NewsGuard

BLACKBIRD.AI

MAN-MACHINE INTELLIGENT SYSTEM FOR SCALABLE IDENTIFICATION OF DIGITAL MEDIA HOAXES AND MISINFORMATION

OCTOBER 28, 2020

1.0 Introduction

The internet has become the primary delivery mechanism for news for a significant portion of the world's population. Social media distributes much of that content due to its low cost, ready accessibility, and optimized mechanisms for sharing. However, those benefits come with large risks: the digital public forum has minimal barriers to entry, including for those seeking to propagate misinformation. Moreover, the lack of governance leaves open the possibility of hoaxes being spread among a population that is vulnerable to being confused or misled because they lack information about the trustworthiness of sources of information. The problem is further compounded by the speed of transmission enabled by the internet: Digital conversations including hoaxes often evolve rapidly, spread widely, and mutate as they are propagated. The dynamics include the mechanisms of social media platforms whose algorithms actually prioritize conspiracies and other highly engaging narratives in order to maximize the platforms' user metrics.

In this report, we address an important aspect of this problem: the issue of hoax propagation, or the promotion of myths online that can have dangerous consequences. We present promising results from a joint study performed by NewsGuard and Blackbird.AI to **explore the combined power of human and artificial intelligence** to provide high integrity assessments of hoax content — at scale, and in a manner that can keep up with the fast pace at which information, or misinformation, is shared anywhere on the internet. We tested a man-machine intelligent system concept and evaluated its effectiveness in tracking the spread of two dangerous myths related to COVID-19.

Our findings illuminate the benefits of Blackbird.AI's artificial intelligence (AI) in providing automatic detection of hoaxes and related stories at scale when combined with NewsGuard's human experts crafting machine-readable narratives that provide a unique identifier, or "fingerprint," to empower Blackbird's AI tools to hunt down online hoaxes.

2.0 Partner Capabilities

The work described in this report was performed jointly by NewsGuard, the leading service providing ratings of the trustworthiness of sources of news and information online, and Blackbird.AI, a recognized innovator in the field of AI-driven analysis of narratives and social media networks for U.S. national security and enterprise. In the following sections, we summarize the capabilities of each organization as they relate to the collaborative work described in this report.

2.1 NewsGuard

NewsGuard's mission is to counter misinformation by rating all of the top sources of news and information based on basic, apolitical criteria of journalistic practice. In addition to creating for each news and information website Green and Red ratings, trust scores of 1-100 and Nutrition Labels with detailed information for news consumers, NewsGuard also produces its Misinformation Fingerprints product, which is an outgrowth of NewsGuard's ratings process and the constant updating of those ratings. These fingerprints leverage NewsGuard's unique bird's eye view of online misinformation from having rated—and continuing to rate—thousands of websites to catalog the most popular hoaxes being published online. The fingerprints are

written in a machine-readable narrative format, with related examples, keywords, hashtags and links to content that contains each hoax. NewsGuard approaches the challenge using a team of trained journalists, who collaborate in the review and assessment process illustrated in Figure 2-1 below.

Figure 2-1: NewsGuard scores sites and hoax content using a carefully curated review process executed by expert journalists.

NewsGuard develops two categories of assessments:

- ***Publisher Site Ratings:*** News and information websites are reviewed in ongoing fashion to assess their overall trustworthiness based on nine criteria relating to credibility and transparency, such as whether each site regularly publishes false content, the credibility of its sources, its publishing history and its ownership and control. These weighted criteria result in a trust score between 0 and 100 and an overall designation as Green (generally trustworthy) or Red (not generally trustworthy) ratings. These are explained in a "Nutrition Label" for each site. The ratings and labels are delivered to information consumers in various ways including through a browser plug-in, mobile version, APIs, and integrations into products such as those by Microsoft and mobile and broadband providers.
- ***Misinformation Fingerprints:*** NewsGuard catalogs all the significant myths or hoaxes found on the thousands of sites its journalists have rated and continue to update and turns each into a Misinformation Fingerprint™, a machine-readable narrative that captures each hoax's unique characteristics, such as narrative language expressing the hoax, multiple associated variations, hashtags related to the hoax, key words, and a narrative debunking the hoax citing authoritative sources.

NewsGuard's approach using trained teams of respected journalists to perform these tasks with precision has enabled the company to develop a strong reputation for being trusted, apolitical and transparent. Its ratings and labels are available through hundreds of public libraries, to tens of millions of students and teachers, to millions of households through digital platforms, and to internet service providers and healthcare systems. Moreover, NewsGuard was used by the World Health Organization to reach more than one billion people who had seen hoaxes in their social-media feeds to deliver trustworthy COVID-19 information to them to mitigate the "infodemic." And, in the process of producing these ratings and labels, NewsGuard has been able to produce these unique Misinformation Fingerprints

2.2 Blackbird.AI

Blackbird.AI is a multidisciplinary team of entrepreneurs, AI engineers and national security experts with an aligned interest around empowering the pursuit of information integrity across public digital media at internet scale. A core area of expertise is Natural Language Processing (NLP), which is applied to discover and analyze the discourse around emergent harmful narratives that appear across sources of Publicly Available Information (PAI).

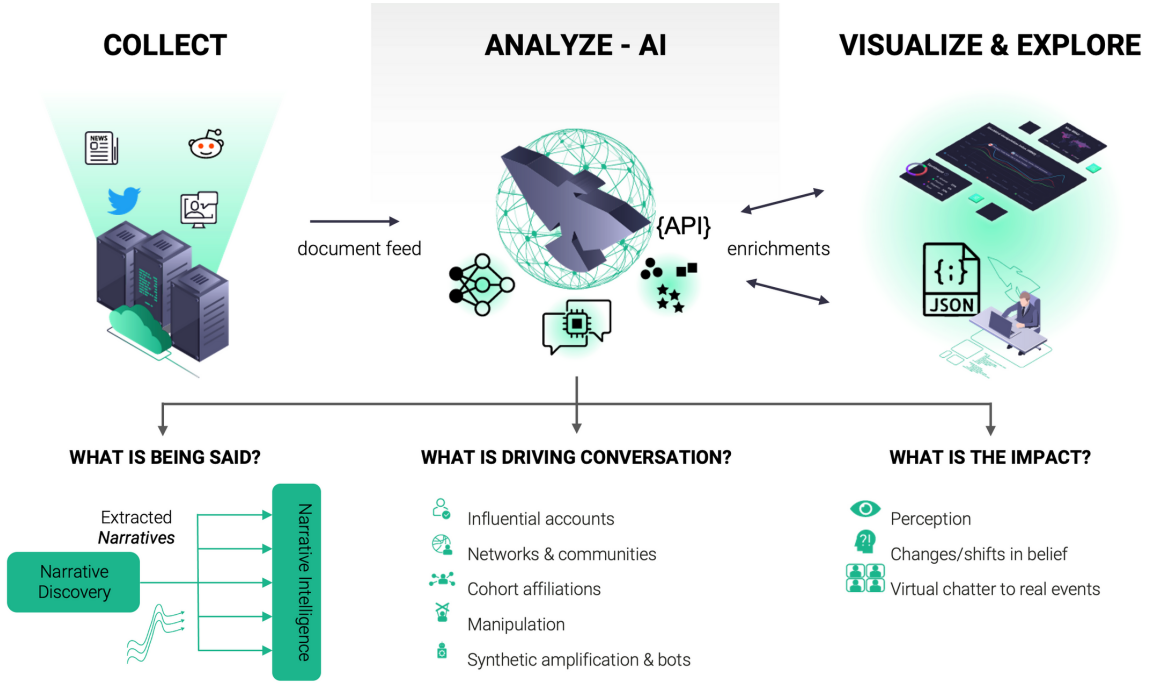

Figure 2-2: Blackbird.AI discovers and analyzes emergent harmful narratives in news and social media using artificial intelligence algorithms. Blackbird's tools aid clients in understanding what is being said, what is driving the conversation, and the resulting impact. Unique signals that are extracted include evidence of manipulation, cohort affiliations, network relationships and comparative intelligence over time.

Blackbird's Constellation Engine™ has a unique capability to detect the signatures of narratives that are being deliberately manipulated and promoted in a "propaganda-like" fashion, which is often evidence of directed misinformation agendas. This capability is based on proprietary measures referred to as the Blackbird Manipulation Index™ (BBMI) and the Blackbird Risk Index™ (BRI), which fuse together evidence based on patterns of propagation, user behavior, content analysis, and engagement. The platform also extracts specialized signals in social, news, and web data based on cohort analysis, user relationships and networks, and narrative intelligence.

3.0 Methodology

NewsGuard and Blackbird collaborated to explore the hypothesis that combining the strengths of their approaches could provide a superior solution to detect and track the spread of digital media hoaxes at a scale beyond even NewsGuard's ratings of more than 6,000 individual news and information sites, to reach the entire publicly available internet. To that end, they defined a system concept and evaluated it through testing hoaxes using the procedures described in the following sections.

3.1 Joint System Concept

A "Scalable Dynamic Fingerprinting System" was proposed with the components illustrated in Figure 3-1.

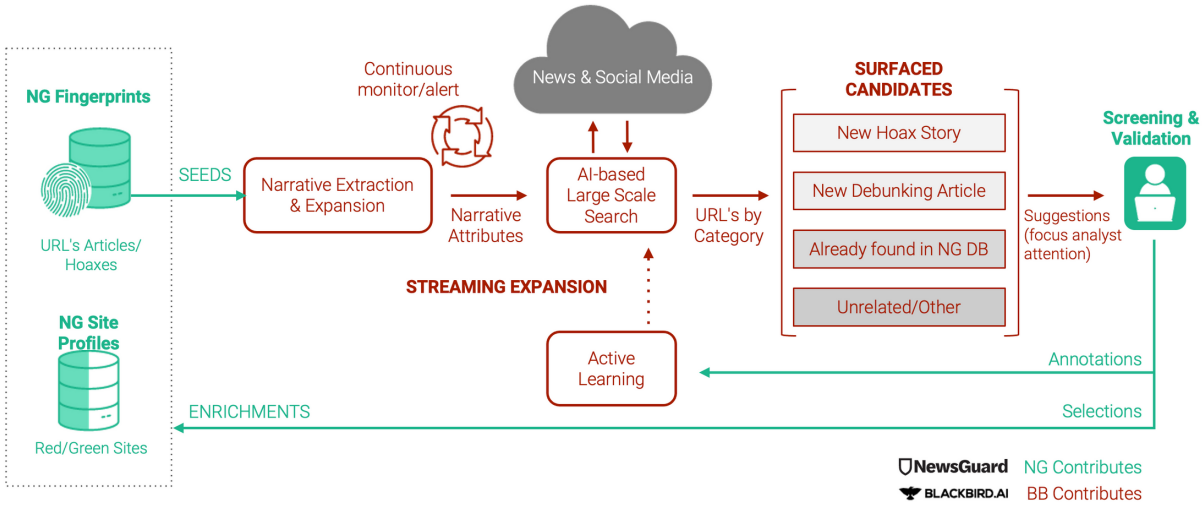

Figure 3-1: ***Scalable Dynamic Fingerprint System*** *Myth descriptions and sample URLs are used to seed an AI-driven narrative-based search of various news and social media platforms. URLs are surfaced that include candidate new example hoax stories and debunking articles to be reviewed by expert analysts, who in turn provide feedback to the system.*

The system is seeded using myth descriptions from the NewsGuard Misinformation Fingerprint database. The seeds include summary descriptions of the myth as well as sample URLs of the myth story and debunking narratives in articles that have been discovered by NewsGuard journalists, as described in Section 3.2 below. Automated analysis is performed on these seeds by Blackbird's system to derive attributes for AI-based narrative search at scale across a variety of news and social media platforms. The automated search process returns a set of URLs grouped by category, including new example stories related to the hoax, as well as new instances of debunking articles. The system may also rediscover examples that already exist in the NewsGuard database, as well as return examples that are not directly related but are similar in some way. In the case of the "unrelated/other" items, these articles may reflect strong mutations of the original hoax, or they may expose a related but new and possibly noteworthy hoax in its own right. The search results are reviewed by expert journalists who can decide to include the new findings in NewsGuard's database, and/or provide instructional annotation to the AI engine that will be used to improve the AI models through an active learning process.

The approach has several desirable characteristics:

- ***Scalable Content Monitoring:*** By combining NewsGuard's unique bird's eye view of misinformation narratives across thousands of news and information sites worldwide with Blackbird's AI capabilities, the system enables an end user to find instances of known, debunked falsehoods and misinformation narratives across social media platforms such as Twitter, YouTube, and Facebook, as well as the open web.
- ***Continuous Screening:*** The AI system can be operated in an "always on" mode continuously processing streaming input data (in near real time) so that it keeps pace with the constant evolution of digital conversation. Similarly, the fingerprints database is constantly updated and expanding as NewsGuard's human analysts identify and catalogue new misinformation narratives. In effect, this means that an end user combining NewsGuard and Blackbird in this way would benefit from a system that alerts them any time a known misinformation narrative appears on social media or the open web, and where it is appearing.
- ***Bootstrapped Learning:*** The AI models in the system receive feedback from the human analysts, which drives their continuous improvement in performance.
- ***Adjustable Specificity:*** AI models may be configured to lock tightly around specified hoaxes to monitor for those hoaxes, or they may be "opened up" to discover related or otherwise similar hoaxes. Likewise, depending on the end user, the system may be configured to focus on a wide stream of content—such as all social media and open web sources—or a more narrow set of content, such as just Twitter posts or just YouTube videos.

3.2 Targeted Myths

For purposes of this study, two myths were selected from NewsGuard's Misinformation Fingerprint database to seed the system. These were analyzed in the August 2020 timeframe.

1. NewsGuard Misinformation Fingerprint: Flu Vaccine Increases Risk of Contracting COVID-19

The Myth: People injected with the influenza vaccine have a 36 percent higher risk of contracting coronavirus, based on a study published in the journal *Vaccine* in January 2020 that used data from the U.S. Armed Forces Health Surveillance Branch.

Summary of Facts: There is no evidence that the flu vaccine either protects against the COVID-19 virus or increases the risk of infection. The claim relies on misrepresenting the January 2020 study, which actually covers seasonal coronaviruses that cause common colds, not the COVID-19 virus. The U.S. military data used in the study was collected during the 2017-18 flu season, well before the COVID-19 virus emerged in late 2019.

"The study does not show or suggest that influenza vaccination predisposes in any way, the potential for infection with the more severe forms of coronavirus, such as COVID-19," the Military Health System, which operates the U.S. Armed Forces Health Surveillance Branch, said in a statement to fact-checking website FactCheck.org in April 2020.

While some studies have found an association between the flu shot and non-influenza respiratory illnesses over the course of a single flu season, that association was not found in a larger study published in the journal Clinical Infectious Diseases in June 2013 that included data from multiple seasons.

According to the U.S. Centers for Disease Control and Prevention, "the preponderance of evidence suggests that this is not a common or regular occurrence and that influenza vaccination does not, in fact, make people more susceptible to other respiratory infections."

Associated Variations:
[Associated Variations are related myths. While each will have its own fingerprint in the NewsGuard misinformation database, they are noted here as well.]

- Judy Mikovits, a discredited molecular biologist and anti-vaccine activist, cited the same study in the 2020 documentary entitled "Plandemic." Mikovits claimed in the segment that "If you've ever had a flu vaccine, you were injected with coronaviruses."

2. NewsGuard Misinformation Fingerprint: <u>Wearing a face mask can cause hypercapnia</u>

The Myth:
Wearing a face mask can cause hypercapnia, a condition involving too much carbon dioxide in the bloodstream. Wearing a face mask to prevent the spread of the COVID-19 virus causes the wearer to continually to breathe in exhaled carbon dioxide, limiting the flow of oxygen and putting people at risk of hypercapnia. Wearing face coverings for long periods can therefore also lead to headaches, shortness of breath, and dizziness.

Summary of Facts:
In a June 2020 blog post on the American Lung Association's website, pulmonologist David G Hill wrote, "We wear masks all day long in the hospital. The masks are designed to be breathed through and there is no evidence that low oxygen levels occur." Mayo Clinic writes on its website that cloth face masks are a "very breathable" option for reducing the spread of the COVID-19 virus in public, noting that "carbon dioxide will freely diffuse through your mask as you breathe."

In May 2020 a representative from the U.S. Centers for Disease Control and Prevention (CDC) told Reuters that "it is unlikely that wearing a mask will cause hypercapnia." The representative noted that although carbon dioxide will build up in the mask over time, particularly if it is a medical-grade respirator, "the level of CO2 likely to build up in the mask is mostly tolerable to people exposed to it."

There is some evidence that wearing an N95 respirator for prolonged periods can cause discomfort. A small study published in the peer-reviewed journal Acta Neurologica Scandinavica in March 2006 found that some health care workers who wore N95 respirators during the 2003 SARS epidemic developed headaches, especially if they had suffered from headaches before the study. A CDC representative told fact-checking site Snopes in May 2020 that healthcare workers may experience headaches and difficulty breathing if N95 respirators are worn continuously for more than an hour, but noted: "To fix the problem of breathing too much CO2 that has built up within the respirator facepiece, a worker can simply remove the respirator."

Associated Variations:

[Associated Variations are related myths. While each will have its own Misinformation Fingerprint in the NewsGuard misinformation database, they are noted here as well.]

- Wearing a face mask causes hypoxia, a condition that arises from a lack of adequate oxygen.

3.3 Blackbird.AI: AI-Driven Narrative Detection

The hoax descriptions provided in Section 3.2, along with example URLs already found by NewsGuard's journalists that were associated with those hoaxes and associated keywords, were ingested by Blackbird's system to extract "narrative descriptors" to support AI-driven search for related content. For this study, the search was primarily focused on assessing the hoax's social media footprint, through Twitter and YouTube in particular, and also included some open web searches.

Blackbird's Constellation Engine™ scoured media for Lookalike Hoaxes found within NewsGuard's Misinformation Fingerprint. Blackbird's platform automatically surfaces relevant news, stories and social media posts that exhibit similar patterns to hoaxes in the NewsGuard Misinformation Fingerprint. The system improves in real-time with an analyst's feedback via active learning mechanisms and will train the platform to improve the detection of Lookalike Hoaxes that are similar to stories in the NewsGuard Misinformation Fingerprint. This system improves with analyst feedback continuously and will operate to collect Lookalike Hoaxes continually and at scale.

4.0 Results

A key measure of success for this initial feasibility study was to evaluate the effective identification of new instances of content, such as social media posts or videos, containing myths catalogued in the NewsGuard Misinformation Fingerprint database through the use of Blackbird's AI automation, backed with attention-focused human review and validation. The following figure diagrams the search expansion parameters for the two hoax fingerprints taken together.

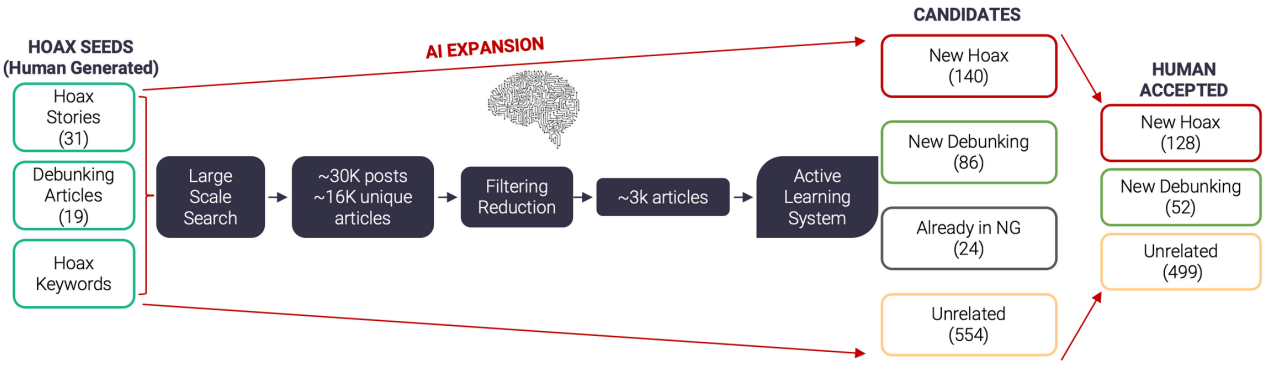

Figure 4-1: Search Parameters for the Two Hoaxes. Numbers shown in (parenthesis) indicate counts for the category. ***Within minutes*** *of document assembly, the system identified several hundred new instances of hoax and debunking articles, with 91% of hoax candidates accepted by analysts as correct (true positives).*

To establish a baseline, the AI pass was run a single time in batch mode over a date range of January 2019 to August 2020. Blackbird's system surfaced ~30,000 social posts (documents) with ~16,000 unique articles matching the narrative criteria. From there the system filtered down to ~3,000 articles displayed at least some level of social interactions (one favorite or retweet). Blackbird's system is capable of processing and assigning risk indicators to ~2,000 documents per second so the classification process once documents were assembled was completed within minutes.

The test of the system using the two example NewsGuard Fingerprints surfaced a wide range of instances of both hoaxes on Twitter, YouTube, and the open web. Efficiency in the system is driven by the number of documents the AI is able to screen and correctly categorize, balanced with the analyst workload to review the results. The operating point of the AI screening can be adjusted to increase detection rate on the hoaxes (surface true positive cases) at the cost of bringing in more false positives. For our initial study, our operating point selection was loose and led to fairly high counts of "unrelated/other" as is evidenced in Figures 4-1, 4-2, which while they might include discovery of interesting variants on the story, also increases analyst workload by providing more candidates needing review. There was no optimization of operating point performed, but we propose to explore this trade space more fully in future work described below.

For the setting of the study, 91% of the candidate hoaxes and 60% of the debunking candidates were accepted by analysts as correct. For example, the system flagged a post from a Twitter account with more than 97,000 followers claiming that flu vaccines increase the risk of COVID-19. The post had garnered over 800 retweets, quote tweets and 821 likes, and had not been taken down or fact-checked by Twitter, despite its policies against COVID-19 misinformation. On YouTube, the system flagged a video from the account for Children's Health Defense, a prominent anti-vaccine group with more than 24,000 YouTube subscribers, using the same hoax to question a statement from Dr. Anthony Fauci urging Americans to get a flu shot. YouTube had not taken any action against the video or against the account, which has published other, similar falsehoods in the past. The system also identified instances of the hoaxes on the open web. For example, it flagged an article from the website sott.net, which scores just 17.5 out of 100 points on NewsGuard's rating system, from April 16, 2020. The article had been posted and spread by Facebook accounts with over 100,000 followers.

As a byproduct of the automated search process, *variants* of the hoax stories may be discovered that are noteworthy in their own right. Some examples surfaced by the AI include:

Hoax - Flu Vaccine Increases Risk of Contracting COVID-19

- The regular "winter" flu vaccine was used "en masse" twice in the two months preceding the COVID outbreak in Bergamo, the Italian epicenter of their outbreak this spring, and by a company that is also working on a COVID vaccine.
- Flu vaccine does not protect against coronavirus but it does make patients that contract COVID much more susceptible to complications and more likely to die.

Hoax - Wearing a Face Mask can cause Hypercapnia

- Masks increase risk of hypoxia and hypercapnia, but also can increase susceptibility and contagiousness of COVID-19, as well as the severity of COVID symptoms once contracted.
- Prolonged wearing of a mask increases the risk of cancer growth due to cancer "growing best in a microenvironment that is low in oxygen," and low oxygen environments are known to promote the growth, invasion and spread of cancers.

5.0 Value Propositions

The ability to augment manual assessment of online claims with AI systems that can automate screening at scale and keep pace with the rapid evolution of online dialogue can provide benefits in several areas:

- ***Continuous Hoax Detection and Mitigation:*** The impact of hoax messaging can be reduced by tracking its evolution in a more granular fashion, including across social media user accounts; this could inform counter-messaging that debunks the myth directly to consumers of the false information, and/or directs countermeasures at the affected audience, such as through personalized advertising. (For example, working with the World Health Organization, NewsGuard was able to help deliver corrective WHO medical advice to more than one billion people who had been exposed to COVID-19 hoaxes in their social media feeds.)
- ***Early Detection of Emergent Harmful Narratives:*** Problematic narratives can spin up in minutes, spread widely and jump platforms quickly. A case of particular concern is a myth that originates in a "dark zone" of the social web, then migrates to a principal platform such as Twitter and finally breaks into mainstream media. AI automation can support human experts with continuous vigilance and early warning for such cases. This capability is of particular interest to entities that need to know what misinformation is spreading about them and how much engagement has resulted. This capability is widely relevant, including for brand-reputation managers and threat-intelligence analysts.
- ***Platform Content Moderation:*** Content distribution platforms are increasingly trying to avoid spreading misinformation. The system described here can assist them with cooperative man-machine problem solving that scales up to the problem in a manageable way and continuously improves their effectiveness. Platforms would be able to mitigate prominent hoaxes before they even become popular on their platforms by making content moderators aware of the current leading disinformation narratives on the internet and their platforms.

6.0 Summary & Future Work

This report has presented promising results from a feasibility study undertaken by NewsGuard and Blackbird.AI to investigate the power of partnered man-machine intelligence addressing the spread of dangerous misinformation in the form of myths or hoaxes on the internet, including through social media. Our chief finding is that a viable system concept can be built around the following man-machine collaboration principles:

- ***Artificial Intelligence Role:*** Support human analysts in screening large volumes of digital media to focus analyst attention on surfaced content.
- ***Human Intelligence Role:*** Provide trustworthy validation based on experience and context-sensitive screening, craft machine-readable narratives and debunkings, and make recommendations that help the AI learn superior models.

The next steps in our collaboration propose to address the following:

- ***Expand the Study:*** Investigate a wider range of hoaxes, monitor across time and continuously scan more platforms, and develop comprehensive performance and runtime metrics.
- ***Track Model Improvement over Time:*** As more data is collected and annotated by experts, rigorously evaluate improvements in AI model performance from active learning mechanisms.
- ***Incorporate Manipulation Attributes:*** Incorporate Blackbird's manipulation indices to detect synthetic amplification of messaging around hoax stories, for example by detecting user accounts that are "forcers," or botlike in nature, in order to provide another index of risk that can be used to prioritize hoax investigation and mitigation efforts.

Longer term, we see opportunities for our joint system concept to benefit from advances in the following technical areas:

- ***Transfer Learning:*** Rapidly bootstrap models for hoax narratives using transfer learning methods, which would be of great value in initializing the system on fresh hoax fingerprints.
- ***Confidence-Weighted Insertion:*** Explore opportunities to improve the scalability of the system through potential risk-versus-effort tradeoff whereby high-confidence detections can be inserted without review and low-confidence detections rejected without review, with ambiguous cases generating requests for analyst review.
- ***Multimodality Content Analysis:*** Extend the narrative search to incorporate additional content across modality, including image/video and audio track.

We live in a world in which the integrity of information has degraded such that the authenticity of much of the news and information on the internet can be called into question. This reality, if left unchecked, poses unprecedented threats to society, leading to erosion of trust in governments, social institutions, and in one another. NewsGuard and Blackbird were both born of a concern that it is essential to create mechanisms that enable trust in what has become a highly untrustworthy, chaotic global information ecosystem. Our teams are working on complementary approaches to the problem, with NewsGuard based on human intelligence and Blackbird based on artificial intelligence. When these approaches are combined appropriately, a responsible man-machine intelligence can be created that can handle the scale and complexity of the internet, continuously evolve to improve its own performance, and provide mechanisms of trustworthiness in a transparent manner that information consumers can audit and understand.


NewsGuard


BLACKBIRD.AI