978-9934-619-05-2

THE ROLE OF AI IN THE BATTLE AGAINST DISINFORMATION

Published by the
NATO Strategic Communications
Centre of Excellence

ISBN: 978-9934-619-05-2
Authors: Alfonsas Juršėnas, Kasparas Karlauskas, Eimantas Ledinauskas, Gediminas Maskeliūnas, Donatas Rondomanskas, Julius Ruseckas
Contributors: Luke Urbonavicius
Design: Kārlis Ulmanis

Riga, February 2022
NATO STRATCOM COE
11b Kalnciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook/stratcomcoe
Twitter: @stratcomcoe

This publication does not represent the opinions or policies of NATO or NATO StratCom COE.

Abstract

Detecting and countering disinformation grows increasingly important as social media sites have become a leading news source for most people. Efficient disinformation campaigns lead to negative real-world consequences on a global scale, both in politics and in society. Machine learning (ML) methods have demonstrated their potential for at least partial automatisation of disinformation detection and analysis. In this report, we review current and emerging artificial intelligence (AI) methods that are used or can be used to counter the spread and generation of disinformation, and briefly reflect on ongoing developments in anti-disinformation legislation in the EU. This overview will shed light on some of the tools that disinformation-countering practitioners could use to make their work easier. [1]

Table of contents

> AI can automate a wide range of specific tasks and significantly increase analysts' research capabilities. However, due to current limitations, AI can play only a supporting role by helping to process vast amounts of information and detecting what requires further attention.

Introduction

Due to the rapid pace of development and adaptation in the field of Artificial Intelligence (AI), the role it plays in *disinformation* practices is gradually increasing, boosting the work of malicious actors and analysts. We begin by defining *disinformation* as false or manipulated information that is created and disseminated in order to deceive,[2] i.e., to mislead public opinion about politics,[3] to divide and polarise society,[4] and to erode trust in public health institutions.[5] Disinformation practices evolve over time and adopt available technological advancements, including advancements in the field of AI.

Over the last ten years, the field of AI has enjoyed a series of significant and transformative breakthroughs, many of which have been applied to solving science and engineering problems.[6] In this paper, we use the following broad definition of AI proposed by the European Commission.[7] The term 'Artificial intelligence system' (AI system) refers to software developed using one or more of the techniques and approaches listed below:

- Machine learning (ML) approaches, including supervised, unsupervised, and reinforcement learning, using a wide variety of methods, including deep neural networks;
- Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning, and expert systems;
- Statistical approaches, Bayesian estimation, search and optimisation methods.

This definition includes both classic AI algorithms[8] and the relatively new methods based on deep artificial neural networks that have led to the most current breakthroughs in the field.[9]

AI-based tools provide practitioners with previously unavailable capabilities for analysing large amounts of data and exploiting complex patterns in large datasets. Currently AI methods are most successful in performing rather narrow, well-defined tasks, e.g., classification, regression, etc. Performing such tasks on large datasets of user activity has made it possible to create recommendation systems that select which information to display to maximise user engagement[10] (e.g., views, shares, likes, comments) and time spent on social media platforms. The AI algorithms used in recommendation systems have also made social network platforms more vulnerable to disinformation campaigns;[11] for example, the spread of highly emotional and divisive content is favoured to maximize user engagement.[12] AI is also used in generating increasingly realistic fake images, audio (voice imitation), video, and text, and can boost the ability of malevolent social bots to imitate human activity more realistically and to generate disinformation content at scale.

Fortunately, AI methods can also be used to counter disinformation. This includes detecting social bots, screening content for potential disinformation, performing deeper analysis that can detect modified versions of already debunked articles, modelling discussed topics, following hostile narratives, identifying AI-generated content (e.g., text, images, audio), and other activities. Narrowly focused AI tools have a great deal of potential for automating the many repetitive and time-consuming tasks performed by analysts countering disinformation.

The factors that continue to limit AI technologies are their reasoning and world-model capabilities. For example, existing AI tools require a substantial number of data examples to learn specific tasks such as recognising objects in a scene (e.g., detecting humans in a photo). Furthermore, AI models cannot filter out physically unrealistic occurrences of objects (e.g., detecting trains in the sky) because they lack a general 'world model'—a model that would contextualize and enable an AI model to identify the objects that exist within an environment and determine the likely dynamic changes these objects undergo. The more complex the environment, the more complex the model must be, so a world model is not usually included in current deep neural networks except in simplistic cases of reinforcement learning,[13] and is not likely to be included soon. This poses significant limitations for AI tools and can result in a 'lack of common sense' in unexpected cases. Researchers from OpenAI prepared an example using hand-written text and photographs of objects to demonstrate how the CLIP model can be fooled into making absurd classifications.[14] In this article, we analyse how AI methods are being used to counter disinformation campaigns, focusing on a group of selected topics for the sake of simplicity.

Detecting disinformation content

Disinformation content that is disseminated in a digital information medium can come from a variety of sources; it can be user-generated, scraped from the web and manually modified, or even computer-generated (synthetic). Analysts must scan vast amounts of information using various tools and software in order to discover and monitor a disinformation campaign. This is done by identifying patterns, classifying textual and audio-visual data, computing similarities between samples of content, and other techniques. AI models can be designed to support all these tasks (with varying success) and, thus, can serve as a powerful tool for analysts. In this section, we describe AI use cases in detecting various forms of malicious content and discuss methods for verifying and ensuring the authenticity of the original data being disseminated.

Text analysis and detection

In its simplest form, a disinformation detector can be thought of as working on text classification problems—an AI model is trained under supervision to classify text, i.e., assign the probability of predefined categories appearing in each piece of text it processes. Indeed, the majority of existing research entails supervised methods (AI models trained on data labelled manually by human annotators); semi-supervised and unsupervised methods are less commonly used.[15,16] Before diving into specific use cases, let's start by emphasizing the utility of representing text, or any other type of data, numerically as feature (embedding) vectors. These numerical representations obtained at word, sentence, or even document level represent spatial (semantic) relationships to other words/sentences/documents, which an AI model can use to cluster, classify, or even compute semantic similarities numerically instead of having to search for and compare specific keywords. Models such as Word2Vec and GloVe can be used to transform words into embedding vectors. For sentence-level embeddings one might use AI models such as Universal Sentence Encoder[17] or another model tailored for a specific language and use case.

In the context of detecting and analysing disinformation activities, both classical Machine Learning (ML) and Deep Learning (DL) models are widely used. However, over the last four years attention-based neural network models in Natural Language Processing called Transformers[18] have demonstrated high levels of accuracy.[19] Some studies demonstrate improved accuracy by developing more efficient ways to utilise meta-data, such as speaker credibility and information about online social interactions.[20]

Zellers et al. proposed a new model, named Grover, for generating fake text using Generative Pre-trained Transformer 2 (GPT2)[21] architecture; they showed that the overall trustworthiness score of disinformation increases when rewritten by the Grover text generator. The authors also found that the Grover neural network is able to detect computer-generated articles effectively. They argue that to combat AI-generated fake news, access to the generators is critical. However, OpenAI[22] challenged the supremacy of GPT-2 generated texts by showing that fine-tuning a RoBERTa-based detector [23] achieved consistently higher accuracy than fine-tuning a GPT-2-based detector with equivalent capacity. Jwa et al.[24] proposed a model for disinformation detection based on BERT transformer architecture[25] and fine-tuned on CNN and Daily Mail news data. The relationship between the headline and the body of the news text was analysed. Research by Marcellino et al.[26] introduced an improved model for efficiently detecting topics related to conspiracy theories based on a hybrid ML model that combined BERT word embeddings (numerical feature vectors) with linguistic stance markers obtained from a RAND-Lex textual ML analysis tool able to combine qualitative content analysis with pattern identification, tone, and sentiment estimation in word use.

Fagni et al.[27] collected a dataset of deepfake tweets—TweepFake—to evaluate 13 computer-generated (deepfake) text detection methods. The results showed that automatically distinguishing between human-composed and computer-generated tweets is challenging due to continual improvements in the performance of generators and to the limited length of the tweet. Disinformation detection exploits the stylistic biases that exist in a text.[28] AI text generators often introduce artefacts into their texts, which can be learned and recognised by discriminators.[29] However, the datasets used to train models are likely to be biased, and this can cause errors in detection.[30]

Furthermore, it is insufficient to approach the detection of disinformation as simply a matter of identifying machine-generated text. The most intrinsic characteristic of disinformation is that 'truth' is made ambiguous by introducing false and/or misleading facts, not whether a text is human- or machine-generated.

Disinformation detection using a knowledge base

Another approach to detecting fake news is to employ a knowledge base of verified facts or articles. Ghosh et al.[31] introduce a fake news detector consisting of two submodules: a veracity detection submodule based on information retrieval models and a style-based submodule. The veracity checking consists of two steps: the most relevant documents are retrieved from a carefully prepared knowledge base, and, given the true (i.e., factual) information those documents contain, the veracity of a claim is inferred. Shaar et al.[32] propose a model that learns to rank relevant documents to detect previously fact-checked claims. The BERT transformer neural network is used as a sentence encoder to obtain a numerical representation for an input text. The cosine similarity is computed for ranking between the numerical representation of input claims and verified claims in the dataset. An interesting product of this kind is called FactSparrow,[33] introduced by Repustar,[34] which uses a Twitter bot as a fact request and delivery mechanism; anyone can mention the FactSparrow bot in social media conversations and retrieve relevant facts of the topic discussed.

Detection using a knowledge base can also greatly benefit from Named Entity Recognition (NER)—a specific task where the AI model extracts useful information (proper names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.) from vast amounts of raw unstructured textual data. Entity-level sentiment analysis[35] (assigning numerical values to expressions of sentiment and aggregating them for analysis so that a document can be given a positive or negative score and high or low sentiment magnitude value) makes it possible to analyse and compare texts at a more granular level. Using AI to further extract entity relations enables the building of advanced knowledge graphs[36] for efficient information extraction and visualization. Thanks to growing shift towards AutoML solutions,[37,38,39] less tech-savvy AI enthusiasts can benefit from state-of-the-art AI models and shift focus from AI model parameter tuning to the development of specific datasets.

Reasoning-based detection of disinformation

Disinformation and misinformation can, in principle, be detected by looking for inconsistencies between the text in question and a list of known facts and user-defined ontologies (a set of facts about objects that exist within a defined "world", their categories and relations).[40] Inconsistency detection may be especially effective when detecting automatically generated text as current generative AI models contradict themselves when generating long texts. Checking for such inconsistencies could, in principle, be done by using reasoning algorithms based on formal logic. In practice, however, it is challenging to construct formalised fact or ontology databases that can be applied to general text on the internet. Nevertheless, such an approach might be practical when used in narrow domains (e.g., COVID-19 or the history of a particular country). Groza and Pop[41,42] have proposed a design and prototype for a reasoning-based disinformation detection system for the medical domain. The system uses natural language processing (NLP) techniques to convert a natural language text into formal logic statements; these statements are double-checked for inconsistencies as mentioned above. Such a formalised fact and ontology database in a narrow domain could be obtained semi-automatically by choosing trusted sources and extracting formal logic statements using NLP techniques. However, the performance of such systems would be highly dependent on the accuracy of the translations from natural language into formal logic constructs, which is a complex and challenging problem. Moreover, this system alone would not be able to detect disinformation aligned with known historical or scientific facts (e.g., new made-up events) as new disputed claims would be checked for consistency with previously confirmed facts. For example, if COVID-19 is a new previously unknown virus, then statements about it would still be consistent with what is known about viruses in general.

Deepfake detection: Images, Audio, Video

Detecting synthetic and manipulated images

With the increasing performance capability and popularity of generative models in computer vision—the field of AI concerned with training computers to interpret and make sense of the visual world—detecting fake images and videos has become an important problem. In 2019, Facebook, Amazon, Microsoft, and the non-profit organisation Partnership on AI announced a public deepfake detection challenge.[43] Since then, efforts have been made to develop better datasets and more reliable detection systems. For example, the FaceForensics++ dataset[44] was created to address several issues relevant to manipulated image detection. The dataset consists of raw, undoctored images taken from 1000 different videos. Then the same images were altered using several popular forgery methods, including face swapping and facial re-enactment (transferring facial expressions from one face to another while retaining identity). Due to compression algorithms pre-process most media uploaded onto social networks, the dataset also contains images of various qualities processed with the commonly used H.264 codec. Unlike humans, an AI can detect fake or manipulated images even with considerable compression. The FaceForensics++ dataset also contains pre-trained models that enable transfer learning, which is crucial as new manipulation methods emerge and pose new threats to learning-based detection methods.[45]

Older deepfakes (such as those appearing on thispersondoesnotexist.com) can be detected through simple manual inspection because of the visual inconsistencies common in images generated using older models, such as background distortion, inconsistent earrings, poorly defined features around teeth, and other tell-tale signs.[46]

One of the most accurate deepfake detectors[47] uses an Expectation Maximization (EM) algorithm to analyse convolutional traces ('sort of a unique feature fingerprint left in the image during the image generation process');[48] it was tested using fake images generated through AttGAN, GDWCT, StarGAN, StyleGAN, and

> ” With the increasing performance capability and popularity of generative models in computer vision […] detecting fake images and videos has become an important problem

StyleGAN2, achieving 99.81% accuracy. Neves et al.[49] performed an experimental assessment of facial manipulation detection performed by different state-of-the-art detection systems in various experimental conditions. In controlled scenarios, the tested systems achieved results similar to the best previous detection studies. However, in more challenging scenarios, their performance quickly declined. One review[50] noted that most existing studies claiming high effectiveness for one or another model in detecting DeepFakes do not generalise well in cases of unseen DeepFakes and are not robust in detecting image/video transformations. The best model from the Facebook deepfake detection contest had an accuracy rate of 82%. However, when the same algorithm was tested against a set of previously unseen deepfakes, its performance dropped to 65%.[51] New techniques are rapidly emerging and deepfake creators might target a particular detector, so the most sophisticated deepfake detectors use multiple detection models.[52]

Detecting synthetic audio

In recent years, voice cloning through voice conversion (changing one person's voice signal into another) and text-to-speech generation methods (synthesising audio that corresponds to text input with the voice of a targeted person) have produced high-quality results.[53] The most advanced of these methods are based on deep learning models,[54] so they are sometimes referred to as deepfake voice technology. Authentication is a form of biometric identification in applications that use automatic speech verification (ASV) systems. For instance, in 2013, attackers tricked an employee into transferring money from their company's bank account using a deepfake voice imitation of that person's boss.[55] Due to the continuous evolution of spoofing attacks, more sophisticated detection algorithms are needed. Research in the automatic detection of spoofed audio has shown that neural network models can achieve a 2.19% error rate when classifying genuine and spoofed speech.[56]

Deep learning models are being used to detect synthesized speech. One approach is to convert the audio recording in question into an image, or spectrogram, by computing the distribution of various sound frequencies over time. A Mel-spectrogram is a 2D graph created from an audio recording showing time as one axis and sound frequency as the other. This turns the fake audio detection problem into an image classification problem that can be solved by a deep learning model.[57] Alternative approaches to synthetic audio and video detection use AI-modelled latent representations (features). An example of such an approach has been demonstrated by Chintha et al.[58] The proposed model, CRNNSpoof, processes raw audio signal directly; it achieved a 4.24% error rate in the ASVspoof 2019 challenge, successfully outperforming the baseline models. Chintha et al.[59] observed a typical AI/ML problem—while fake voice detection methods are very accurate when used with training data, they perform poorly on examples that differ from those demonstrated in the training dataset.

Detecting synthetic or manipulated video

Fake videos have become increasingly realistic to the point where it is often difficult to spot them with the naked eye. A study by FaceForensics++[60] demonstrated that some outputs of forgery methods (e.g., Face2Face)[61] were difficult for human observers to detect; in these cases, the detection rate was close to random guessing. It is likely that, in a mundane context, such videos would not raise suspicion among casual viewers. However, various ML methods can be used to detect extremely subtle clues. A number of fake video detection models are publicly available. For example, fake video forensics tools allow users to input video URLs and receive an evaluation of a video's authenticity.[62,63]

Yang, Li, and Lyu[64] have developed a method for detecting deepfake videos based on inconsistencies in the predicted 3D orientations of people's faces. The positions of facial landmarks (corners of the lips, the tip of the nose, etc.) are not necessarily preserved when a deepfake model transforms one face into another. Deepfake models generate the central part of the target face and then superimpose it onto the original footage.

Temporal information can also be helpful when detecting manipulated videos, even if isolated frames of a video are hyper-realistic.[65] For example, mouth movement not matching the speech, unnatural eye blinking patterns, etc.

Fingerprinting data to preserve authenticity

Deepfake image detection based on generative model artefacts is unsustainable in the long run because of the improving performance of generative AI models and evolving detection countermeasures (e.g., Carlini and Farid).[66] A more viable alternative might be to ensure that generated content can be traced to its source via fingerprinting techniques. Yu et al.[67] propose a general method for artificially fingerprinting images generated by GANs. This 'fingerprinting' procedure uses deep learning-based steganography to embed hidden information into the training data of the GAN models; during training they learn to embed these signatures into the images they generate. Such signatures could easily be detected and identified using a specific decoder neural network. Importantly, they are undetectable and non-removable without access to the decoder. Authors claim that this approach is very general and could prevent the malign use of published pre-trained models.

Yu et al.[68] argue that GANs leave specific fingerprints even without using such special measures. They show that the fingerprints depend on the architecture of the model used and various other details related to training. This means a classifier can be trained to attribute images to the specific GANs that generated them. However, this technique can only detect images generated by the specific GANs that were used to train the classifier. Thus, to successfully apply this technique, one would need access to a generative model or to many generated images. Authors claim this can be done by querying different services for generated images and labelling them with the name of the service. Then, the classifier could be trained to test image authenticity by predicting whether an image has been created by a specific service. Moreover, it could test service authenticity by checking if the generated images they provide contain the appropriate signature.

Another possibility for combating disinformation and misinformation using data fingerprinting techniques is to establish credibility by linking media content to verified sources; then published content can be compared to the original to determine if it has been modified. Project Origin[69] (founded by Adobe, Arm, BBC, Intel, Microsoft, and Truepic) aims to create an open standard for measuring accountability through media provenance. They propose that media items be registered using digital fingerprints as part of the publication process. Content creators would receive a certification of authenticity, or digital fingerprint that is stored in a cryptographically secure distributed ledger. This fingerprint would be embedded into each media publication before distribution so that a web browser or a dedicated application could automatically compare the fingerprint of the publication with the original stored in the distributed ledger and ensure the credibility of the data.

Detecting how disinformation spreads: bots and sockpuppets

Monitoring websites outside social media

Web scraping enables the automated monitoring of disinformation campaigns outside of social media platforms. One example of such monitoring would be identifying and clustering web pages that distribute near-verbatim copies of the same article. This could be done by scraping selected websites and looking for similarities in their source code patterns (e.g., HTML, JS) using popular scraping tools (e.g., Scrapy)[70] or browser automation tools (e.g., Selenium).[71] The similarity between articles can be measured using deep learning or classical NLP tools.[72] There are also many online tools that detect similar websites (e.g., SimilarSites),[73] but it is better to scrape more dynamic websites for changes, since the online tools might not frequently update their databases.

If scrapping is not an option, then news monitoring datasets could be used. The largest dataset is the Global Database of Events, Language, and Tone (GDELT),[74] which scrapes news worldwide and stores this data in a compact Conflict and Mediation Event Observations (CAMEO)[75] format.[76] GDELT data can be accessed directly or by using cloud services such as Google BigQuery[77] and Amazon S3.[78] The downside to cloud services is that analysts are unable to use network analysis algorithms efficiently due to the complexity and limitations of SQL queries.[79] Another drawback is the cost of data processing in cloud services. Researchers from the Simula Research Laboratory in Norway have developed a data mining system for the GDELT dataset to counter these limitations.[80] They have also proposed looking for disinformation by clustering news sources that frequently report the same events. This approach has enabled them to analyse about one billion news articles over the last four years.

Sockpuppet detection

A sockpuppet is a non-automated fake online social network (OSN) account created by a single nefarious entity, or puppetmaster. Sockpuppets can be used, for example, to spread spam or disinformation, or to impersonate an organic discussion with many participants, intended to produce a misleading impression of a prevailing consensus opinion (astroturfing).[81]

As a single entity usually controls numerous accounts, sockpuppets can be detected by looking for similarities in account-generated verbal content and activity patterns[82] linked to certain account profiles[83] or social network structures.[84,85] Most of these approaches use machine learning methods.[86] The social network structure approach assumes that a single sockpuppet account will create similar social connections to other accounts administered by the same puppetmaster. This detection method analyses pairs of accounts and looks for similarities in their social graphs.[87]

To detect stolen accounts being used as sockpuppets, analysts can look for behavioural anomalies.[88] For example, user activity (posting, page visits) within a social network can be tracked and collected, and then anomaly detection models can be applied to identify when a user account was hijacked.[89]

Puppetmasters anticipate the possibility of being caught and take steps to modify their behaviour to avoid detection. They may change the profile information and the verbal features of a stolen account and are likely to alter their writing style intentionally.[90] If non-verbal activity patterns are the basis of detection (e.g., posting behaviours such as post frequency, timing, and user activity patterns),[91] then puppetmasters can adjust these patterns to better match the behaviour of real users.

Due to the wide variety of real-user behaviour and the effectiveness of sockpuppet operators, all of these approaches can produce false positives, i.e., real users being labelled as sockpuppet accounts. The results demonstrated in recent studies are encouraging, but it is likely that detection tools must be regularly retrained to keep up with novel user activity patterns and must also be specialised for specific countries and languages. False positives remain a problem in bot and sockpuppet detection.[92]

Bot detection

A 'social bot' is a fully-automated social network account; partially-automated accounts are usually called 'cyborgs'.[93] There are numerous scientific publications proposing methods for bot detection. Orabi et al.[94] counted at least 53 articles published between 2010 and 2019 that focus on fully automated accounts only.

In reviews carried out by Cresci and Orabi et al.,[95] bot detection methods that use machine learning are first categorised according to whether they are supervised, unsupervised, or semi-supervised, and then further divided into content-based or behaviour-based detection methods. Both reviews also mention detection methods that do not depend on ML techniques; these are crowdsourcing-based methods and human-crafted network structure analysis.

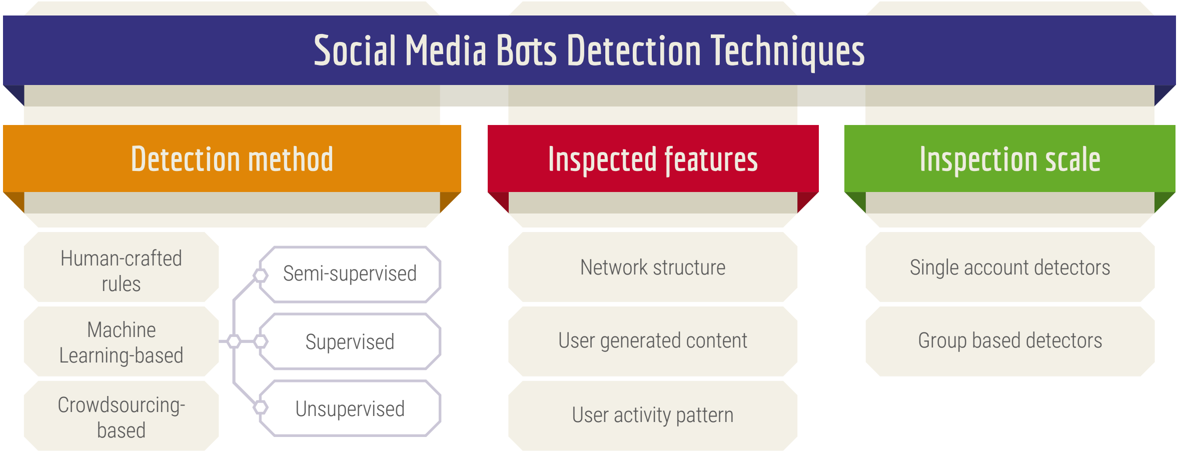


Figure 1. Social media bot detection techniques categorised by method, inspected features, and inspection scale. Partially adopted from Cresci[96] and Orabi et al.[97]

Single bot detectors: identifying single-user accounts

A popular single bot detector is a binary classification model named Botometer[98] (formerly BotOrNot). Botometer provides a public application programming interface (API) that any Twitter user could use to estimate the likelihood of an account being a bot. Because this is a binary classifier, users set the probability threshold value for determining whether an account is a bot; this means the threshold can be fine-tuned to match the population of interest. However, as users are setting this value freely, a degree of uncertainty should be expected. Binary classifiers trained on specific datasets were shown to have generalisation problems and should be carefully tested and/or retrained before use.[99]

Group bot detectors: identifying the global coordination of multiple accounts

The difficulty in detecting bots one by one led to experiments in identifying anomalous synchronisation patterns among multiple OSN user accounts.[100] One way to measure the similarities among account activity patterns is to measure the longest common subsequence of identical actions (e.g., post, share, like, etc.) that accounts share.[101] Unusually long sequences of actions indicate that a single operator controls all of the bots that share this behaviour. Such accounts can be clustered together and mapped to identify an entire botnet. Another study[102] of coordinated bot detection examined 160,000,000 tweets from ten state-attributed campaigns by training bot network detectors on statistical features extracted from social network activity (e.g.,

> ”Social media companies are best positioned to create accurate bot detectors because they have full access to the users' information [...]. However, independent analysis is essential if bot detection is to become more transparent.

retweets, co-tweets, co-mentions, and others).[103] The study concluded that botnet tactics vary over time and strongly differ between campaigns. There was a reduction in bot detection performance when countering novel bot networks. Even so, global bot detection is a viable alternative to single account analysis, especially as the identification of an entire bot network is more valuable than detecting multiple bots that are not necessarily related to each other.

Difficulties in bot detection

Difficulties in developing datasets suitable for bot detection

Obtaining detailed social media data is difficult due to social network policies and legitimate concerns about user privacy. Most bot detection publications focus on Twitter[104] because the platform is comparatively open.[105] The majority of bot detectors use a supervised learning methodology that requires AI models to be trained on manually labelled datasets.[106] Thus supervised learning methods currently require intensive work from human annotators and the datasets must be sufficiently large and diverse to avoid bias but not so large as to overfit the data necessary for the model.

Furthermore, it is necessary to update these datasets regularly to reflect the latest patterns in social media as bot algorithms are continually being improved.[107] The task is further complicated by the issue of human bias in labelling during dataset development. To properly analyse a dataset, precise definitions of bots and regular accounts must be used. This is especially problematic when detecting non-binary cases (e.g., cyborgs). It is likely that smaller countries and communities will need to develop customised datasets relevant to them so they can detect bots within their local environments.

The transferability of bot detection models and their stability over time

As demonstrated in a study by Rauchfleisch & Kaiser,[108] there is a drop in performance when a bot detection model created in one country is used for a different country or language (e.g., Germany instead of the US). Different cultures may exhibit significant differences in communication style; therefore, supervised learning models tend not to generalize well across multiple cultures and languages.

Similar deficiencies can be expected in most supervised binary classifiers trained on limited data. Difficulties in detecting bots also increase as public discourses evolve, vocabulary changes, and improvements are made to bot software.

Some humans act like bots

Some groups of human accounts, such as campaigning politicians and social activists, are very active on social media and can sometimes reach a level of activity similar to that of social bots. Furthermore, privacy-conscious users prefer profile names containing random digits and characters and use profile pictures without a human face, which makes them appear similar to simple bots.[109] However, since even bot-like individuals are not likely to behave like coordinated bots in a botnet, techniques that detect the anomalous coordination of multiple accounts will not return false positives for such accounts.

Improvements in bot software

Some studies have shown that it is difficult for even tech-savvy users to identify the most advanced bots (24% accuracy), while the same users could spot older bots with 91% accuracy.[110] We expect bot developers to take advantage of the growing number of open-source AI tools for content generation, which will make detection more challenging.

Bot detection scores obtained from bot detector tools and software should be interpreted carefully and treated as indicative. As Rauchfleisch & Kaiser state,[111] detectors must be periodically revalidated due to shifting behaviour in both users and bots. Social media companies are best positioned to create accurate bot detectors because they have full access to the users' information (e.g., patterns of communication with other users, IP address, browser information). However, independent analysis is essential if bot detection is to become more transparent.

Explainability of disinformation detections by AI methods

Once an AI system has detected a potential source of disinformation, it is crucial that the decisions made by the AI are explained to platform administrators; this will help them assess the likelihood of algorithmic bias in a particular decision. Such an open approach will increase trust and give administrators more confidence in AI tools. Information sharing can also help platform administrators explain why an account has been classified as committing inauthentic actions.

The limitations of eXplainable Artificial Intelligence (XAI) methods

Many methods have been developed to explain the decisions made by AI algorithms. One method overlays heatmaps onto suspected fake images or tabular data to show which regions (or cells) contributed most to the image or table being placed in a particular class (e.g., what areas show that the image is a deepfake).[112] However, XAI methods are still maturing. Some methods in AI are explainable by design, for example, decision trees. But the more explainable other models become, the more their performance is reduced.[113,114] There are several points worth mentioning here. First, there is a need to define what counts as an explanation and how detailed an explanation should be; having reached a certain level of detail, an explanation becomes difficult or impossible to interpret. Second, explanations might be leaked to adversarial actors who can exploit them to improve their disinformation techniques.

Adversarial attacks against explainability

Deep neural networks are sensitive to specific selected changes in the input data. This sensitivity can be exploited by so-called adversarial examples[115]—input that is changed to maximally distort predictions. Adversarial examples also exist for AI explainability: these are modified inputs (e.g., images or other data) that have been tampered with to cause the method of explanation to fail and indicate incorrect explanations.[116,117] Methods to counter adversarial XAI examples are also being developed.[118,119]

Current XAI methods have significant limitations, but in certain cases XAI models can provide valuable information to platform administrators so they can judge the reliability of the AI decisions that detected a disinformation campaign. XAI methods could potentially be used to create more transparent recommendation systems (e.g., ads, YouTube videos, and other types of content) and provide greater AI model transparency for platform administrators.

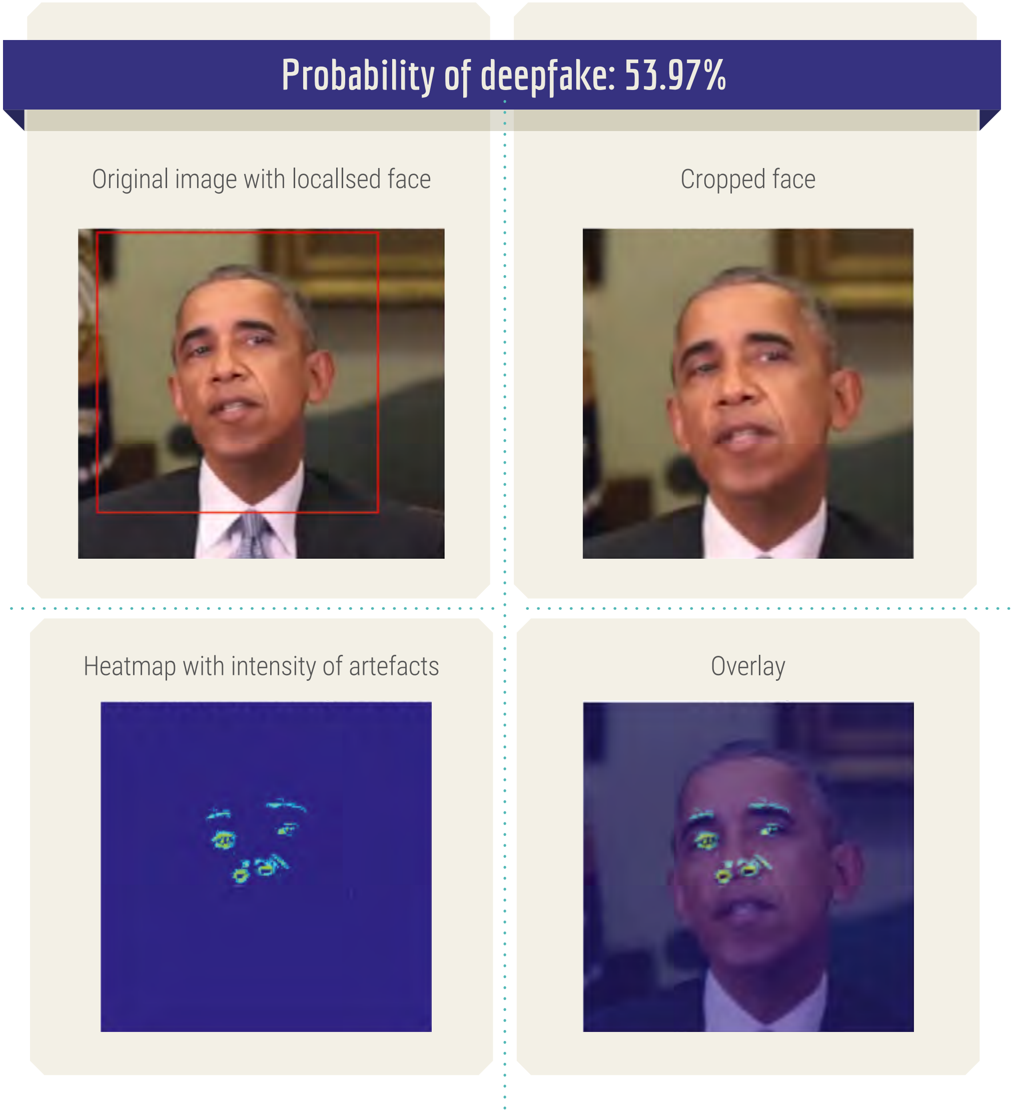

Probability of deepfake: 53.97%
Original image with locallsed face
Cropped face
Heatmap with intensity of artefacts
Overlay

Figure 2. Example of a heatmap showing which regions contributed to an image being classified as a deepfake. Source: Reprinted with permission from DuckDuckGoose.nl.

Legal frameworks against disinformation

There is an ongoing discussion about the responsible development of AI and how (much) it should be regulated. On the one hand, regulations that are too strict might impede the development of AI technologies and limit their benefit to society and to various critical infrastructure stakeholders. A much-debated example is the General Data Protection Regulation (GDPR), adopted by the EU in 2016. It obliges data brokers to collect and process personal data responsibly. It puts limitations on sharing and selling personal data, which could create a relative disadvantage for European companies developing face recognition algorithms or similar AI-based technologies due to the large quantities of personal data needed to train AI models.

On the other hand, it is important to develop AI responsibly and inclusively. Algorithms might privilege certain content as they are trained on data produced by humans who have inherent biases. For instance, algorithms that disproportionately flag content from specific political, religious, or ethnic groups may be inappropriate or even discriminatory;[120] this should be avoided. The first regulation addressing such issues and seeking to harmonise rules regarding AI systems is being discussed in the European Union. If adopted, the document would classify AI practices into four risk groups and place requirements on each group according to its risk level, including the prohibition of specific AI practices harmful to safety and to fundamental human rights.[121] Regulations on data policy and AI will also be relevant for disinformation campaigns, as data availability or algorithms (dis) privileging a certain kind of content could empower or prevent AI-based targeting and the dissemination of disinformation. Therefore, any regulation concerning AI must take into account the impact it might have on disinformation practices.

Governments could make better use of legislation to fight (AI-powered) disinformation. While it is hard to punish malevolent online actors due to the cross-border nature of virtual space and the challenge of attribution, the digital platforms (which these actors use) could be obliged to take specific measures to fight disinformation better themselves or to provide others with disinformation countermeasures. Examples include disinformation reporting tools, stricter data user authentication, content monitoring and authentication systems, greater transparency regarding data sources shared on the platform, educational campaigns on disinformation, data sharing with official authorities, and other measures.[122]

However, implementing stricter authentication protocols would complicate access and might discourage some users from using a particular platform out of concern regarding platform access to private data and would place a greater burden on the platforms to administer sensitive personal data responsibly. Sharing data with scholars or official institutions might lead to the loss of commercial advantage, could potentially be used by authoritarian regimes to track and persecute users, or might help malevolent actors discover and exploit a platform's weaknesses. Therefore, the measures enforced by law should be well-considered and seek to find the correct balance between addressing the need to fight disinformation and respecting the rights and legitimate interests of online platforms and their users.

Testing AI legislation through dialogue with online platforms. In 2018, the European Commission and major online actors agreed to test potential legislation by adopting the Code of Practice on Disinformation. This is a regularly reviewed voluntary document that sets standards for fighting disinformation. The signatories adopt roadmaps, choose measures, and provide reports about what they have done to make their platforms more resilient against disinformation and how well these measures have worked. As of 2021, Meta, Alphabet, Twitter, Microsoft, Mozilla, TikTok, and others, have all become signatories of the Code. This document should become a co-regulatory instrument for the digital market, along with the Digital Services Act,[123] when the latter is adopted. There are three reasons the Code is a useful example when discussing further legislation. First, such mechanisms allow online platforms to be heard, to actively set standards, and to test how they work in practice. Second, they give regulating authorities an opportunity to learn what works before including new standards in legally binding regulations. Third, the experiment shows how much can be achieved through voluntary means. Thus, by entering into dialogue with the industry, regulators could gather enough test data to dismiss unfounded concerns (for users and the industry) before enacting anti-disinformation legislation and could pinpoint those areas where voluntary mechanisms might be not enough to produce the desired result.

Discussion: How AI can boost the work of analysts

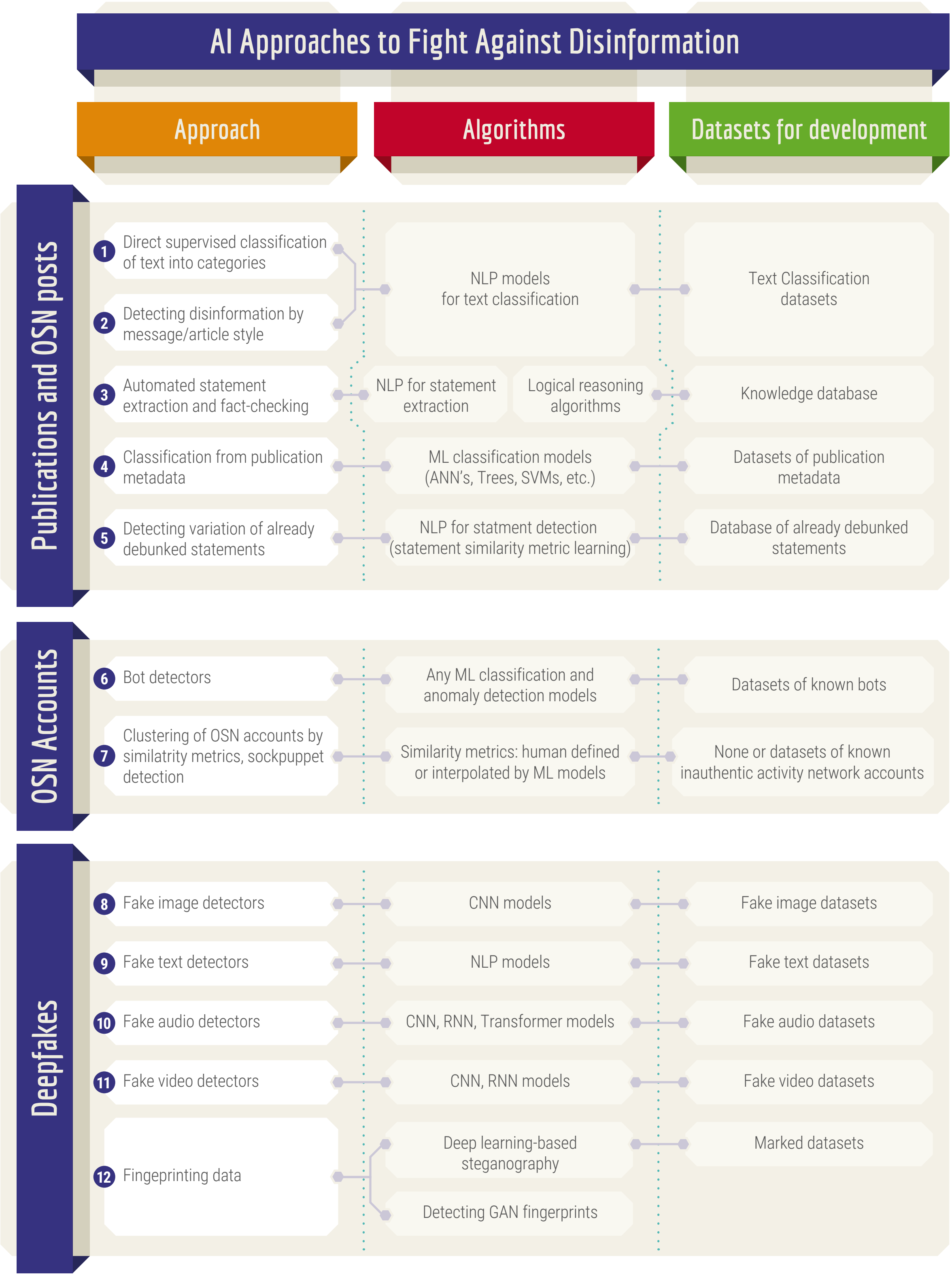
AI can automate a wide range of specific tasks and significantly increase analysts' productivity. However, due to current limitations, AI can play only a supporting role by helping to process vast amounts of information and detecting what requires further attention. However, social and online media monitoring tools are increasingly implementing AI-based solutions, which make it possible to search for information at a more granular level. Soon we also expect to see semantic search functionality that takes into consideration the intent and contextual meaning of the words used in a search. Here the GDELT database must again be mentioned; GDELT incorporates cutting-edge AI-based information extraction methods to allow users to search its massive online repository of news media. Google BigQuery is another AI-based platform that allows customers to benefit from entity extraction, key phrase and n-gram extraction, tone and sentiment estimation, or even use document-level embeddings in multi-lingual spaces to search for semantically similar articles.

We see that fundamental theoretical breakthroughs are still needed for AI to reach or exceed human capabilities in decision-making tasks. Most of the recent AI milestones have been achieved by big tech companies. However, the vast computing resources and access to data that make these companies important in the current landscape may not be as important for fundamental breakthroughs in the future. If so, the importance of other actors could increase in that context.

Bot detectors and NLP-based disinformation discovery tools tend to be specialised for specific countries, languages, and datasets. To detect disinformation campaigns in countries with less-common languages, one will need models and tools tailored to the languages and public discourse specifics of these countries. Even if big tech companies invest resources into developing more multilingual AI models, the pre-trained models they might release would still have limited performance capability. Further improvements and breakthroughs made in the NLP domain could change this linguistic AI capability bias as, for example, English language datasets could be efficiently reused in other language contexts, although they would still lack local cultural information.

Based on our research, the main areas in which AI can facilitate the work of analysts are the following:

1. ***Detecting suspicious content for further inspection***: deepfake



AI Approaches to Fight Against Disinformation

		Approach	Algorithms	Datasets for development
Publications and OSN posts	1	Direct supervised classification of text into categories	NLP models for text classification	Text Classification datasets
	2	Detecting disinformation by message/article style		
	3	Automated statement extraction and fact-checking	NLP for statement extraction Logical reasoning algorithms	Knowledge database
	4	Classification from publication metadata	ML classification models (ANN's, Trees, SVMs, etc.)	Datasets of publication metadata
	5	Detecting variation of already debunked statements	NLP for statment detection (statement similarity metric learning)	Database of already debunked statements
OSN Accounts	6	Bot detectors	Any ML classification and anomaly detection models	Datasets of known bots
	7	Clustering of OSN accounts by similatrity metrics, sockpuppet detection	Similarity metrics: human defined or interpolated by ML models	None or datasets of known inauthentic activity network accounts
Deepfakes	8	Fake image detectors	CNN models	Fake image datasets
	9	Fake text detectors	NLP models	Fake text datasets
	10	Fake audio detectors	CNN, RNN, Transformer models	Fake audio datasets
	11	Fake video detectors	CNN, RNN models	Fake video datasets
	12	Fingeprinting data	Deep learning-based steganography Detecting GAN fingerprints	Marked datasets

Figure 3. Potential AI approaches to fighting disinformation that have been identified in this study. This table is not exhaustive.

detection, scoring and sorting articles based on veracity, and detecting similar content to help evaluate the spreading dynamics of a particular narrative.

2. ***Detecting the spread of disinformation:*** monitoring topics, extracting named entities, building knowledge-graphs, and identifying bots, sockpuppets, and cyborgs based on their activity patterns, produced content, social network structure, coordination, etc.

3. ***Partial automation for efficient disinformation content analysis***: the classification of articles into disinformation narratives, automatic fact-checking (currently very limited), opinion mining, sentiment analysis, detection of emotional statements, topic modelling, and other features of the style typical of propaganda.

Figure 3 shows what components various disinformation-countering AI-based approaches require in terms of datasets and algorithms. Each approach is paired with the algorithm used (horizontally) and the dataset needed to develop an AI model. The vertical text to the left groups the approaches by area of application. There are many other small opportunities for the automation of disinformation analysis that we do not mention here because they depend on a highly specific analysis pipeline.

Conclusions

Current AI models require large, high-quality datasets. At present, deep learning AI models using neural networks are data-demanding, but they are becoming less data intensive due to transfer learning and other advancements. Even so, AI models require specific high-quality and unbiased datasets to learn complex tasks. In AI research, creating new databases for machine learning is expensive and difficult. For this reason, researchers use and reuse a small number of benchmark datasets, many of them originating from following institutions—Stanford, Microsoft, Princeton, Meta, Alphabet, Max Planck, AT&T.[124]

AI models fail to generalise if they are trained on datasets that are biased or too specific; this makes the usefulness of pre-trained models that lack fine-tuning somewhat limited in real-world situations. Due to the continuous arms race between attackers (disinformation actors) and defenders (analysts, fact checkers), it is also challenging to keep datasets and models up to date. Training data must evolve over time and follow the evolution of the disinformation landscape.

Probably the most critical limitation of artificial neural networks is that they still lack ontology (common terminology and vocabulary) and common-sense reasoning. These limitations hinder the ability of current AI models to do fact-checking, evaluate logical consistency, and deal with indirect statements (Aesopian language, metaphors, etc.). Despite scepticism about the usefulness of autonomous AI systems in disinformation detection and monitoring, current AI can do the heavy lifting for data collection, cleaning, categorisation, and translation. Models from domains such as Natural Language Processing and Computer Vision help build powerful tools for extracting information from large unstructured data collections, creating advanced knowledge graphs, tagging and classifying images/videos, etc. This allows analysts to efficiently navigate vast amounts of scraped multilingual information and search for patterns of interest.

Disinformation is a broad concept and hard to define (algorithmically). It is challenging and, in some cases, algorithmically unrealistic to define disinformation. Nevertheless, AI can help in monitoring hostile narratives defined by an analyst. At the current level of development, we predict that the relationship between AI and disinformation analysis will likely remain similar for the next 2–5 years.[125]

Detecting generated content by identifying artefacts may soon become impractical. In the context of deepfakes and AI-generated text, some disinformation content is of such limited size or infidelity, for example short posts or small resolution profile images, that it is impossible to distinguish if it was generated using AI. Moreover, improvements in generative models and the increasing realism of generated content, may make it become impossible for even a perfect AI agent to recognise the fakes. ***One solution would be to adopt data fingerprinting measures, such as deep learning-based steganography*** (see Fingerprinting data for authentication).

International cooperation in developing AI-based models, tools, and datasets is crucial, as is proper cooperation among governments and online platforms. Analysts can be taught the fundamentals and limitations of current AI methods, so that we can manage our expectations of AI tools and software, avoid biased decisions, and provide a deeper understanding of data necessary for training new AI models. Having a better understanding of AI capabilities and limitations could improve communication between analysts and AI engineers, allowing them to spot opportunities to facilitate the analysts' work, especially in data collection, cleaning, fusion, and other processing tasks. Governments, institutions, and military organisations must develop and integrate strategies regulating the responsible use of AI to avoid biased decisions during peacetime and in crisis situations. Such strategies should address AI integration from various perspectives, including responsibility, reliability, explainability, and legislation.[126]

Governments should enact legislation to fight disinformation. It is expected that the use of AI-powered tools in disinformation campaigns will gradually increase. Legislation should limit the potential harm of disinformation campaigns by establishing proper transparency and security standards for online platforms. Here the transparency of online platforms and a positive dialogue between platforms and analysts is essential to ensure that new legislation does not infringe on the valid interests of users and companies; online platforms must be given incentives to cooperate. Strengthening cooperation with online platforms is especially important for closed platforms that do not permit third-party access to their data via APIs. More transparent reports and deeper collaboration will enable the public, governments, and research institutions to assess the nature and scale of disinformation activities. Public challenges and hackathons are positive and interactive forums for improving the efficiency of digital platforms in countering disinformation activities, but their continued lack of transparency hinders these platforms from being more trusted despite their proposed cutting-edge AI solutions.

Proposed regulations can be successfully tested through voluntary mechanisms such as the EU's Code of Practice on Disinformation. Therefore, a reduction in the spread of disinformation can be achieved through a deeper integration of AI tools in the work of analysts and the establishment of a legal framework that properly addresses the concerns of all stakeholders.

Appendix

Here we include a list of tools and organizations that might be of interest to online media analysts and AI enthusiasts who are engaged in countering disinformation. We are also developing a review comparing different monitoring tools, which we expect to publish in one of the upcoming reports in 2022.

Tool	Applications
Spark Toro	Helps identify target audience sources of media consumption by entering keywords/ phrases into a search engine based on the behaviour of the TA.
BuzzSumo	Algorithms and analytics-based research tools are used to identify trends, buzzwords, trending social media stories, etc.
Meltwater	A media monitoring company that helps manage and monitor social media presence, engagement, influence, as well as measuring performance to identify trends.
Brandwatch	Market analysis tool with a large dataset used as a search engine to segmentise and analyse consumer trends.
Botometer	The free-to-use tool monitors a Twitter account's activity, followers, and friends and gives it a score to rank its likelihood of being a bot.

Table 1. Examples of tools that can be used for disinformation-countering analyses.

Organisation	Provided services
U.S. State Department Global Engagement Center	Emphasis on coordinating US Federal efforts in exposing disinformation campaigns in the public sector.

Organisation	Provided services
Debunk.eu	Research disinformation and media literacy education. Uses AI tools in a feedback loop with human experts—AI tools continually learn and try to automate parts of the analysis pipeline.
Graphika	Market research, strategic message planning, and disinformation detection.
Poynter	Media literacy, education, funding for other fact-checking organisations.
Sentinel	"Headquartered in Tallinn, Estonia, Sentinel works with governments, media and defence agencies to help protect democracies from disinformation campaigns, synthetic media and information operations by developing a state-of-the-art AI detection platform." https://thesentinel.ai/about.html
Snopes	Fact-checking organisation, media publications, quotes from key political players.
PolitiFact	Emphasis on non-partisan fact-checking and providing consultations to media organisations.
FullFact	Fact-checking claims made by various organisations with an emphasis on automation and ML.

Table 2. Selection of organisations that work with disinformation countering analytics

Bibliography

1 The authors would like to thank the following disinformation experts for sharing their useful insights about the field: Mr Lukas Andriukaitis, Mr Viktoras Daukšas, Ms Alyssa Kann, and Mr Kamil Mikulski. Of course, any misinterpretations and errors are entirely the authors' fault.

2 Akers, J., Bansal, G., Cadamuro, G., Chen, C., Chen, Q., Lin, L., Mulcaire, P., Nandakumar, R., Rockett, M., Simko, L., Toman, J., Wu, T., Zeng, E., Zorn, B., & Roesner, F. Technology-Enabled Disinformation: Summary, Lessons, and Recommendations. ArXiv, 2018.

3 Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election. 2017.

4 Tucker, J.A., Guess, A.M., Barberá, P., Vaccari, C., Siegel, A.A., Sanovich, S., Stukal, D., & Nyhan, B. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. 2018.

5 Broniatowski, D.A., Jamison, A., Qi, S., Alkulaib, L., Chen, T., Benton, A., Quinn, S., & Dredze, M. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. American Journal of Public Health, 2018, 108, 1378-1384.

6 Goodfellow, I., Bengio, Y., & Courville, A.C. Deep Learning. Nature, 2015, 521, 436-444.

7 Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts, Brussels, European Commission, 21 April 2021.

8 Haugeland, J. Artificial intelligence: The very idea. MIT press. 1989, 116.

9 Goodfellow, I., Bengio, Y., & Courville, A.C. Deep Learning. Nature, 2015, 521, 436-444.

10 Stöcker, C. How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation. In Multidisciplinary International Symposium on Disinformation in Open Online Media. Springer, Cham, February 2019, 129-149.

11 Stöcker, C. How Facebook and Google Accidentally Created a Perfect Ecosystem for Targeted Disinformation. In Multidisciplinary International Symposium on Disinformation in Open Online Media. Springer, Cham, February 2019, 129-149.

12 Runcieman, M. YouTube, the Great Radicalizer. The New York Times, 10 March 2018, sec. Opinion.

13 Deepmind. MuZero: Mastering Go, chess, shogi and Atari without rules. Deepmind blog, 23 December 2020.

14 Attacks in the Wild, OpenAI blog, [Accessed 20 December 2021]

15 Oshikawa, R., Qian, J., & Wang, W.Y. A Survey on Natural Language Processing for Fake News Detection. LREC, 2020.

16 Elhadad, M.K., Li, K.F., & Gebali, F. Fake News Detection on Social Media: A Systematic Survey. 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2019, 1-8.

17 Cer, Daniel, et al. "Universal sentence encoder." arXiv preprint arXiv:1803.11175 (2018).

18 Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

19 Khan, J.Y., Khondaker, M.T., Afroz, S., Uddin, G., & Iqbal, A. A benchmark study of machine learning models for online fake news detection, arXiv, 2021.

20 Oshikawa, R., Qian, J., & Wang, W.Y. A Survey on Natural Language Processing for Fake News Detection. LREC, 2020.

21 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I., "Language models are unsupervised multitask learners." OpenAI blog, 2019.

22 Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., & Wang, J. Release strategies and the social impacts of language models. Pre-print, arXiv, 2019.

23 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. Pre-print, arXiv, 2019.

24 Jwa, H., Oh, D., Park, K., Kang, J., & Lim, H. exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). Applied Sciences, 2019, 9, 4062.

25 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., Bert: Pre-training of deep bidirectional transformers for language understanding. Pre-print, arXiv, 2018.

26 Marcellino, W. Detecting Conspiracy Theories on Social Media Improving Machine Learning to Detect and Understand Online Conspiracy Theories. RAND CORP SANTA MONICA CA, 2021.

27 Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. TweepFake: About detecting deepfake tweets. PLoS ONE, 2021, 16.

28 Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. Automatic Detection of Fake News. ArXiv, 2017.

29 Gordon, R. Better fact-checking for fake news. MIT News.

30 Gordon, R. Better fact-checking for fake news. MIT News.

31 Ghosh, S., & Shah, C. Toward Automatic Fake News Classification. In Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019.

32 Shaar, S., Martino, G.D., Babulkov, N., & Nakov, P. That is a Known Lie: Detecting Previously Fact-Checked Claims. ACL.

33 FactSparrow. Repustar. [Accessed 25 August 2021]

34 Repustar. [Accessed 25 August 2021]

35 Analyzing Entity Sentiment, Google Cloud, [Accessed 20 December 2021]

36 Build a Knowledge Graph using NLP and Ontologies, Neo4j, [Accessed 20 December 2021]

37 AutoML solutions by Google, [Accessed 20 December 2021]

38 AutoML solutions by Microsoft, [Accessed 20 December 2021]

39 AutoML solutions by Amazon, [Accessed 20 December 2021]

40 Russell, S.J., & Norvig, P. Artificial intelligence - a modern approach. Prentice Hall series in Artificial Intelligence. 2003, 4th Edition, 314.

41 Groza, A. Detecting fake news for the new coronavirus by reasoning on the Covid-19 ontology. ArXiv, 2020.

42 Groza, A., & Pop, A. Fake news detector in the medical domain by reasoning with description logics. 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), 2020, 145-152.

43 Schroepfer, M. Creating a dataset and a challenge for deepfakes. Facebook AI blog.

44 Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images, IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1-11.

45 Hwang, T. Deepfakes - Primer and Forecast. Riga: NATO Strategic Communications Centre of Excellence, 2021.

46 Deconstructing Deepfakes-How do they work and what are the risks?. WatchBlog: Official Blog of the U.S. Government Accountability Office, 13 October 2020.

47 Guarnera, L., Giudice, O., & Battiato, S. DeepFake Detection by Analyzing Convolutional Traces. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, 2841-2850.

48 Guarnera, L., Giudice, O., & Battiato, S. DeepFake Detection by Analyzing Convolutional Traces. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, 2841-2850.

49 Neves, J., Tolosana, R., Vera-Rodríguez, R., Lopes, V., Proencca, H., & Fierrez, J. GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection. IEEE Journal of Selected Topics in Signal Processing, 2020, 14, 1038-1048.

50 Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. Countering Malicious DeepFakes: Survey, Battleground, and Horizon. ArXiv, 2021.

51 Knight, W. Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them. Wired, 12 June 2020.

52 Knight, W. Deepfakes Aren't Very Good. Nor Are the Tools to Detect Them. Wired, 12 June 2020.

53 Zhang, M., Wang, X., Fang, F., Li, H., & Yamagishi, J. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet. Pre-print, arXiv, 2019.

54 Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L. J. A review of deep learning based speech synthesis. Applied Sciences, 2019, 9(19), 4050.

55 David, D. Council Post: Analyzing The Rise Of Deepfake Voice Technology. Forbes. 10 May 2021.

56 Zhang, Y., Jiang, F., & Duan, Z. One-class learning towards generalized voice spoofing detection. arXiv, 2020.

57 Dessa. Detecting Audio Deep Fakes With AI. Medium. 17 April 2020.

58 Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., & Ptucha, R. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5), 1024-1037.

59 Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., & Ptucha, R. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(5), 1024-1037.

60 Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. ArXiv, 2019.

61 Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. ArXiv, 2020.

62 FakeVideoForensics, Github, [Accessed 20.12.2021.]

63 Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. ArXiv, 2019.

64 Yang, X., Li, Y., & Lyu, S. Exposing Deep Fakes Using Inconsistent Head Poses. ArXiv, 2018.

65 Mirsky, Y., & Lee, W. The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys, 2021, 54(1), 1–41.

66 Carlini, N., & Farid, H. Evading Deepfake-Image Detectors with White- and Black-Box Attacks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, 2804-2813.

67 Yu, N., Skripniuk, V., Abdelnabi, S., & Fritz, M. Artificial GAN Fingerprints: Rooting Deepfake Attribution in Training Data. arXiv, 2020.

68 Yu, N., Davis, L., & Fritz, M. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, 7555-7565.

69 Microsoft Innovation. Project Origin. [Accessed 15 July 2021].

70 Scrapy, [Accessed 20 December 2021]

71 Selenium, [Accessed 20 December 2021]

72 Guan, R., Zhang, H., Liang, Y., Giunchiglia, F., Huang, L., & Feng, X. Deep Feature-Based Text Clustering and Its Explanation. IEEE Transactions on Knowledge and Data Engineering, 2020.

73 Similarsites, [Accessed 20 December 2021]

74 Leetaru, K., Schrodt, P.A., "Gdelt: Global data on events, location, and tone, 1979–2012." ISA annual convention. Vol. 2. No. 4. Citeseer, 2013.

75 The GDELT event database data format codebook, 2015

76 Gerner, D.J., Abu-Jabr, R., Schrodt, P.A., & Yilmaz, Ö. Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions. 2002.

77 Serverless, highly scalable, and cost-effective multicloud data warehouse designed for business agility. Google BigQuery, [Accessed 20 December 2021]

78 GDELT Open data registry on Amazon [Accessed 20 December 2021]

79 Pogorelov, K., Schroeder, D. T., Filkukova, P., & Langguth, J. A System for High Performance Mining on GDELT Data. In 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), May 2020, 1101-1111.

80 Pogorelov, K., Schroeder, D. T., Filkukova, P., & Langguth, J. A System for High Performance Mining on GDELT Data. In 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), May 2020, 1101-1111.

81 Kumar, S., Cheng, J., Leskovec, J., & Subrahmanian, V. S. An army of me: Sockpuppets in online discussion communities. In Proceedings of the 26th International Conference on World Wide Web, April 2018, 857-866.

82 Tsikerdekis, M., & Zeadally, S. Multiple Account Identity Deception Detection in Social Media Using Nonverbal Behavior. IEEE Transactions on Information Forensics and Security, 9, 2014 1311-1321.

83 Kešelj, V., Peng, F., Cercone, N., & Thomas, C. N-Gram-Based Author Profiles for Authorship Attribution. In Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING, August 2003, Vol. 3, 255-264.

84 Wang, J., Zhou, W., Li, J., Yan, Z., Han, J., & Hu, S. An online sockpuppet detection method based on subgraph similarity matching. In 2018 IEEE Intl. Conf. on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, December 2018, 391-398.

85 Pisciotta, G., Somenzi, M., Barisani, E., & Rossetti, G. Sockpuppet Detection: a Telegram case study. Pre-print. arXiv 2021.

86 Yamak, Z., Saunier, J., & Vercouter, L. SocksCatch: Automatic detection and grouping of sockpuppets in social media. Knowledge-Based Systems, 2018, 149, 124-142.

87 Wang, J., Zhou, W., Li, J., Yan, Z., Han, J., & Hu, S. An online sockpuppet detection method based on subgraph similarity matching. In 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, December 2018, 391-398.

88 XRuan, X., Wu, Z., Wang, H., & Jajodia, S. Profiling online social behaviors for compromised account detection. IEEE transactions on information forensics and security, 2005, 11(1), 176-187.

89 XRuan, X., Wu, Z., Wang, H., & Jajodia, S. Profiling online social behaviors for compromised account detection. IEEE transactions on information forensics and security, 2005, 11(1), 176-187.

90 Wang, J., Zhou, W., Li, J., Yan, Z., Han, J., & Hu, S. An online sockpuppet detection method based on subgraph similarity matching. In 2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, December 2018, 391-398.

91 Tsikerdekis, M., & Zeadally, S. Multiple Account Identity Deception Detection in Social Media Using Nonverbal Behavior. IEEE Transactions on Information Forensics and Security, 2014, 9, 1311-1321.

92 Authors' interviews with Ms Alyssa Kann, Research Assistant, Digital Forensic Research Lab, The Atlantic Council, 23 June 2021 and Mr Kamil Mikulski, Analyst, Kosciuszko Institute, 15 June 2021.

93 Cresci, S. A decade of social bot detection. Communications of the ACM, 2020, 63(10), 72–83.

94 Orabi, M.S., Mouheb, D., Aghbari, Z., & Kamel, I. Detection of Bots in Social Media: A Systematic Review. Information Processing and Management, 2020, 57.

95 Orabi, M.S., Mouheb, D., Aghbari, Z., & Kamel, I. Detection of Bots in Social Media: A Systematic Review. Information Processing and Management, 2020, 57.

96 Cresci, S. A decade of social bot detection. Communications of the ACM, 2020, 63(10), 72–83.

97 Orabi, M.S., Mouheb, D., Aghbari, Z., & Kamel, I. Detection of Bots in Social Media: A Systematic Review. Information Processing and Management, 2020, 57.

98 Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A. Online human-bot interactions: Detection, estimation, and characterization. In Proceedings of the International AAAI

Conference on Web and Social Media, May 2017, Vol. 11, No. 1.

99 Rauchfleisch, A., & Kaiser, J. The False Positive Problem of Automatic Bot Detection in Social Science Research. PLOS ONE, 2020, 15(10).

100 Cresci, S. A decade of social bot detection. Communications of the ACM, 2020, 63(10), 72–83

101 Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. IEEE Transactions on Dependable and Secure Computing, 2017, 15(4), 561-576.

102 Vargas, L., Emami, P., & Traynor, P. On the detection of disinformation campaign activity with network analysis. In Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, November 2020, 133-146.

103 Vargas, L., Emami, P., & Traynor, P. On the detection of disinformation campaign activity with network analysis. In Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop, November 2020, 133-146.

104 Cresci, S. A decade of social bot detection. Communications of the ACM, 2020, 63(10), 72–83

105 Authors' online interview with Mr Lukas Andriukaitis, Associate Director, Digital Forensic Research Lab, Atlantic Council, 10 June 2021.

106 Cresci, S. A decade of social bot detection. Communications of the ACM, 2020, 63(10), 72–83

107 Cresci, S. A decade of social bot detection. Communications of the ACM, 2020, 63(10), 72–83.

108 Rauchfleisch, A., & Kaiser, J. The False Positive Problem of Automatic Bot Detection in Social Science Research. PLOS ONE, 2020, 15(10).

109 Roth, Y., & Pickles, N. Bot or not? The facts about platform manipulation on Twitter. Twitter blog. 18 May 2020.

110 Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In Proceedings of the 26th International Conference on World Wide Web Companion, April 2017, 963-972.

111 Rauchfleisch, A., & Kaiser, J. The False Positive Problem of Automatic Bot Detection in Social Science Research. PLOS ONE, 2020, 15(10).

112 Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. Layer-Wise Relevance Propagation: An Overview. Explainable AI. 2019.

113 Gunning, D., & Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 2019, 40, 44-58.

114 Rokach, L., & Maimon, O. Decision trees. In Data mining and knowledge discovery handbook. Springer, Boston, MA, 2005, 165-192.

115 Goodfellow, I. J., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples. Pre-print. arXiv, 2015.

116 Dombrowski, A. K., Alber, M., Anders, C. J., Ackermann, M., Müller, K. R., & Kessel, P. Explanations can be manipulated and geometry is to blame. Pre-print, arXiv, 2019.

117 Heo, J., Joo, S., & Moon, T. Fooling Neural Network Interpretations via Adversarial Model Manipulation. Advances in Neural Information Processing Systems, 2019, 32, 2925-2936.

118 Rieger, L., & Hansen, L. K. A simple defense against adversarial attacks on heatmap explanations. Pre-print, arXiv, 2020.

119 Wang, Z., Wang, H., Ramkumar, S., Mardziel, P., Fredrikson, M., & Datta, A. Smoothed Geometry for Robust Attribution. In Advances in Neural Information Processing Systems, 2020, 13623–13634.

120 Facebook AI. How we're using Fairness Flow to help build AI that works better for everyone. Facebook AI blog, 31 March 2021.

121 Proposal for a Regulation of The European Parliament and of The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Commission. [Accessed 15 July 2021].

122 Akers, J., Bansal, G., Cadamuro, G., Chen, C., Chen, Q., Lin, L., Mulcaire, P., Nandakumar, R., Rockett, M., Simko, L., Toman, J., Wu, T., Zeng, E., Zorn, B., & Roesner, F. Technology-Enabled Disinformation: Summary, Lessons, and Recommendations. ArXiv, 2018.

123 Facebook AI. How we're using Fairness Flow to help build AI that works better for everyone. Facebook AI blog, 31 March 2021.

124 Koch, B., Denton, E., Hanna, A., & Foster, J. G., Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. Pre-print arXiv, 2021.

125 Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 2018, 62, 729-754.

126 NATO, Summary of the NATO Artificial Intelligence Strategy, 22 Oct.

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org