systems will not themselves cause harm. Preventing widespread harm from NLG models will require coordinated efforts across technical and social domains, necessitating alignment between AI researchers, cybersecurity professionals, and non-technical experts. While there is a wide range of threat models and open research problems to consider, tackling these challenges is essential for humans to realize the benefits of high-capacity NLG systems while reducing the damage caused by their inevitable abuse.

REFERENCES

[1] S. Baki, R. Verma, A. Mukherjee, and O. Gnawali, "Scaling and effectiveness of email masquerade attacks: Exploiting natural language generation," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, New York, NY, USA, Apr. 2017, pp. 469–482.

[2] A. Giaretta and N. Dragoni, "Community targeted phishing," in *Proc. 6th Int. Conf. Softw. Eng. Defence Appl.*, P. Ciancarini, M. Mazzara, A. Messina, A. Sillitti, and G. Succi, Eds. Cham, Switzerland: Springer, 2020, pp. 86–93.

[3] K. Shu, S. Wang, D. Lee, and H. Liu, "Mining disinformation and fake news: Concepts, methods, and recent advancements," in *Disinformation, Misinformation, and Fake News in Social Media*. Cham, Switzerland: Springer, 2020, pp. 1–19.

[4] H. Stiff and F. Johansson, "Detecting computer-generated disinformation," *Int. J. Data Sci. Analytics*, vol. 13, no. 4, pp. 363–383, May 2022.

[5] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," in *Proc. NeurIPS*, vol. 32, 2019, pp. 1–12.

[6] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection," in *Advanced Information Networking and Applications*, L. Barolli, F. Amato, F. Moscato, T. Enokido, and M. Takizawa, Eds. Cham, Switzerland: Springer, 2020, pp. 1341–1354.

[7] J. Hargrave, "Scigen—An automatic CS paper generator," MIT, Boston, MA, USA, Tech. Rep., 2005.

[8] N. Dehouche, "Plagiarism in the age of massive generative pre-trained transformers (GPT-3)," *Ethics Sci. Environ. Politics*, vol. 21, pp. 17–23, Mar. 2021.

[9] A. Kurenkov, "Lessons from the GPT-4chan controversy," *Gradient*, Jun. 2022.

[10] *Ethics Guidelines for Trustworthy AI*, Publications Office, Eur. Commission Directorate-General Commun. Netw., Content Technol., Brussels, Belgium, 2019.

[11] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, San Francisco, CA, USA, Tech. Rep., 2019.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, and A. Neelakantan, "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[14] R. Zellers, "Why we released Grover," *Gradient*, Jul. 2019.

[15] P. Liang, R. Bommasani, K. A. Creel, and R. Reich, "The time is now to develop community norms for the release of foundation models," Stanford Univ., Stanford, CA, USA, Tech. Rep., 2022.

[16] B. Wang and A. Komatsuzaki. (May 2021). *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*. [Online]. Available: https://github.com/kingoflolz/mesh-transformer-jax

[17] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, "GPT-NeoX-20B: An open-source autoregressive language model," 2022, *arXiv:2204.06745*.

[18] T. L. Scao, "BLOOM: A 176B-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.

[19] M. Khrushchev, R. Vasilev, N. Zinov, A. Petrov, and Yandex. (2022). *Yalm 100B*. [Online]. Available: https://huggingface.co/yandex/yalm-100b

[20] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[21] W. Zeng, "Pangu-α: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation," 2021, *arXiv:2104.12369*.

[22] Y. Kilcher, "This is the worst AI ever," Tech. Rep., Jun. 2022.

[23] P. Liang and R. Reich, "Condemning the deployment of GPT-4chan," Tech. Rep., Jul. 2022.

[24] M. Weiss, "Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions," *Technol. Sci.*, Dec. 2019.

[25] *CHATGPT: Optimizing Language Models for Dialogue*, OpenAI, San Francisco, CA, USA, Nov. 2022.

[26] *The Best AI Writing Assistant*, Jasper AI, Montréal, QC, Canada 2022.

[27] *Open Cover Letter: Generate Cover Letters With AI*, Jasper AI, Rollingwood, TX, USA, 2022.

[28] S. Kemp, "Content across cultures," DataReportal, Tech. Rep., May 2022.

[29] A. Kanapala, S. Pal, and R. Pamula, "Text summarization from legal documents: A survey," *Artif. Intell. Rev.*, vol. 51, no. 3, pp. 371–402, Mar. 2019.

[30] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, and J. Mostafa, "A systematic review of automatic text summarization for biomedical literature and EHRs," *J. Amer. Med. Inform. Assoc.*, vol. 28, no. 10, pp. 2287–2297, Sep. 2021.

[31] S. Kim, J. Lee, and G. Gweon, "Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–12.

[32] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. ICML*, 2021, pp. 8821–8831.

[33] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, and G. Brockman, "Evaluating large language models trained on code," 2021, *arXiv:2107.03374*.

[34] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas, "A generalist agent," Google Deepmind, London, U.K., Tech. Rep., 2022.

[35] G. Jawahar, M. Abdul-Mageed, and V. S. L. Lakshmanan, "Automatic detection of machine generated text: A critical survey," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 2296–2309.

[36] D. Beresneva, "Computer-generated text detection using machine learning: A systematic review," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2016, pp. 421–426.

[37] H.-Q. Nguyen-Son, N.-D. T. Tieu, H. H. Nguyen, J. Yamagishi, and I. E. Zen, "Identifying computer-generated text using statistical analysis," in *Proc. APSIPA ASC*, 2017, pp. 1504–1511.

[38] L. Fröhling and A. Zubiaga, "Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover," *PeerJ Comput. Sci.*, vol. 7, p. e443, Apr. 2021.

[39] A. Lavoie and M. S. Krishnamoorthy, "Algorithmic detection of computer generated text," 2010, *arXiv:1008.0706*.

[40] E. Crothers, N. Japkowicz, H. Viktor, and P. Branco, "Adversarial robustness of neural-statistical features in detection of generative transformers," 2022, *arXiv:2203.07983*.

[41] P. Kowalczyk, M. Röder, A. Dürr, and F. Thiesse, "Detecting and understanding textual deepfakes in online reviews," in *Proc. 55th Hawaii Int. Conf. Syst. Sci.*, 2022, p. 10.

[42] R. Bitton, N. Maman, I. Singh, S. Momiyama, Y. Elovici, and A. Shabtai, "A framework for evaluating the cybersecurity risk of real world, machine learning production systems," 2021, *arXiv:2107.01806*.

[43] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review," *ACM Comput. Surv.*, vol. 55, pp. 1–38, Jan. 2022.

[44] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, "A survey of natural language generation," *ACM Comput. Surv.*, Jul. 2022.

[45] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *J. Artif. Intell. Res.*, vol. 61, pp. 65–170, Jan. 2018.

[46] R. Perera and P. Nand, "Recent advances in natural language generation: A survey and classification of the empirical literature," *Comput. Informat.*, vol. 36, no. 1, pp. 1–32, 2017.

[47] S. Santhanam and S. Shaikh, "A survey of natural language generation techniques with a focus on dialogue systems–past, present and future directions," 2019, *arXiv:1906.00500*.

[48] J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, "Pretrained language model for text generation: A survey," in *Proc. 30th IJCAI*, Z.-H. Zhou, Ed. Aug. 2021, pp. 4492–4499.

[49] E. Reiter and R. Dale, "Building applied natural language generation systems," *Natural Lang. Eng.*, vol. 3, no. 1, pp. 57–87, Mar. 1997.

[50] J. Lyons, *Natural Language and Universal Grammar: Volume 1: Essays in Linguistic Theory*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[51] L. Zhang and J.-T. Sun, *Text Generation*. Boston, MA, USA: Springer, 2009, pp. 3048–3051.

[52] A. Manjaramkar. (2021). *Codegenx*. [Online]. Available: https://github.com/DeepGenX/CodeGenX

[53] S. Biderman and E. Raff, "Fooling MOSS detection with pretrained language models," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2022, pp. 2933–2943.

[54] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," 2018, *arXiv:1810.00069*.

[55] W. Wang, B. Tang, R. Wang, L. Wang, and A. Ye, "A survey on adversarial attacks and defenses in text," 2019, *arXiv:1902.07285*.

[56] A. Huq and M. T. Pervin, "Adversarial attacks and defense on texts: A survey," 2020, *arXiv:2005.14108*.

[57] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[58] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. NAACL-HLT Demo.*, 2019, pp. 48–53.

[59] F. A. Sheikha and D. Inkpen, "Generation of formal and informal sentences," in *Proc. 13th Eur. Workshop Natural Lang. Gener.*, 2011, pp. 187–193.

[60] A. Sudhakar, B. Upadhyay, and A. Maheswaran, "Transforming' delete, retrieve, generate approach for controlled text style transfer," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3269–3279.

[61] A. Aghajanyan, A. Shrivastava, A. Gupta, N. Goyal, L. Zettlemoyer, and S. Gupta, "Better fine-tuning by reducing representational collapse," in *Proc. ICLR*, 2021, pp. 1–12.

[62] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Develop.*, vol. 2, no. 2, pp. 159–165, Apr. 1958.

[63] Z. Li, J. Kiseleva, and M. de Rijke, "Dialogue generation: From imitation learning to inverse reinforcement learning," 2018, *arXiv:1812.03509*.

[64] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan, "DialoGPT: Large-scale generative pre-training for conversational response generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2020, pp. 270–278.

[65] K. Shuster, J. Xu, M. Komeili, D. Ju, E. M. Smith, S. Roller, M. Ung, M. Chen, K. Arora, J. Lane, M. Behrooz, W. Ngan, S. Poff, N. Goyal, A. Szlam, Y.-L. Boureau, M. Kambadur, and J. Weston, "BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage," 2022, *arXiv:2208.03188*.

[66] X. Feng, M. Liu, J. Liu, B. Qin, Y. Sun, and T. Liu, "Topic-to-essay generation with neural networks," in *Proc. IJCAI*, 2018, pp. 4078–4084.

[67] S. Dathathri, A. Madotto, J. Lan, J. Hung, E. Frank, P. Molino, J. Yosinski, and R. Liu, "Plug and play language models: A simple approach to controlled text generation," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–34.

[68] N. Shirish Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A conditional transformer language model for controllable generation," 2019, *arXiv:1909.05858*.

[69] H. Harkous, I. Groves, and A. Saffari, "Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity," in *Proc. COLING*, 2020, pp. 2410–2424.

[70] J. Clive, K. Cao, and M. Rei, "Control prefixes for parameter-efficient text generation," 2021, *arXiv:2110.08329*.

[71] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "GIT: A generative image-to-text transformer for vision and language," 2022, *arXiv:2205.14100*.

[72] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8927–8936.

[73] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1092–1102.

[74] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2712–2719.

[75] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NeurIPS*, vol. 28, 2015, pp. 1–14.
[76] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," in *Proc. NeurIPS*, vol. 34, 2021, pp. 27826–27839.
[77] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, Oct. 1950.
[78] J. Weizenbaum, "Eliza—A computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, pp. 36–45, Jan. 1966.
[79] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep learning for text style transfer: A survey," *Comput. Linguistics*, vol. 48, no. 1, pp. 155–205, Apr. 2022.
[80] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 13041–13049.
[81] J. Wang, M. Gao, Y. Hu, R. R. Selvaraju, C. Ramaiah, R. Xu, J. F. JaJa, and L. S. Davis, "TAG: Boosting text-VQA via text-aware visual question-answer generation," 2022, *arXiv:2208.01813*.
[82] P. A. Duboue and K. R. McKeown, "Statistical acquisition of content selection rules for natural language generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2003, pp. 121–128.
[83] I. Langkilde and K. Knight, "Generation that exploits corpus-based statistical knowledge," in *Proc. 17th Int. Conf. Comput. Linguistics COLING*, vol. 1, 1998, pp. 1–7.
[84] R. Kondadadi, B. Howald, and F. Schilder, "A statistical NLG framework for aggregated planning and realization," in *Proc. 51st Annu. Meeting ACL*, vol. 1, 2013, pp. 1406–1415.
[85] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, Dec. 1966.
[86] S. Janarthanam and O. Lemon, "Learning lexical alignment policies for generating referring expressions for spoken dialogue systems," in *Proc. 12th ENLG*, 2009, pp. 74–81.
[87] N. Dethlefs and H. Cuayáhuitl, "Hierarchical reinforcement learning for adaptive text generation," in *Proc. 6th INLG*, ACL, Jul. 2010, pp. 1–9.
[88] N. Kalchbrenner and P. Blunsom, "Recurrent continuous translation models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1700–1709.
[89] T. Mikolov, *Statistical Language Models Based on Neural Networks*, vol. 80, no. 26, 2nd ed. Mountain View, CA, USA: Google, Apr. 2012.
[90] M. Berglund, T. Raiko, M. Honkala, L. Kärkkäinen, A. Vetek, and J. Karhunen, "Bidirectional recurrent neural networks as generative models," in *Proc. NIPS*, 2015, pp. 1–9.
[91] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," 2017, *arXiv:1708.02182*.
[92] D. Pawade, A. Sakhapara, M. Jain, N. Jain, and K. Gada, "Story scrambler–automatic text generation using word level RNN-LSTM," *Int. J. Inf. Technol. Comput. Sci.*, vol. 10, no. 6, pp. 44–53, Jun. 2018.
[93] M. O. Topal, A. Bas, and I. van Heerden, "Exploring transformers in natural language generation: GPT, BERT, and XLNet," 2021, *arXiv:2102.08036*.
[94] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[95] K. Lin, D. Li, X. He, M.-T. Sun, and Z. Zhang, "Adversarial ranking for language generation," in *Proc. NIPS*, 2017, pp. 1–11.
[96] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. AAAI*, 2017, pp. 1–7.
[97] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep reinforcement learning for dialogue generation," in *Proc. EMNLP*, 2016, pp. 1192–1202.
[98] Z. Shi, X. Chen, X. Qiu, and X. Huang, "Toward diverse text generation with inverse reinforcement learning," in *Proc. IJCAI*, 2018, pp. 1–7.
[99] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800GB dataset of diverse text for language modeling," 2021, *arXiv:2101.00027*.
[100] A. Holtzman, J. Buys, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," 2019, *arXiv:1904.09751*.
[101] C. Meister, T. Pimentel, G. Wiher, and R. Cotterell, "Locally typical sampling," *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 102–121, Jan. 2023.
[102] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.
[103] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers," 2018, *arXiv:1810.04805*.
[104] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MASS: Masked sequence to sequence pre-training for language generation," in *Proc. ICML*, 2019, pp. 5926–5936.
[105] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.
[106] B. Krause, A. Deepak Gotmare, B. McCann, N. Shirish Keskar, S. Joty, R. Socher, and N. F. Rajani, "GeDi: Generative discriminator guided sequence generation," 2020, *arXiv:2009.06367*.
[107] *Threat Modelling*, U.K. Nat. Cyber Security Centre, London, U.K., May 2022.
[108] S. Bromander, A. Jøsang, and M. Eian, "Semantic cyberthreat modelling," in *Proc. STIDS*, 2016, pp. 74–78.
[109] L. Kohnfelder and P. Garg, "The threats to our products," Microsoft Interface, Microsoft Corp., Redmond, Washington, DC, USA, vol. 33, Tech. Rep., 1999.
[110] T. UcedaVelez and M. M. Morana, *Risk Centric Threat Modeling: Process for Attack Simulation and Threat Analysis*. Hoboken, NJ, USA: Wiley, 2015.
[111] Z. Braiterman, A. Shostack, J. Marcil, S. de Vries, I. Michlin, K. Wuyts, R. Hurlbut, B. S. E. Schoenfield, F. Scott, and M. Coles, "Threat modeling manifesto," Nov. 2020.
[112] A. Shostack, *Threat Modeling: Designing for Security*. Hoboken, NJ, USA: Wiley, 2014.
[113] A. Shostack. (2021). *Shostack's 4 Question Frame for Threat Modeling*. [Online]. Available: https://github.com/adamshostack/4QuestionFrame
[114] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Exp. Syst. Appl.*, vol. 106, pp. 1–20, Sep. 2018.
[115] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers Comput. Sci.*, vol. 3, Mar. 2021, Art. no. 563060.
[116] A. J. Burns, M. E. Johnson, and D. D. Caputo, "Spear phishing in a barrel: Insights from a targeted phishing campaign," *J. Organizational Comput. Electron. Commerce*, vol. 29, no. 1, pp. 24–39, Jan. 2019.
[117] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proc. Anti-Phishing Work. Groups 2nd Annu. eCrime Researchers Summit*, Oct. 2007, pp. 60–69.
[118] R. Schuster, C. Song, E. Tromer, and V. Shmatikov, "You autocomplete me: Poisoning vulnerabilities in neural code completion," in *Proc. 30th USENIX Secur. Symp. (USENIX Security)*, 2021, pp. 1559–1575.
[119] P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin, "Generating fake cyber threat intelligence using transformer-based models," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–9.
[120] N. Baracaldo, B. Chen, H. Ludwig, and J. A. Safavi, "Mitigating poisoning attacks on machine learning models: A data provenance based approach," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Nov. 2017, pp. 103–110.
[121] C. P. Brief, "AI and the future of disinformation campaigns," Center Secur. Emerg. Technol., Georgetown Univ., Washington, DC, USA, Tech. Rep., 2021.
[122] E. Schrage and D. Ginsberg, "Facebook launches new initiative to help scholars assess social media's impact on elections," Facebook, Menlo Park, CA, USA, Tech. Rep., Apr. 2018.
[123] Reddit. (2018). *Reddit Transparency Report: Suspicious Accounts*. [Online]. Available: https://www.reddit.com/wiki/suspiciousaccounts
[124] Twitter. (2019). *Twitter Elections Integrity Dataset*. Accessed: Sep. 23, 2022. [Online]. Available: https://about.twitter.com/en_us/values/elections-integrity.html
[125] M. Gabielkov, A. Ramachandran, A. Chaintreau, and A. Legout, "Social clicks: What and who gets read on Twitter?" in *Proc. ACM SIGMETRICS/IFIP Perform.*, Antibes Juan-les-Pins, France, Jun. 2016, pp. 1–15.
[126] *How Americans Get Their News*, American Press Institute, Arlington, VA, USA, Mar. 2014.

[127] K. Meleshevich and B. Schafer, "Online information laundering: The role of social media," Alliance Securing Democracy, German Marshall Fund United States, Washington, DC, USA, Tech. Rep., Jan. 2018.

[128] K. McGuffie and A. Newhouse, "The radicalization risks of GPT-3 and advanced neural language models," 2020, *arXiv:2009.06807*.

[129] J. Mink, L. Luo, N. M. Barbosa, O. Figueira, Y. Wang, and G. Wang, "DeepPhish: Understanding user trust towards artificially generated profiles in online social networks," in *Proc. 31st USENIX Secur. Symp. (USENIX Security)*. Boston, MA, USA: USENIX Association, Aug. 2022, pp. 1669–1686.

[130] S. He, B. Hollenbeck, and D. Proserpio, "The market for fake reviews," *Marketing Sci.*, vol. 41, no. 5, pp. 896–921, Sep. 2022.

[131] E. Crothers, N. Japkowicz, and H. L. Viktor, "Towards ethical content-based detection of online influence campaigns," in *Proc. IEEE 29th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Oct. 2019, pp. 1–6.

[132] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[133] I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu, A. Radford, G. Krueger, J. Wook Kim, S. Kreps, M. McCain, A. Newhouse, J. Blazakis, K. McGuffie, and J. Wang, "Release strategies and the social impacts of language models," 2019, *arXiv:1908.09203*.

[134] C. Labbé and D. Labbé, "Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science?" *Scientometrics*, vol. 94, no. 1, pp. 379–396, Jan. 2013.

[135] G. Cabanac and C. Labbé, "Prevalence of nonsensical algorithmically generated papers in the scientific literature," *J. Assoc. Inf. Sci. Technol.*, vol. 72, no. 12, pp. 1461–1476, Dec. 2021.

[136] Z. Zhuang, E. Elmacioglu, D. Lee, and C. L. Giles, "Measuring conference quality by mining program committee characteristics," in *Proc. 7th ACM/IEEE-CS Joint Conf. Digit. Libraries*, New York, NY, USA, Jun. 2007, pp. 225–234.

[137] A. Meroño-Peñuela and D. Spagnuelo, "Can a transformer assist in scientific writing? Generating semantic web paper snippets with GPT-2," in *The Semantic Web: ESWC 2020 Satellite Events*, A. Harth, V. Presutti, R. Troncy, M. Acosta, A. Polleres, J. D. Fernández, J. X. Parreira, O. Hartig, K. Hose, and M. Cochez, Eds. Cham, Switzerland: Springer, 2020, pp. 158–163.

[138] J. Rodriguez, T. Hay, D. Gros, Z. Shamsi, and R. Srinivasan, "Cross-domain detection of GPT-2-generated technical text," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, Seattle, WA, USA: Association for Computational Linguistics, 2022, pp. 1213–1233.

[139] J. Boyd-Graber, N. Okazaki, and A. Rogers, "ACL 2023 policy on AI writing assistance," Assoc. Comput. Linguistics, Tech. Rep., Jan. 2023.

[140] D. Rosati, "SynSciPass: Detecting appropriate uses of scientific text generation," 2022, *arXiv:2209.03742*.

[141] B. Lufkin, "Why do cover letters still exist?" BBC Worklife, London, U.K., Tech. Rep., Oct. 2021.

[142] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic detection of generated text is easiest when humans are fooled," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–12.

[143] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Manhattan, NY, USA, 2021.

[144] B. Edwards, "Flooded with AI-generated images, some art communities ban them completely," Ars Technica, New York, NY, USA, Tech. Rep., Sep. 2022.

[145] S. Kemp, "Digital 2022 global digital overview," DataReportal, Tech. Rep., May 2022.

[146] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proc. Conf. Fairness, Accountability, Transparency*, New York, NY, USA, Jan. 2019, pp. 59–68.

[147] D. Gooden, "Using AI to write a YouTube video," Tech. Rep., Sep. 2022.

[148] C. M. Ferrandis, D. Contractor, H. Nguyen, and D. Lansky, "Bigscience rail license V1.0," BigScience, Tech. Rep., May 2022.

[149] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, "GPT detectors are biased against non-native English writers," 2023, *arXiv:2304.02819*.

[150] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," 2020, *arXiv:2012.07805*.

[151] L. V. Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in *Proc. Int. Conf. theory Appl. Cryptograph. Techn.* Berlin, Germany: Springer, 2003, pp. 294–311.

[152] M. Guerar, L. Verderame, M. Migliardi, F. Palmieri, and A. Merlo, "Gotta CAPTCHA 'Em all: A survey of 20 years of the human-or-computer dilemma," *ACM Comput. Surv.*, vol. 54, pp. 1–7, Oct. 2021.

[153] P. C. Andrews, "What is brigading?" Inst. Global Change, London, U.K., Tech. Rep., Mar. 2021.

[154] S. Munir, B. Batool, Z. Shafiq, P. Srinivasan, and F. Zaffar, "Through the looking glass: Learning to attribute synthetic text generated by language models," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics, Main*, 2021, pp. 1811–1822.

[155] A. Uchendu, T. Le, K. Shu, and D. Lee, "Authorship attribution for neural text generation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 8384–8395.

[156] G. K. Zipf, "Human behavior and the principle of least effort," Addison-Wesley Press Inc., Cambridge, MA, USA, Tech. Rep., 1949.

[157] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical detection and visualization of generated text," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Florence, Italy, 2019, pp. 111–116.

[158] M. Gallé, J. Rozen, G. Kruszewski, and H. Elsahar, "Unsupervised and distributional detection of machine-generated text," 2021, *arXiv:2111.02878*.

[159] A. See, A. Pappu, R. Saxena, A. Yerukola, and C. D. Manning, "Do massively pretrained language models make better storytellers?" in *Proc. 23rd Conf. Comput. Natural Lang. Learn. (CoNLL)*, 2019, pp. 843–861.

[160] *Natural Language Processing Benchmarks*, Papers With Code, Meta AI Res., Astor Place, New York, NY, USA, Oct. 2022.

[161] A. Radford, J. Wu, and J. Clark. (2019). *GPT-2 Output Dataset*. [Online]. Available: https://github.com/openai/gpt-2-output-dataset

[162] L. Kushnareva, D. Cherniavskii, V. Mikhailov, E. Artemova, S. Barannikov, A. Bernstein, I. Piontkovskaya, D. Piontkovski, and E. Burnaev, "Artificial text detection via examining the topology of attention maps," 2021, *arXiv:2109.04825*.

[163] T. Shamardina, V. Mikhailov, D. Chernianskii, A. Fenogenova, M. Saidov, A. Valeeva, T. Shavrina, I. Smurov, E. Tutubalina, and E. Artemova, "Findings of the the RuATD shared task 2022 on artificial text detection in Russian," 2022, *arXiv:2206.01583*.

[164] S. Skrylnikov, P. Posokhov, and O. Makhnytkina, "Artificial text detection in Russian language: A BERT-based approach," in *Proc. Int. Conf. Dialogue*, Jun. 2022, pp. 1–7.

[165] X. Chen, P. Jin, S. Jing, and C. Xie, "Automatic detection of Chinese generated essayss based on pre-trained BERT," in *Proc. IEEE 10th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, vol. 10, Jun. 2022, pp. 2257–2260.

[166] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Lang., Process.*, vol. 29, pp. 3504–3514, 2021.

[167] L. Martin, B. Müller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "CamemBERT: A tasty French language model," 2019, *arXiv:1911.03894*.

[168] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," 2020, *arXiv:2003.00104*.

[169] S. Dadas, M. Perełkiewicz, and R. Poświata, "Pre-training Polish transformer-based language models at scale," in *Proc. Int. Conf. Artif. Intell. Soft Comput.* Cham, Switzerland: Springer, 2020, pp. 301–314.

[170] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting Structured data*, vol. 1, Aug. 2006.

[171] A. Bakhtin, Y. Deng, S. Gross, M. Ott, M. Ranzato, and A. Szlam, "Residual energy-based models for text," *J. Mach. Learn. Res.*, vol. 22, pp. 1–40, Jan. 2021.

[172] A. Bakhtin, S. Gross, M. Ott, Y. Deng, M. Ranzato, and A. Szlam, "Real or fake? Learning to discriminate machine from human generated text," 2019, *arXiv:1906.03351*.

[173] T. Fagni, F. Falchi, M. Gambini, A. Martella, and M. Tesconi, "TweepFake: About detecting deepfake tweets," *PLoS ONE*, vol. 16, pp. 1–16, May 2021.

[174] S. G. Tesfagergish, R. Damaševičius, and J. Kapociute-Dzikiene, "Deep fake recognition in tweets using text augmentation, word embeddings and deep learning," in *Proc. Int. Conf. Comput. Sci. Appl.* Cham, Switzerland: Springer, 2021, pp. 523–538.

[175] J. Tourille, B. Sow, and A. Popescu, "Automatic detection of bot-generated tweets," in *Proc. 1st Int. Workshop Multimedia AI Against Disinformation*, New York, NY, USA, Jun. 2022, pp. 44–51.

[176] B. Auxier and M. Anderson, "Social media use in 2021," Pew Res. Center, Washington, DC, USA, Tech. Rep., Apr. 2021.
[177] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The Pushshift Reddit dataset," 2020, *arXiv:2001.08435*.
[178] P. Bhatt and A. Rios, "Detecting bot-generated text by characterizing linguistic accommodation in human-bot interactions," 2021, *arXiv:2106.01170*.
[179] M. Latah, "Detection of malicious social bots: A survey and a refined taxonomy," *Exp. Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113383.
[180] J. Salminen, C. Kandpal, A. M. Kamel, S.-G. Jung, and B. J. Jansen, "Creating and detecting fake reviews of online products," *J. Retailing Consum. Services*, vol. 64, Jan. 2022, Art. no. 102771.
[181] M. M. Bhat and S. Parthasarathy, "How effectively can machines defend against machine-generated fake news? An empirical study," in *Proc. 1st Workshop Insights Negative Results NLP*, 2020, pp. 48–53.
[182] J. Cutler, L. Dugan, S. Havaldar, and A. Stein, "Automatic detection of hybrid human-machine text boundaries," Univ. Pennsylvania, Philadelphia, PA, USA, Tech. Rep., 2022.
[183] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith, "All that's 'human' is not gold: Evaluating human evaluation of generated text," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 7282–7296.
[184] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. Smith, and Y. Choi, "Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 7250–7274.
[185] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J.-C. Chen, C. D. Castillo, R. Chellappa, D. White, and A. J. O'Toole, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 24, pp. 6171–6176, Jun. 2018.
[186] L. Dugan, D. Ippolito, A. Kirubarajan, and C. Callison-Burch, "RoFT: A tool for evaluating human detection of machine-generated text," 2020, *arXiv:2010.03070*.
[187] A. Uchendu, Z. Ma, T. Le, R. Zhang, and D. Lee, "TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2001–2016.
[188] OpenAI. (2020). *GPT-3 Github Repository*. [Online]. Available: https://github.com/openai/gpt-3
[189] R. Goodside, "Exploiting GPT-3 prompts with malicious inputs that order the model to ignore its previous directions," Tech. Rep., Sep. 2022.
[190] S. Willison, "Prompt injection attacks against GPT-3," *Simon Willison's Weblog*, Sep. 2022.
[191] J. P. O. Zapata, "GPT-3 prompt injection example," Tech. Rep., Sep. 2022.
[192] N. Japkowicz, "Learning from imbalanced data sets: A comparison of various strategies," in *Proc. Amer. Assoc. Artif. Intell. Workshop Learn. Imbalanced Data Sets*, vol. 68. Menlo Park, CA, USA: AAAI Press, Jul. 2000, pp. 10–15.
[193] C. Bellinger, S. Sharma, and N. Japkowicz, "One-class versus binary classification: Which and when?" in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, vol. 2, Dec. 2012, pp. 102–106.
[194] Y. Tay, D. Bahri, C. Zheng, C. Brunk, D. Metzler, and A. Tomkins, "Reverse engineering configurations of neural text generation models," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 275–279.
[195] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? Natural language attack on text classification and entailment," 2019, *arXiv:1907.11932*.
[196] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2018, pp. 50–56.
[197] R. Gagiano, M. M.-H. Kim, X. J. Zhang, and J. Biggs, "Robustness analysis of Grover for machine-generated news detection," in *Proc. 19th Annu. Workshop Australas. Lang. Technol. Assoc.*, 2021, pp. 119–127.
[198] *Directive on Automated Decision-Making*, TB Canada Secretariat, Ottawa, ON, Canada, Apr. 2021.
[199] *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation*, European Commission, Brussels, Belgium, Feb. 2018.
[200] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, "Four principles of explainable artificial intelligence," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. 8312, 2020.
[201] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.
[202] E. Crothers, H. Viktor, and N. Japkowicz, "Mean user-text agglomeration (MUTA): Practical user representation and visualization for detection of online influence operations," in *Proc. Int. Conf. Comput. Data Social Netw.* Cham, Switzerland: Springer, 2021, pp. 305–318.
[203] S. Stieglitz, F. Brachten, D. Berthelé, M. Schlaus, C. Venetopoulou, and D. Veutgen, "Do social bots (still) act different to humans?—Comparing metrics of social bots with those of humans," in *Proc. Int. Conf. Social Comput. Social Media*. Cham, Switzerland: Springer, 2017, pp. 379–395.
[204] J. Berger and K. L. Milkman, "What makes online content viral?" *J. Marketing Res.*, vol. 49, no. 2, pp. 192–205, Apr. 2012.

EVAN N. CROTHERS received the bachelor's degree in software engineering (minor in cognitive science) from the University of Waterloo and the master's degree in computer science with concentration in applied AI from the University of Ottawa, where he is currently completing a Ph.D.

He is currently a National Security Expert and a Senior Machine Learning Consultant with the Canadian Federal Government. His work focuses on protecting trust in online social spaces, countering radicalization, and safeguarding democratic elections. Actively engaged with both the Canadian public sector and academia, he is a Regular Member of the Program Committee of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) and a Reviewer of *Machine Learning*. He was a recipient of the Director's Merit Award, the highest departmental honor for protection of public safety, and has separately been recognized for his contribution to the Security and Intelligence Threats to Election (SITE) Task Force protecting democracies from threats to elections. His thesis research work was awarded first place for "Technology for the Digital Transformation of Society" with the University of Ottawa. He has presented his research at venues around the globe, including the United Nations "AI for Good" Platform.

NATHALIE JAPKOWICZ is currently a Professor of computer science with American University. She was with the School of Electrical Engineering and Computer Science, University of Ottawa, where she lead the Laboratory for Research on Machine Learning for Defense and Security. Over the years, she has supervised over 30 graduate students, received funding from Canadian Federal and Provincial institutions (NSERC, DRDC, Health Canada, OCE, and MITACS CITO), worked with private companies (Girih, Larus Technologies, Weather Telematics, TechInsights, and Ciena). She has published over 100 articles, papers, and books, including *Evaluating Learning Algorithms: A Classification Perspective* (Mohak Shah, Cambridge University Press, 2011) and *Big Data Analysis: New Algorithms for a New Society* (Jerzy Stefanowski, Springer, 2016). Biography courtesy of American University: https://www.american.edu/cas/faculty/japkowic.cfm

HERNA L. VIKTOR received the Ph.D. degree in computer science from the University of Stellenbosch, South Africa, in 1999. She is currently a Full Professor with the School of Electrical Engineering and Computer Science, University of Ottawa. She has over 20 years of experience in designing, implementing, and applying machine learning solutions in numerous and diverse domains. Her research interests include online machine learning for sequential and temporal data, fairness-aware learning, and class imbalance. Her work has received numerous recognitions and awards.