Corrosive AI: Emerging Effects of the Use of Generative AI on Political Trust

~by~

Riley Lankes

SN 22206048

University College Dublin

School of Politics and International Relations

17th August 2023

~

This thesis is submitted in partial fulfillment of the requirements for the degree of

Master of Arts in International Relations

Abstract

The rapid rate of development of generative artificial intelligence (generative AI) technologies in recent years has led many scholars and industry experts to express concern about potential negative externalities stemming from the technology. This framing of the issue assumes that AI provides a net benefit to society, but that unforeseen side effects may pose varied risks to existing economic, political, and social structures. This paper offers a different view of risks associated with generative AI; like any technology, the effects of generative AI on existing structures depend on how it is used. The question is not *how AI will affect existing structures,* but *how the use of AI will affect them.* As such, this paper asks: in what ways could the use of generative AI corrode political trust? To answer this question, this paper identifies four salient factors related to the use of generative AI that have the potential to damage political trust. In turn, these four factors are translated into sub-questions which inform the questions posed in the expert interviews conducted for this project. Insights gained from these interviews lend support to the idea that generative AI has the potential to corrode political trust. Interviews also identified gaps in the current understanding of how generative AI is used, and aid in setting an agenda for related future research.

Word Count: 12,488

Keywords: Disinformation, generative AI, political trust, deepfakes.

Table of Contents

Introduction

Artificial intelligence is not a new idea. Dating back to at least the Ancient Greek myth of Talos, a giant bronze construct which was said to guard the island of Crete, humans have imagined the possibility of created intelligences. In the 21st century, these imaginings have become reality. In the last 5 years, the rate of development of artificial intelligence technologies has grown exponentially. This is largely thanks to the rise of Generative Large Language Multi-modal Model AI (GLLMM), a technology which allows researchers to use language-based models to approach nearly any problem. Rather than various fields of research developing different AI technologies, the rise of GLLMMs has led nearly all AI research to focus on language-based models, rapidly accelerating the rate of development.[1]

While rapid development of a new technology is not necessarily dangerous, what concerns many scholars and technologists is the rapid rate of *deployment* of AI. Before a new medication can be rolled out to patients, it goes through years of testing so that efficacy and safety can be ensured. The rapid development of AI has resulted in a race between the companies and states developing it, with a rush to create and deploy the technology to avoid falling behind.

As with the development of any new technology which sees widespread adoption, the effects which AI will have on the structures of society are unclear. Experts from across numerous fields have expressed concern that the rapid deployment of AI without small-scale testing and precursory regulation may lead to unintended consequences across the existing structures of global society. Some technologists have even gone as far as to predict that AI will "…break the

[1] Center for Humane Technology (2023)

operating system of humanity" due to how fundamentally it may change how people interact and self-govern.[2]

While there are many ways in which AI may affect the realm of international relations, this paper is focused on the relationship between AI and trust. As such, one type of AI is of particular concern: *generative AI.* Generative artificial intelligence is a subtype of AI which uses deep learning models to generate original content. Generative AI is "trained" on a collection of data, which it then uses to "…generate statistically probable outputs when prompted."[3] With recent advances in this technology, these generated outputs can take many forms: text, images, audio, and even video. One application of generative AI is particularly important to this paper: AI-powered deepfakes. Deepfakes are a technology which can create imitations of the likeness of an individual, with the ability to create video facsimiles of facial features and voices. In recent years, deepfakes have begun to incorporate generative AI models. With this incorporation of generative models, AI-powered deepfakes make it possible to generate videos of individuals doing things they never did, or saying things they didn't say.

Exponential growth in the rate of development of generative AI has led many to speculate that we may be headed towards a crisis of trust.[4] Considered alongside the explosion of digital disinformation in the age of social media, as well as a trend of declining trust in government since the mid-20th century[5], predictions that AI will erode trust appear plausible. The novel contribution put forth in this paper is that AI may be *corrosive* to trust. This paper coins the term *corrosive AI* – the idea that rather than eroding trust slowly over time, the injection of generative

[2] Center for Humane Technology (2023)
[3] IBM (2023)
[4] Center for Humane Technology (2023); Dobber et al. (2021)
[5] Hetherington & Husser (2012); Keele (2007)

AI into an information paradigm that is already struggling to combat disinformation is likely to corrode trust far more quickly than ever before.

Broadly speaking, this paper offers an examination of the relationship between the use of generative AI and public trust in government. More specifically, this paper is concerned with political trust. Political trust is best understood as an evaluation of government action against people's expectations of their government. Trust is formed over time through iterated interaction. If government entities fail to meet expectations, trust is lost. Critically, research shows that political scandals to have a significant negative impact on political trust.[6]

Widespread access to generative AI has the potential to allow for the generation of political scandals. With software capable of generating deepfakes widely available, the ability to create a sandal by falsifying a video or audio recording of a given political actor is now in the hands of millions of people. These AI generated videos or recordings do not need to be true to induce a scandal, they only need to be believed. AI may also erode trust in the long-term, as citizens could lose confidence in the authenticity of video and audio content featuring members of government. This loss of trust could have broad policy implications, as political trust has been shown to be an important determinate of foreign and domestic policy.

This paper begins with a review of relevant studies from both the trust and disinformation literatures. In the next section, an overview of research and recent developments in AI technologies is presented. The unprecedented rate of development of AI necessitates a distinct section; an understanding of generative AI and its technical capabilities is necessary for understanding the central arguments of this paper. Building on the foundation provided by these various bodies of research, the third section presents the core claim of this paper: widespread

[6] Keele (2007)

access to generative AI has the potential quickly to corrode political trust. The fourth section reviews insights gained from interviews with subject-matter experts in areas including AI development, political trust, and disinformation. The paper concludes with a discussion of findings and potential areas for future research.

Literature Review – Trust & Disinformation

Trust in Politics, Trust in Content

The idea of corrosive AI is grounded in the assumption that generative AI will quickly corrode the public's trust in government, as well as trust in the authenticity of content. Examining AI's potential relationship with trust requires first defining trust. Trust is the basis of cooperative interactions between actors but proves difficult to precisely define. Game Theory provides a starting point for building a suitably precise definition. In the game theoretical sense, trust is the level of subjective probability with which actor/group A assesses that actor/group Z will perform a particular action, before actor A can monitor said action.[7] In this sense, the statement "I trust Z" made by A can be understood as a belief that Z will likely perform an action that benefits A, to the point that cooperation appears beneficial. Trust is a belief that expected actions will be undertaken by another actor or group. When the subjective probability of expected actions seems high, we say we "trust" someone. When that probability seems low, we say someone is "untrustworthy", and to tend to refrain from cooperation.[8]

[7] Gambetta (2000)
[8] Gambetta (2000)

Applying this conceptualization of trust to the realm of politics, trust in government actors and institutions is the result of people's perceptions of government behavior relative to their expectations.[9] According this definition, political trust can be understood to increase when government actions match people's expectations. However, this definition does not consider the temporal dimension of trust formation. According to *Social Learning Theory*, trust is formed via iterated interactions.[10] Beyond conceptual understanding, an empirical study of group trust formation using data from the Social Trust Survey found empirical support for this model of trust formation.[11] Synthesizing this with the game theoretical definition detailed above, political trust is extrapolated from multiple instances of expectation-vs-action analysis over time. This fits with the working definitions of political trust used in many empirical studies, which treat changes in trust as a function of perceived performance on salient political issues.[12]

The loss of trust is currently an area of focus in political trust literature. While many scholars agree that political trust has declined in the United States since the 1960s, identifying root causes of this loss of trust is the subject of ongoing study.[13] One phenomenon has gained particular attention: scandal. Luke Keele's work in this area is particularly notable, focusing on the effect which scandals have on political trust, while also considering perceptions of government performance over time. Through analysis of macro-level data on trust in government in the United States, Keele demonstrates that high-profile scandals involving political actors have a significant negative impact political trust.[14]

[9] Hetherington & Husser (2012)
[10] Glanville & Paxton (2007); Hardin (2002)
[11] Glanville & Paxton (2007)
[12] Hetherington & Husser (2012); Hetherington (1998)
[13] Hetherington & Husser (2012); Keele (2007)
[14] Keele (2007)

With political trust understood to be formed via repeated interactions and expectations-vs-reality checks, it's important to understand the context and nature of these interactions. With the prolific nature of social media and television today, a significant portion of political trust-forming interactions involve video content of some sort.[15] This is where it becomes important to understand the second type of trust relevant to this paper: trust in the authenticity of content.

The assumption that video content can contribute to the formation of political trust is well grounded in existing literature. Past scholarship suggests that individuals tend to form nonreciprocal connections with media figures or political opinion leaders whom they have never met, but whose content they consume frequently.[16] This type of one-sided connection is referred to as a *parasocial relationship.* There is also strong evidence to suggest that the formation of parasocial relationships can impact political trust.[17] This assumption is further made by several studies on disinformation in politics, is rarely explicitly defined within these.[18] Instead, many studies tend to lump video-based and text-based news coverage together in a term like "news coverage". This lumping together of two very different types of content is problematic, due to the difference in perceived trustworthiness between the two mediums. Recent experimental evidence from studies on text-based disinformation suggests that the medium may not be as effective as previously assumed.[19] On the other hand, videos tend to be regarded by many as the gold standard in authentically portraying events, were "seeing is believing".[20]

Despite this common belief that "seeing is believing", there is evidence to suggest that videos are not always authentic representations of the world. The context within which a photo or video

[15] Verma (2023)
[16] Hoffner & Bond (2022)
[17] Liu (2023)
[18] Freelon & Wells (2020); Erlich & Garner (2023)
[19] Elrich & Garner (2023)
[20] Dobber et al. (2021); Verma (2023)

is situated is critical for determining the meaning which individuals exposed to the content ascribe to it. The meaning communicated by a photo or video can be altered by changing the context of the content, without changing the content itself.[21] With the advent of deepfakes, bad actors now have the capability to alter *both* the context and content of videos. As the next section will explore, altering the context of photos and videos is one of many methods for creating disinformation. The study undertaken by this paper is motivated by the idea that widespread access to deepfakes, which can modify the *content* of video, may increase both the effectiveness and volume of disinformation published digitally.

Disinformation

Misinformation and disinformation can both be thought of as violations of content trust, as they are by nature *not* authentic representations of the world. Whereas misinformation is generally defined as inaccurate or false content that misleads, disinformation is content which is *intentionally designed* to mislead. Misinformation is incorrect by accident; disinformation is incorrect by intent.[22] This paper is concerned with the effects which widespread access to AI generated content may have on political trust. The creation of this content is by nature intentional, and so is *disinformation.*

Disinformation as a phenomenon has been extensively studied. Current scholarship across political science and informatics suggests a rise in the publication of disinformation and misinformation in news and media.[23] One study which focused on UN Peacekeeping found evidence to suggest that peacekeeping operations have faced a growing number of targeted

[21] Verma (2023)
[22] Lanoszka (2019)
[23] Bennett & Livingston (2020); Martens et al. (2018)

disinformation efforts in recent years.[24] Further studies in the field have focused on the effectiveness of policy used to combat disinformation in democratic states[25], or attempted to study the comparative effectiveness of disinformation across various political and economic topics.[26]

With disinformation research in international relations having tended towards studying disinformation as a phenomenon, the area of study has only recently begun to receive theoretical attention. The result of this trend is a well fleshed-out understanding of the facts of disinformation campaigns, but a gap in understanding surrounding the knock-on effects of disinformation. In turn, this gap limits the ability to accurately assess the potential effects of disinformation empowered by generative AI. One recent attempt at situating the phenomenon of disinformation into theory posits that the ultimate aim of disinformation is to "…affect the target state's ability to generate military capabilities or willingness to align itself with others against the disinforming state…."[27] According to this decidedly rationalist understanding of disinformation, the disinforming state's goal is to affect the target state's ability to support its own foreign policy and military strategy. Disinformation only seeks to influence the national discourse or diminish the target state's legitimacy to serve the disinforming state's strategic goals.[28]

The rationalist understanding of disinformation has merit but is flawed. The critical issue is in the state-centric approach of this understanding, where only states are considered as the architects of disinformation campaigns. While most research on disinformation campaigns has focused on state actors as disinformers, specifically Russia[29], the assumption that only states can

[24] Trithart (2022)
[25] Tenove (2020)
[26] Erlich & Garner (2023)
[27] Lanoszka (2019)
[28] Lanoszka (2019)
[29] Erlich & Garner (2023); Tenove (2020); Freelon & Wells (2020)

coordinate disinformation campaigns is not supported by the literature. There is evidence to suggest that nonstate actors such as corporate entities and political groups actively engage in spreading disinformation in the pursuit of their interests.[30]

With the scope of which actors can engage in spreading disinformation expanded to include nonstate actors, the rationalist understanding of disinformation falls more in line with the evidence put forward by contemporary research. This paper operates under the assumption that the purpose of publishing disinformation is to affect a target group's ability to support its own strategic goals. Defining the purpose of disinformation broadly here allows for the consideration of individual actors, groups, or even entire state governments as the targets of disinformation.

Current State of AI

The rate of development of artificial intelligence technology has rapidly accelerated in the last decade. At present, some experts suggest that AI development has reached an exponential rate of growth.[31] The rapid rate of change in what AI is capable of is what necessitates a section of this paper devoted to discussing its current capabilities. Furthermore, the rapid rate of development means that any sufficiently peer-reviewed studies involving AI are necessarily out of date by the time they are published. This section attempts to contextualize past scholarship on AI with the current capabilities of AI technologies. This is done with the understanding that this paper is not immune to the mechanisms described above and will need to be updated to remain relevant.

[30] Franta (2021); Hameleers (2020)
[31] Center for Humane Technology (2023)

Foundational Knowledge and Current Capabilities

While there are many types of AI, this study focuses on generative AI. Generative artificial intelligence is a subtype of AI which uses deep learning models to generate original content. Generative AI is "trained" on a collection of data, which it then uses to "…generate statistically probable outputs when prompted."[32] With recent advances in this technology, these generated outputs can take many forms: text, images, audio, and even video. One application of generative AI is particularly important to this paper: AI-powered deepfakes. AI-powered deepfakes are highly realistic videos and images generated by AI using deep learning technology, with the ability to create video facsimiles of facial features and voices. In short, deepfakes make it possible to generate videos of individuals doing things they never did or saying things they never said.[33]

The rate of development of AI has skyrocketed in the last decade. AI experts often ascribe this rapid growth to the development of Generative Large Language Multi-modal Models (GLLMM), which allow for the use of language-based models to approach almost any type of issue. Widespread adoption of GLLMMs has streamlined AI research, so research across various fields all contributes to the advancement of language-based models, rapidly accelerating the rate of development.[34] This acceleration in the rate of development of AI has led to countless advances in what AI technologies are capable of. Two of these advances are particularly relevant to the claims made by this paper: emergent abilities, and the capacity for self-improvement.

[32] IBM (2023)
[33] Chesney & Citron (2019)
[34] Center for Humane Technology (2023)

Emergent abilities are a phenomenon wherein large language model AI develops new capabilities when scaled up which it did not initially possess.[35] For an example of this phenomenon, one only needs to turn to the increasingly popular ChatGPT, powered by the GPT-3 model. While GPT-3 did not initially have the capability, researchers discovered in early 2023 that ChatGPT had developed the ability to correctly answer questions about research-grade chemistry.[36] The use of the term "discovered" in the previous sentence is appropriate, given that there is currently no method for predicting or detecting emergent abilities in AI.[37] As Jeff Dean, Google's Chief Scientist stated, "Although there are dozens of examples of emergent abilities [in AI], there are currently few compelling explanations for why such abilities emerge." In short, AI language models appear to be developing new capabilities on their own, and researchers do not fully understand why this is happening.

The second advancement in AI technology worth noting in this paper is the ability to self-improve. In 2022, researchers from the University of Illinois and Google were able to prompt Large Language Model AI (LLM) to generate datasets, which the model was then able to train itself on. The researchers found that the model was capable of improving its performance on other reasoning datasets by training on the data which it generated.[38] In short, AI is capable of making itself better at accomplishing certain tasks.

It is important to note that much of the current research on AI in the social sciences tends to ascribe agency to AI, with predictions about how AI may or may not affect various aspects of international political, economic, or social structures. These studies, whether intentionally or not, are often rooted in *technological determinism*, or the idea that technological development is to

[35] Wei et al. (2023)
[36] Jablonka et al. (2023)
[37] Wei et al. (2023)
[38] Huang et al. (2022)

some degree autonomous. A deterministic approach would hold that the development of technology shapes human society, not the other way around. While popular depictions of AI in fiction often have agency (think HAL 9000), the reality in our world is that AI has not reached that point. Generative AI is a tool that requires input, and the effects which it has on political, economic, and social structures are largely determined by *how people choose to use it.*[39]

AI-Powered Disinformation

Research on the relationship between AI and disinformation has focused on AI as a potential tool for spreading disinformation. New research suggests that technology's ability to analyze vast amounts of metadata and serve users content specifically designed to appeal to them may allow disinformation efforts to become both more accurate and more destabilizing.[40] However, this research focuses on the use of AI in content serving algorithms, not on generative AI.

Recent developments in both natural language and image processing have led to the incorporation of generative AI into deepfake technology, which is capable of "learning" what an individual looks and sounds like using real images and recordings of the individual. Once an individual has been "learned", the AI-powered program can be used to create images, videos, or audio recordings of them which appear genuine.[41] Generative AI technology is now more accessible than ever, with nearly any individual with internet access able to use it. Companies have also begun to incorporate these functions into existing software, as seen in Microsoft's incorporation of an "AI-powered assistant" into the Windows search bar. This has led to concerns among analysts and researchers that widespread access to generative AI technology

[39] See Aspray & Doty (2023)
[40] Horowitz et al. (2018)
[41] Schippers (2020)

may result in an erosion of truth.[42] At the time of writing, there have already been instances of deepfakes being used to attack the reputation of political figures, with a deepfake video of U.S. Congresswoman Nancy Pelosi (D-CA) drunkenly rambling gaining widespread attention in 2019.[43] Further examples of this phenomenon have been observed more recently, with AI generated images of Donald Trump being arrested, as well as generated video clips of the former President hugging Dr. Anthony Fauci going viral in 2023.[44]

Instances of AI generated disinformation have led to efforts to examine trust in video as a medium, as well as trust in visual journalists. A study on AI-powered deepfakes found that deepfake technology "…threatens the individual credibility of both visual journalists and visual media because of the uncertainty, polarization, and misinformation…." This same study also predicts that an increase in AI generated content about political actors "…could further diminish trust in politicians and political processes."[45]

Model of Corrosive AI

Generative AI Corrodes Trust

Researchers and technologists alike have predicted that the advent of AI, specifically generative AI, may erode political trust.[46] This claim, while plausible, is imprecise for two

[42] Kertysova (2018)
[43] Schippers (2020)
[44] Schwartz (2023)
[45] Verma (2023)
[46] Verma (2023); Center for Humane Technology (2023)

reasons. First, the claim that "*AI will erode trust*" ascribes agency to AI, a capability which AI does not yet possess. Ascribing agency or autonomy to any technology is a hallmark of technological determinism. Technological determinism is best understood as an approach which emphasizes the autonomous and social-shaping tendencies of technology, while de-emphasizing the role of human input.[47] This paper rejects technological determinism, instead understanding technology as both the product of human design, as well as a tool which requires human input. While a technological determinist might state that "*AI will erode trust",* the argument put forth in this paper is more in line with the statement "the *misuse of AI will erode trust".* Here human input in decisions as to how AI should be regulated are determinates of how AI will affect structures of society like trust.[48]

The second issue with the statement "*AI will erode trust"* is in the choice of the verb "erode". Erosion denotes a slow decline over a long period of time, in the way that a river slowly erodes rock to form a canyon over thousands of years. The novel claim at the center of this paper is that AI will likely *corrode* political trust in only a matter of years, rather than the long-term breakdown denoted by "erosion". With these two clarifications made, the research question at the heart of this paper can be defined:

RQ1: In what ways could the use of generative AI corrode political trust?

Answering this question requires studying both the actual capabilities of AI, as well as the contexts within which the technology is used. Emergent patterns in the literature help define four

[47] Dafoe (2015)
[48] See McKnight et. Al (2011)

potential factors that may contribute to the corrosive nature of generative AI. These four salient factors are: 1) the open access model of generative AI deployment and development; 2) the relationship between generative AI and trust in video, the news media, and politics; 3) the potential use of generative AI to manufacture political scandals; and 4) the potential use of generative AI to enhance disinformation. These factors are expanded into research sub-questions, which inform the questions posed in interviews conducted by this study.

The foundations for answering these four sub-questions can be found in existing literature. However, none of these questions can be sufficiently answered by existing research alone. Below, each sub-question is defined, existing evidence from literature and news media is explored, and gaps in understanding are identified. The following section then details insights gained from elite interviews which attempt to fill these gaps in understanding.

SQ1: How could open access to generative AI make the technology corrosive to political trust?

Generative AI is accessible for anyone with an internet connection. With only a quick Google search, any individual has access to hundreds of websites hosting AI image generators, chatbots, and more. These websites tend to provide their services for free, with users only needing to create an account to access generative AI. Beyond these websites, which allow users to utilize AI software hosted by a 3rd party (such as ChatGPT from OpenAI), generative AI software is also available for download by users. Many examples of AI software available for download are also open source, meaning that users can modify the software at will. Meta's Llama-2 is the latest large example of an open-source large language model AI, being made available for free in July of 2023.[49]

[49] Meta (2023); Isaac & Metz (2023)

The obvious risk with open access to generative AI, especially open-source access, is that the original creators of the software cannot control how their product is used. The potential for misuse is high when users are given access to a technology, when the potential use cases of that technology are not fully understood. The use of deepfake technology in creating porn is already an issue, arising from the fact that open source deepfake software is widely available for download.[50] This issue has likely been made worse by the incorporation of generative models into existing deepfake programs. Furthermore, the publication of software online cannot be retracted. Once a piece of software is made available to the public, it is potentially out in the world for good.

Beyond this open access to generative AI, the way in which deepfake models function contributes to their potential to damage political trust. Deepfakes require source material to train on, wherein the generative AI model used to generate the deepfake will utilize numerous images of the subject to "learn" what their face looks like.[51] More data to train on improves the quality of the deepfake. With political figures necessarily being public figures, anyone attempting to create a deepfake of a political figure will enjoy a wealth of images and videos to train their model on.

Of the four sub-questions, the question of open access to generative AI is the most thoroughly understood. The prominent issue of deepfake porn provides evidence that open-sources generative AI software can be modified and used in ways outside of its intended purpose. What remains unclear is how prevalent AI generated content of political actors is currently, or how prevalent it will become.

[50] Gosse & Burkell (2020)
[51] Somers (2023)

SQ2: How could the use of generative AI affect existing trends of declining trust in government, media, and video as a medium?

It is well established by past scholarship that political trust generally has been declining in the United States since the 1960s.[52] A previous section of this paper established that political trust is formed via repeated interaction and expectation-vs-reality testing, and that many trust forming interactions in the 21st century involve video content. Video is now central to modern society's information ecology, and new research suggests that deepfake technology threatens the credibility of video as a medium. The same study also concluded that deepfakes present a threat to the credibility of visual journalists, due to the uncertainty that deepfakes were shown to create.[53]

Existing literature supports the idea that generative AI will likely exacerbate existing trends of declining trust in government. This is because generative AI has been shown to have the capability to damage trust in the content and mediums that play a part in building trust. If people distrust the authenticity of video content, trust formation in political actors based upon that content will be impeded. Similar reasoning can be applied to declining trust in visual journalists, as their work can be assumed to play a similar role in trust formation as video content does (in many cases, they are the creators of this content). Indeed, a handful of existing studies have suggested that deepfake technology has the capability to damage political trust.[54] One such study went as far as to predict that "An increase in the number of fake videos about political personalities—at all levels of national and regional politics—could further diminish trust in politicians and political processes."[55]

[52] Hetherington & Husser (2012); Keele (2007)
[53] Verma (2023)
[54] Dobber et al (2021)
[55] Verma (2023)

What remains unclear in this area is evidence for the connection between trust in video and visual journalists, and political trust. To be able to say that generative AI has the capacity to exacerbate the existing downward trend in political trust, further evidence showing that trust in video and trust in visual journalists affect political trust is needed.

SQ3: In what ways could generative AI be used to create scandals which damage political trust?

Research on trust suggests that scandals tend to be damaging to political trust.[56] Scandals are highly visible failures for political actors, with their negative effect on political trust likely tied to their high visibility. It stands to reason that if generative AI can create fake content which causes a scandal, the subsequent scandal would damage political trust in the political group or figure at the center of it.

At the time of writing, there is some evidence to suggest that AI generated content may be capable of causing a political scandal, with a modified video of U.S. Congresswoman Nancy Pelosi appearing to drunkenly slur words gaining widespread attention in 2019.[57] More recently, AI generated images of Donald Trump being arrested went viral, though it is unclear whether either of these examples constituted a true scandal.[58] In the realm of international relations specifically, the gap in understanding in this area surrounds whether AI generated content has become sophisticated enough to mislead people on such a wide scale that a scandal occurs. In short, has generative AI become good enough to fool people on a wide scale? If so, has this already occurred, or is it likely to occur in the near future?

[56] Keele (2007)
[57] Schippers (2020)
[58] Schwartz (2023)

SQ4: How could disinformation campaigns empowered by generative AI damage political trust?

Generative AI likely has the capability to empower disinformation campaigns. Generative AI can be used to create images, videos, and text which existing research suggests are capable of misleading people.[59] Indeed, the use of AI to enhance disinformation has been documented by previous research. AI-powered content targeting allows for disinformation to become more effective by tailoring messages to specific individuals, based upon user profiles. The use of AI for this type of "micro-targeting" is well documented, with the technology proving capable of using the demographic information and online habits of users to "…deliver highly personalized content, and thereby target with maximum effectiveness those who are most vulnerable to influence."[60]

As deepfake technology and AI-powered content targeting continue to become more accessible and sophisticated, their use in disinformation campaigns may promote distrust in traditional media.[61] If people begin to doubt the authenticity of any content, there is a risk that political polarization will intensify as individuals become more and more skeptical of information that does not confirm their existing beliefs. This phenomenon is known as an echo chamber, wherein people only consume content from sources which reaffirm their existing worldview. Targeted disinformation campaigns could leverage AI technologies to further the creation of these echo chambers, exacerbating political polarization. Indeed, there is evidence to suggest this is has already happened.[62]

AI is understood to be able to enhance the effectiveness of disinformation by micro-target users to increase the campaign's chances of successfully misleading as many individuals as

[59] Dobber et al. (2021)
[60] Kertysova (2018)
[61] Arsenault & Kreps (2022)
[62] Kertysova (2018): Verma (2023)

possible. Furthermore, past studies have suggested that generative AI could be used to generate misleading content as "evidence" to support untrue claims. As one study on deepfakes stated, "…deepfake technology can exacerbate political and cultural insularity by customizing content that has congruence with the prior (political or ideological) beliefs held by individuals…."[63]

The gap in understanding in this area surrounds whether this use case for generative AI has already been explored. Is the use of generative AI to empower disinformation campaigns plausible, and if so, has this already occurred?

Methodology

Qualitative Approach

The research questions explored above seek to understand the ways in which the use of generative AI could corrode political trust. Pursuant to this goal, the previous section defined and explored four sub-questions. Though an exploration of existing research and news stories yielded some answers to these questions, there are still gaps in the current understanding of the relationship between generative AI and political trust. Seeking answers to this study's four sub-questions requires exploring a wide range of disciplines, including international relations, political science, information science, and computer engineering involved in AI development. As such, this study pursues a qualitative methodology, as a quantitative hypothesis-testing approach is not well suited to answering such broad questions. In attempting to study concepts as nebulous

[63] Verma (2023)

and subjective as trust, and in seeking to describe the social context in which such concepts exist, qualitative methods are understood to be most appropriate.[64]

The rapid pace of development of AI technologies presents a challenge to understanding how the use of these technologies may affect a concept such as political trust. With the capabilities of generative AI changing on almost a weekly basis, qualitative interviews with subject-area experts were identified as the most appropriate method for seeking answers to this study's research questions. Experts are assumed to be well informed about the realities and capabilities of AI within their given field, and may have access to new information which has yet to reach the public. It is also worth noting a study similar to this one, which examined the relationship between deepfakes and trust in video, also identified a qualitative interview methodology as the most appropriate approach for collecting data.[65]

Participant Selection

With the goal of interviewing a broad selection of knowledgeable experts from across multiple disciplines, initial inquiries were sent out to a selection of individuals with relevant expertise. These initial individuals were selected based upon a combination of their experience working with issues related to generative AI within their individual areas of expertise. It is also worth noting that this study was conducted pursuant to the author's Masters thesis, and so the feasibility of finding willing interviewees played a part in participant selection. Many of the individuals within the initial inquiry group were selected due to an existing professional connection, or an introduction made by an existing connection.

[64] Wildemuth (2016)
[65] Verma (2023)

This study also employed snowball sampling, wherein participants were asked to refer individuals with relevant experience from their own professional networks to the researcher, in order to expand the pool of potential participants. The initial potential participant group was made up of 12 individuals, 5 of whom agreed to be interviewed for this study. From these 5 interviews, 1 additional participant was identified using snowball sampling. The final sample of 6 participants is detailed below in *Table 1*.

Participant Name	Participant Title	Organizational Affiliation	Area(s) of Expertise
Dr. Andrea Hickerson	Dean, School of Journalism and New Media	University of Mississippi	Journalism, Disinformation, Deepfakes
Dr. Kenneth Fleischmann	Professor and Director of Undergraduate Studies	University of Texas Austin	Information Science, AI Ethics
Dr. Christopher Schwartz	Postdoctoral Research Associate, Department of Cybersecurity	Rochester Institute of Technology	Journalism, Cybersecurity, Disinformation
Quentin Miller	Principal Program Manager, AI	Microsoft	AI Development, AI Ethics
Dr. Nitin Verma	Postdoctoral Fellow	School for Future Innovation in Society	Content Trust, Deepfakes, Information Science
Professor Chris Johnson	Faculty Pro-Vice-Chancellor, School of Electronics, Electrical Engineering and Computer Science	Queen's University Belfast	Cybersecurity, Tech Development

Table 1: Expert Interview Participants

Consent & Disclosure

All interviews conducted for this study were done in compliance with University College Dublin's ethical research guidelines. Ahead of each interview, participants were asked to read and sign a consent form detailing the purposes of the study, nature of their participation, potential risks, and information related to data security. This form also included an option for participants to remain anonymous, although all participants who were interviewed chose to be credited by name.

Interview Structure

Interviews conducted for this study were shaped by the information need imposed by the four sub-questions detailed in the previous section. These research questions are open ended, and cannot be answered by a simple yes or no. With such a broad information need, a semi-structured interview format was appropriate. The semi-structured interviews conducted by this study strike a balance between giving participants the latitude to express their views and provide context to them, while also allowing the researcher to steer the conversation to ensure that research goals are achieved.[66] Questions posed were open-ended, with the goal being to facilitate a conversation between the researcher and participants.[67]

Questions posed in interviews fall into two categories: constant questions posed to every participant, and participant-specific questions. Participant-specific questions were included due to the diverse nature of subject-area expertise covered by the participants. For example, a

[66] Creswell & Creswell (2018)
[67] Aberbach & Rockman (2002)

question about the use of generative AI in journalism is appropriate when interviewing a journalist, but not relevant when the participant is a cybersecurity expert.

Several interview participants occupy positions in the tech industry, or act as advisors to various governments and private corporations. To allow participants to speak about these issues as freely as possible, interviews were not recorded. In lieu of recordings, the researcher conducting interviews took notes on ideas expressed by participants, as well as direct quotes. Following each interview, participants were given the opportunity to review a detailed summary of all notes taken during the interview. Only sentiments and quotes which were reviewed and approved by individual interviewees are included in this paper.

Results

Results from the expert interviews conducted in this study serve as a snapshot of informed opinions on the current state of generative AI, as well as how the technology may affect existing social structures. While the sample of experts interviewed for this study is diverse in the fields of study represented by it, the sample size itself is small. Definitive conclusions cannot be confidently drawn from such a small sample, however, the information gained from these expert interviews is useful in identifying issues and setting an agenda for future research. As this section will detail, insights gained from the expert interviews conducted for this study suggest that the concept of *Corrosive AI* is plausible, and that many of the dynamics suggested by this study's research questions appear to already be at play in the world.

Information and insights gained from interviews will be organized by the research sub-questions detailed in the previous section. Each sub-question will be explored individually,

building towards an examination of this study's main research question which incorporates information from all expert interviews.

SQ1: How could open access to generative AI make the technology corrosive to political trust?

As detailed previously, the facts of open public access to various generative AI technologies are relatively well understood. The predominant access model today involves users accessing a website to submit requests for content. These requests are then sent to the service hosting the generative model to be executed, with the resulting content finally being sent back to the user. For generative AI services that use this access model, the actual generation of content is not done locally on the user's device, due to the high computer power required to run large generative models. As a result, the ability to generate content using generative AI is available to nearly any user with an internet connection – powerful computing hardware is *not required.*

Several interviewees expressed that this access model, as well as the larger business model built around it, may be problematic. In particular, experts expressed that the current models promote constant development, but fail to consider unforeseen consequences. Dr. Kenneth Fleischmann of the University of Texas Austin, described the current system of AI development and access as:

"...put it out there and fix what breaks, where your users become part of your design team."

Dr. Fleischmann further expressed that this model is potentially dangerous. Fixing issues as they arrive has the potential to cause significant harm when the technology in question has the potential to significantly influence how users perceive the world around them. According to Fleischmann, this business and development model plays a part in what he describes as a *"potentially existential crisis of trust"* related to generative AI.

Interviewees from within the tech sector also expressed misgivings about the open access model of AI deployment and development. Quentin Miller, Principal Program Manager of Microsoft's AI Platform, stated while discussing deployment models for generative AI:

"I don't think there has been sufficient thinking about how these new technologies can be used…we're putting it [generative AI] out there and making it cheap to use, and someone will misuse it."

Miller also described attempts by Microsoft to mitigate potential harm caused by the misuse of various generative AI technologies with the company's "gated technologies" classification. As detailed by Miller, gated AI technologies are specific applications of AI which Microsoft's internal ethics teams have deemed unsuitable for open public release. The process behind determining which technologies become gated is not publicly available, but Miller was able to provide examples of current gated AI technologies being developed by Microsoft. These included a *custom neural voice model* used to generate speech in the pattern of specific individuals, as well as *real-time video and voice synthesis models*. Miller detailed an example of a prototype of this last type of model being used by a colleague to make themselves look and speak like a specific celebrity on a Microsoft Teams call, in real time.

These snapshots of expert opinion suggest that there is hesitation about the current open access model of generative AI development and deployment among both information scientists and developers in the industry. Open access may allow for more rapid development, but there appear to be misgivings among those most familiar with the process about potential negative externalities resulting from this model.

SQ2: How could the use of generative AI affect existing trends of declining trust in government, media, and video as a medium?

Recent research has established that generative AI is likely to damage trust in video content generally, as well as trust in visual journalism.[68] In a previous section, this paper suggested that a relationship may exist between trust in video, trust in journalism, and political trust. The thinking goes that if video and journalism play a part in the formation of political trust, and generative AI has been shown to damage trust in video and journalism, then generative AI may exacerbate existing declining trends in political trust.

Interviews provided ample evidence to support the existence of a link between trust in video, trust in journalism/media, and political trust. Beginning with media-political trust relationship, Dr. Andrea Hickerson of the University of Mississippi stated when asked about the relationship between news media and political trust:

"There is a correlation between the loss of trust in media and loss of trust in government."

Hickerson went on to express that trust in the media was already low before the advent of widespread access to generative AI, and that she believes that media trust has been falling since the 1990s. This sentiment was echoed by Dr. Christopher Schwartz of the Rochester Institute of Technology, a former journalist focused on counter-disinformation. While discussing the relationship between news media and political trust, Schwartz stated that there *"…isn't much trust in journalists to begin with."* When the conversation shifted to a discussion of deepfakes specifically, Schwartz stated that deepfakes and AI-powered disinformation will "*exacerbate*" the existing situation of low trust in news media. Schwartz went on to say, *"I think it's more*

[68] Verma (2023)

about a loss of trust in institutions generally", expressing that outside of any specific relationship between media trust and political trust, there may also be a crisis of trust generally. Schwartz highlighted mistrust of government institutions and large corporations as examples.

Pivoting towards the video-political trust relationship, information on in this area predominantly comes from one interview subject who has conducted extensive research on deepfakes and trust in video. Dr. Nitin Verma is a Postdoctoral Fellow at the School for Future Innovation in Society, and recently successfully defended a dissertation on the relationship between deepfake technology and public trust in video.[69] While discussing his research on the relationship between generative AI and trust in video and images, Verma expressed his belief that everyday exposure to AI image editing technology, such as the AI-powered photo editing software now built into Google Photos, may damage public trust in photo and video content. Verma stated:

"Even harmless use [of AI-powered photo editing] instills the idea that images can be edited, and it's chipping away at trust."

Verma was also able to provide informed insight into the nature of the relationship between trust in video and political trust. When asked to describe the nature of the relationship, Verma stated:

"...media and video are a large part of how people access politicians. Since we cannot be there to experience politicians and political events in person, we need media to form political opinions."

[69] See Verma (2023)

This statement supports the dynamic described in this paper; wherein video content is integral to the formation of political trust in society today.

Interviews also revealed an additional factor to consider when examining the relationship between AI and political trust: trust in tech companies. While discussing the current state of AI research and development, interviewee Quentin Miller of Microsoft explained the actual hardware needed to develop and test generative AI models is prohibitively expensive, leading the majority of research to be done by large companies or governments. Miller went on to state:

"Furthering the power of generative AI is now in the hands of companies like Microsoft and Google, most research now happens in partnership with them."

Miller went on to express his perception that the tech sector is currently dealing with "*issues of trust*", as well as his belief that considerations for trust broadly are lacking in AI development, saying *"I don't think enough thought is being put into the trust side of it [AI]."* Millers' sentiment was echoed by Dr. Andrea Hickerson, who stated:

"People equate successful tech companies with trust – if you see someone doing well, you tend to assume that they must know what they're doing."

When asked about what role governments should play in regulating generative AI, Hickerson went on to express world governments should begin with *"...regulating tech companies themselves."*

Interviews were particularly useful in filling gaps in understanding related to *SQ2.* Information gained from experts suggests that media and video play roles in the formation of political trust. Political trust is formed via interaction with political figures and institutions, and experts appear to agree that many of these interactions involve video content and/or news media.

Experts expressed that generative AI is already leading to a decline in both trust in video as a medium. Other experts expressed that generative AI may be exacerbating already declining trust in news media. With both news media and video playing roles in political trust formation, it stands to reason that a decline in trust in either of these areas could result in a decline in political trust.

SQ3: In what ways could generative AI be used to create scandals which damage political trust?

The potential of generative AI to be used to manufacture scandals is not currently well understood. This likely results from the rapid pace of development of generative AI, combined with the relatively slow pace of substantive research on uses of the technology. Has generative AI become sophisticated enough to fool people on a wide scale? If so, has this already occurred, or is it likely to occur soon? Expert interviews yielded valuable insight in this area.

A common thread found among participant sentiment regarding the use of generative AI to manufacture scandals is that the technology has the *potential* to be used in this way. However, interviewees also tended to express that we have yet to see this potential fully realized. Dr. Andrea Hickerson of the University of Mississippi stated:

"Deepfakes certainly have the potential to cause scandal, but as of right now it's still just a potential."

This sentiment was echoed by Dr. Nitin Verma, who stated:

"The potential has been demonstrated, but we're still waiting for a headline event."

It is important to note that these statements were made by participants speaking about scandals generally, not specifically about political scandals.

When asked about the potential use of generative AI for creating political scandals, several participants shifted the conversation towards deepfakes, citing the technology as particularly relevant in politics. As discussed previously in this paper, deepfakes are a specific type of generative AI technology that can be used to create fake videos of people doing or saying things they never did or said. To create a deepfake of a subject, the model must be trained on existing authentic content of that subject. In a previous section, this paper suggested that the nature of how deepfakes are trained may make political figures particularly easy subjects to create deepfake of, as there is a wealth of video content available to train models on. Insights gained from expert interviews support this claim. Interviewee Quentin Miller stated when asked about the technical realities of training deepfakes on political actors:

"For politicians, there's a wealth of content to train models on."

While discussing the same topic with Dr. Kenneth Fleischmann of the University of Texas Austin, Fleischmann stated:

"'Seeing is believing' makes deep faking political leaders a very powerful strategy."

Fleischmann's reference to the idea of *"seeing is believing"* here is important, alluding to his belief many people still trust that video content is an authentic representation of the world as it is. This could lead many people to be mislead by deepfake content; as detailed in the discussion of *SQ2* above, the advent of sophisticated generative AI may already be creating a crisis of trust in video.

While several interviewees expressed that generative AI could be used to manufacture content to induce political scandal, the efficacy of AI-generated is not well understood. This is likely due to a lack of real-world examples; we have yet to see generative AI used in this way at scale. When asked about the capability of AI generated content to reliably fool people on a large scale, Dr. Andrea Hickerson expressed that the effectiveness of deepfakes in fooling people may depend both on the realistic look of the deepfake content, and on the believability of the statement made. Dr. Nitin Verma also alluded to other factors at play in determining the believability of AI generated content. When asked about the potential use of generative AI in manufacturing scandals, Verma stated:

"Yes, generative AI could be used to manufacture scandal, especially for those already primed and willing to believe conspiracy theories. These people also tended to have an anachronistic understanding of the technology [deepfakes]."

Verma went on to express a novel point of view regarding the effectiveness of creating political scandal with AI generated content. Verma stated that political figures are among the easiest individuals to deepfake from a technical standpoint, owing to wealth of content available to train models on. At the same time, Verma expressed his belief that political deepfakes are also among the easiest to debunk, attributing this to the "…*high visibility of politicians and they paper trails they leave."*

The consensus among experts on the use of generative AI in manufacturing scandals appears to be that the technology has the potential to be used in this way, but that this potential use case has yet to become prominent. Experts also suggested that the believability of AI generated content is not purely determined by the sophistication of the models used to generate

it, but by other factors such as familiarity with the subject's identity and opinions, or the existing views of the individual viewing the generated content. Interviews further revealed that the effectiveness of AI-generated content in creating political scandal has yet to be demonstrated, owing to a lack of real-world examples. The findings related to SQ3 are efficiently summarized in a statement made by Dr. Christopher Schwarts of the Rochester Institute of Technology:

"We aren't seeing large scale generated content yet, but we're heading towards it."

SQ4: How could disinformation campaigns empowered by generative AI damage political trust?

While the use of generative AI in creating a headline-grabbing political scandal has yet to occur, there appears to be consensus among the experts interviewed for this study that generative AI is already being used to spread disinformation. Knowledge of the use of AI in creating and spreading disinformation was a common thread connecting 5 of the 6 experts participating in this study. As such, these 5 experts were asked: *Do you see AI powered disinformation as a major issue?*

Responses to this question varied in where each participant chose to steer the conversion, but critically, every individual who was asked this question expressed their belief that AI powered disinformation *is already an issue*. Dr. Kenneth Fleischmann of the University of Texas Austin stated:

"[AI powered disinformation is] already an issue and will likely become worse over time."

A similar sentiment was expressed by Dr. Chris Schwartz, who stated his belief that AI-powered disinformation already is an issue, citing the use of machine learning algorithms used to target users with content with social media. While this example of AI-powered disinformation does not feature generative AI, Schwartz went on to express his belief that generative AI is likely

currently being used to create disinformation related to the War in Ukraine, but that examples of it are currently difficult to identify. Dr. Nitin Verma also expressed his belief that generative AI is likely being used to create disinformation related to the Ukraine War, but that *"…we have yet to see a big scandal."* Verma went on to discuss the role which generative AI could play in future disinformation campaigns, saying:

"Generative AI can be used to manufacture flimsy evidence, both as text and media."

The use of generative AI to create this "flimsy evidence" to support misleading claims lends support to the idea that generative AI can be used to empower disinformation campaigns.

With multiple experts stating that generative AI is likely being used to generate disinformation already, why do there appear to be so few examples? The answer appears to be that AI-generated content which is intended to mislead people is very difficult to detect. According to multiple experts, several generative models have reached a level of sophistication that makes the content they are capable of generating extremely difficult to identify. While discussing AI generated speech with Microsoft's Quentin Miller, who holds a Master of Electrical Engineering in speech synthesis, Miller stated:

"You'd have to be an expert to recognize synthesized speech."

Miller went on to explain that generative models are currently capable of generating convincing speech in the style of a specific individual, and that these models require very little data to train on. Miller further expressed that the potential for misuse of this technology led Microsoft to designate their custom neural voice model as a *gated technology*, not available to the public. Dr. Kenneth Fleischmann also stated his belief that AI generated content is difficult to identify, referencing the difficulty in creating AI detectors. He stated:

"With generative AI and AI detectors, we're in a 'one step ahead of the spider' situation."

The selection of experts interviewed for this study appear to agree that AI is currently being used to empower disinformation campaigns. Generative AI was not specifically referenced in this context by every expert, with some instead highlighting the role of machine learning algorithms in allowing disinformers to target specific users with specific content. However, 3 interviewees did express their belief that generative AI has the potential to be used in ways that could empower disinformation campaigns, most notably in the creation of synthesized speech that can convincingly mimic individual voices. While some experts stated generative AI is likely already being used to create and spread disinformation, examples have proven difficult to identify, likely due to the sophistication of current generative models. Critically, insights gained from interviewing developers from within the tech sector reveal that even the companies developing generative models recognize the potential for misuse. All evidence gained from interviews related to generative AI and disinformation seems to suggest that generative AI has the potential to enhance disinformation campaigns, and that this may already be happening.

RQ1: In what ways could the use of generative AI corrode political trust?

Based on a review of the literature, the author of this paper concluded that the research question at the core of this paper would best be answered by exploring four salient factors related to the use of generative AI that have the potential to damage political trust. These four factors, explored above in the form of research sub-questions, informed the questions posed to the experts interviewed for this study. With information from expert interviews leveraged to discuss this study's four sub-questions, this paper can now attempt to answer its core research question: *in what ways could the use of generative AI corrode political trust?*

The first salient factor, explored in *SQ1*, surrounds the current open access business and development model adopted by numerous organizations developing generative AI. This paper predicted that an open access model would lead to misuse, as the developers of generative models cannot fully predict or control how end users will use the technology. In interviews, several experts expressed concerns with the open access models for generative AI development and deployment. Many expressed a feeling that negative externalities and ethical considerations are not being made, and that the potential for misuse is high. As one senior developer at Microsoft stated while discussing this topic, "*we're putting it [generative AI] out there and making it cheap to use, and someone will misuse it."*[70]

Regarding the relationship between trust in video, trust in media, and political trust explored in *SQ2*, this study again found consensus among the sample of experts interviewed. Interviewees with specific knowledge about the relationship between generative AI and trust in video expressed that generative AI is already damaging people's trust in the authenticity of video content. Experienced journalists interviewed expressed that generative AI is likely exacerbating already declining trust in news media. Experts also suggested that both trust in video and trust in the media should affect political trust, as the process of trust formation today involves experiencing political actors and events through video content, often distributed and contextualized by the news media. These findings suggest that generative AI is corroding political trust by damaging trust in the content used to form trust, and the organization that broadcast and contextualize that content.

Regarding *SQ3* and the use of generative AI to manufacture scandals, experts agreed that generative AI has the *potential* to create scandal. However, there was a lack of consensus among

[70] See interviewee Quentin Miller

interviewees regarding the effectiveness of AI-generated content in fooling people at scale. This lack of consensus seems to result from a lack of real-world examples: we have yet to see a headline-grabbing scandal based on AI generated content. Interviews further suggested that the lack of concrete evidence in this area is likely due to the difficulty of identifying AI generated content. This combination of a lack of consensus and a lack of evidence makes studying the use of generative AI in creating political scandals a exciting area for future research. With the information available now however, experts agree that while AI generated scandals are possible, they remain purely a possibility for now.

Finally, regarding the use of generative AI to enhance disinformation explored in *SQ4,* there was consensus among the sample of experts interviewed that AI powered disinformation is already an issue. Multiple interviewees expressed that many current examples of AI powered disinformation do not necessarily include generative AI, referencing the incorporation of machine learning into the algorithms that target users with content. However, several experts expressed that generative AI has the potential to be used to create convincing disinformation, specifically in the recreation of individual voice patterns. Combined with the open access model explored in *SQ1,* it appears that the ability to generate convincing disinformation at a rapid pace will soon be in the hands of millions of users. Interviews further revealed that the organizations developing generative models at least partially understand the potential for misuse of the technology, as seen in Microsoft's "gated technologies" designation. In all, interviews showed a consensus among experts that generative AI can be used to enhance the quality of disinformation, and that open access to the technology may expand the scope of *who* can produce misleading content that could be used to disinform.

The findings of this study suggest that generative AI has the potential to be used in multiple ways which could damage political trust. Expert interviews revealed support for the idea that generative AI could be used to manufacture political scandals, which are understood to be damaging to political trust.[71] This study also found support for the idea that generative AI can be used to create and spreading disinformation, and that this potential use case is at least partially understood by the organizations developing it. Experts also expressed that the current open access model of generative AI development and deployment makes misuse probable. Considering the open access model alongside the potential use cases of generative AI in manufacturing scandals and disinformation, it seems clear that generative AI has the potential to damage political trust. What makes generative AI potentially *corrosive* to political trust is the fact that the technology already appears to be rapidly damaging trust in the content and organizations involved in the political trust formation process. Interviews and existing literature demonstrate that generative AI is likely damaging to trust in the video content people use to form political trust, and damaging to trust in the news media organizations through which that content is broadcast and contextualized.

Conclusion

This paper sought to better understand the relationship between varied uses of generative AI technologies and political trust. An examination of the literature suggested that rather than

[71] Keele (2007)

eroding trust slowly, various factors related to the development, deployment, and use of generative AI mean that the technology could *corrode* political trust more quickly than ever before. In seeking to answer the question "*In what ways could the use of generative AI corrode political trust?",* this paper identified four salient factors related to the use of generative AI which appear to have the potential to damage political trust. These factors were expanded into research sub-questions, which informed the questions posed to the sample to experts interviewed in the study conducted for this paper.

Expert interviews yielded information which supports the model of corrosive AI presented by this paper. Interviews also identified a gap in the current understanding of how generative AI is being used. This gap became evident when experts were asked about the potential use of generative AI to enhance disinformation, as well as the technology's potential to manufacture scandals. In both cases, experts expressed their belief that both of these use cases are probable, but that there is a lack of real-world examples available for study. Future scholarship could therefore focus on identifying instances of AI generated disinformation, or AI generated content which leads to political scandal. With few examples currently available, the identification and study of instances of generative AI being used in these ways would enhance current understanding of how the technology is being used in the real world. A study seeking to identify generative AI being used in these ways could also potentially provide evidence to support the model of corrosive AI proposed in this paper.

Finally, with support evident among interview participants for the corrosive potential of generative AI, information from both interviews and the literature was used to identify a number of potential solutions. With the potential for AI generated content to be used to mislead people, in both the context of political scandals and disinformation, the addition of digital identifiers to

generated content may offer a way help people differentiate between generated and authentic content. Digital identifiers embedded within generated content do not necessarily need to be visible watermarks; invisible embedded identifiers could similarly offer a solution to issues in identifying AI generated content when that content is used to spread disinformation. However, both of these potential solutions require cooperation from the organizations developing generative models. Government regulation seems the obvious answer here, however, interview participants expressed hesitation regarding the efficacy of governments in regulating such a new and rapidly changing technology. In the words of interviewee Dr. Christ Schwartz, the likely result of regulation efforts would be "Politicians with no clue what's going on, versus developers who have all the information and vested interests the technology." Finding a solution to mitigate the corrosive potential of generative AI likely requires further research. As this study has demonstrated, while the potential of generative AI to corrode public trust appears clear, the extent to which that potential is being realized in the real world is not well understood.

References

Aberbach, Joel D., and Bert A. Rockman. 2002. "Conducting and Coding Elite Interviews." *PS: Political Science & Politics* 35(4): 673–76.

Arsenault, Amelia C., and Sarah E. Kreps. "AI and International Politics." In *The Oxford Handbook of AI Governance*, eds. Justin B. Bullock et al. Oxford University Press, 0. https://doi.org/10.1093/oxfordhb/9780197579329.013.49 (April 14, 2023).

Aspray, William, and Philip Doty. "Does Technology Really Outpace Policy, and Does It Matter? A Primer for Technical Experts and Others." *Journal of the Association for Information Science and Technology* n/a(n/a). https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24762 (May 17, 2023).

Bennett, W. Lance, and Steven Livingston, eds. 2020. *The Disinformation Age*. Cambridge: Cambridge University Press. https://www.cambridge.org/core/books/disinformation-age/1F4751119C7C4693E514C249E0F0F997 (July 21, 2023).

Chesney, Bobby, and Danielle Citron, eds. 2019. "Deep Fakes: A Looming Challenge for Privacy." *California Law Review*.

Creswell, John W., and J. David Creswell. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.

Dafoe, Allan. 2015. "On Technological Determinism: A Typology, Scope Conditions, and a Mechanism." *Science, Technology, & Human Values* 40(6): 1047–76.

Dobber, Tom et al. 2021. "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?" *The International Journal of Press/Politics* 26(1): 69–91.

Erlich, Aaron, and Calvin Garner. 2023. "Is Pro-Kremlin Disinformation Effective? Evidence from Ukraine." *The International Journal of Press/Politics* 28(1): 5–28.

Freelon, Deen, and Chris Wells. 2020. "Disinformation as Political Communication." *Political Communication* 37(2): 145–56.

Gambetta, Diego. 2000. "Can We Trust Trust." In *Trust: Making and Breaking Cooperative Relations,* Blackwell, 213–37. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.5695&rep=rep1&type=pdf.

Glanville, Jennifer L., and Pamela Paxton. 2007. "How Do We Learn to Trust? A Confirmatory Tetrad Analysis of the Sources of Generalized Trust." *Social Psychology Quarterly* 70(3): 230–42.

Gosse, Chandell, and Jacquelyn Burkell. 2020. "Politics and Porn: How News Media Characterizes Problems Presented by Deepfakes." *Critical Studies in Media Communication* 37(5): 497–511.

Hameleers, Michael. 2020. "Populist Disinformation: Exploring Intersections between Online Populism and Disinformation in the US and the Netherlands." *Politics and Governance* 8(1): 146–57.

Hetherington, Marc J. 1998. "The Political Relevance of Political Trust." *American Political Science Review* 92(4): 791–808.

Hetherington, Marc J., and Jason A. Husser. 2012. "How Trust Matters: The Changing Political Relevance of Political Trust." *American Journal of Political Science* 56(2): 312–25.

Hoffner, Cynthia A., and Bradley J. Bond. 2022. "Parasocial Relationships, Social Media, & Well-Being." *Current Opinion in Psychology* 45: 101306.

Horowitz, Michael C et al. "Artificial Intelligence and International Security."

Huang, Jiaxin et al. 2022. "Large Language Models Can Self-Improve." http://arxiv.org/abs/2210.11610 (July 21, 2023).

Isaac, Mike, and Cade Metz. 2023. “Meta Unveils a More Powerful A.I. and Isn’t Fretting Over Who Uses It.” *The New York Times*. https://www.nytimes.com/2023/07/18/technology/meta-ai-open-source.html (July 22, 2023).

Jablonka, Kevin Maik, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2023. “Is GPT-3 All You Need for Low-Data Discovery in Chemistry?” https://chemrxiv.org/engage/chemrxiv/article-details/63eb5a669da0bc6b33e97a35 (July 21, 2023).

Keele, Luke. 2007. “Social Capital and the Dynamics of Trust in Government.” *American Journal of Political Science* 51(2): 241–54.

Kertysova, Katarina. 2018. “Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation Is Produced, Disseminated, and Can Be Countered.” *Security and Human Rights* 29(1–4): 55–81.

Lanoszka, Alexander. 2019. “Disinformation in International Politics.” *European Journal of International Security* 4(2): 227–48.

Liu, Piper Liping. 2023. “Parasocial Relationship in the Context of the COVID-19 Pandemic: A Moderated Mediation Model of Digital Media Exposure on Political Trust among Chinese Young People.” *Computers in Human Behavior* 141: 107639.

Martens, Bertin, Luis Aguiar, Estrella Gomez-Herrera, and Frank Mueller-Langer. 2018. “The Digital Transformation of News Media and the Rise of Disinformation and Fake News.” https://papers.ssrn.com/abstract=3164170 (July 21, 2023).

Mcknight, D. Harrison, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. “Trust in a Specific Technology: An Investigation of Its Components and Measures.” *ACM Transactions on Management Information Systems* 2(2): 12:1-12:25.

“Meta and Microsoft Introduce the Next Generation of Llama.” 2023. *Meta*. https://about.fb.com/news/2023/07/llama-2/ (July 22, 2023).

Schippers, Birgit. 2020. “Artificial Intelligence and Democratic Politics.” *Political Insight* 11(1): 32–35.

Schwartz, Christopher. 2023. “Events That Never Happened Could Influence the 2024 Presidential Election – a Cybersecurity Researcher Explains Situation Deepfakes.” *The Conversation*. http://theconversation.com/events-that-never-happened-could-influence-the-2024-presidential-election-a-cybersecurity-researcher-explains-situation-deepfakes-206034 (July 19, 2023).

Somers, Meredith. 2023. “Deepfakes, Explained.” *MIT Sloan*. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained (July 22, 2023).

Tenove, Chris. 2020. “Protecting Democracy from Disinformation: Normative Threats and Policy Responses.” *The International Journal of Press/Politics* 25(3): 517–37.

The A.I. Dilemma. 2023. https://www.youtube.com/watch?v=xoVJKj8lcNQ (July 19, 2023).

Trithart, Albert. 2022. “Disinformation against UN Peacekeeping Operations.” *International Peace Institute*.

Verma, Nitin. 2023. “Deepfake Technology and the Future of Public Trust in Video.” Dissertation. University of Texas Austin.

Wei, Jason et al. 2022. “Emergent Abilities of Large Language Models.” http://arxiv.org/abs/2206.07682 (July 21, 2023).

“What Is Artificial Intelligence (AI) ? | IBM.” https://www.ibm.com/topics/artificial-intelligence (June 4, 2023).

Wildemuth, Barbara M. 2016. *Applications of Social Research Methods to Questions in Information and Library Science, 2nd Edition*. ABC-CLIO.

APPENDIX

Item 1. Survey Participant Informed Consent Form

I am asking you to participate in a research study as a part of a master's thesis project, titled "Corrosive AI: How the use of Generative Artificial Intelligence Threatens Trust in Government". I will describe this study to you and answer any of your questions. This study is being led by Riley Lankes. The Faculty Advisor for this study is Dr. Stefan Muller, University College Dublin, School of Politics and International Relations.

Purpose of the Study

The purpose of this research is to gauge expert opinion on the oncoming impacts of generative AI across a selection of subject matter experts in AI development, technology regulation, information science, and international politics.

Nature of Participation

I will ask you a series of questions related to your thoughts on how generative AI is currently impacting trust on social media platforms and news organizations. I will also ask for your thoughts on how the technology may progress in the next few years, as well as your opinions on how it should be regulated.

Risks and Discomforts

I do not anticipate any risks from participating in this research.

Audio/Video Recording

Audio and video recordings will **not** be made of any interviews. The interviewer (Riley) will take text notes of some statements made and sentiments expressed. After the interview has concluded, participants will receive a follow-up message detailing notes taken during their interview. Participants will have the opportunity to change or retract any statements made.

Only statements which participants review and explicitly approve for inclusion will be used in the final paper.

Anonymity

Any participant may choose to remain anonymous. Participants who choose to do so will only be referred to as "an expert in [X field]" within the final publication, with no identifiable information included. Identifiable data for all participants will be kept secure, as described in the "Privacy/Confidentiality/Data Security" section below. Participants may also outline any specific requests about the degree to which they wish to remain anonymous.

Privacy/Confidentiality/Data Security

All interview notes will be stored in a de-identified file for each individual. A list identifying each subject will be kept in a separate file. All files will be encrypted and stored securely in a password-protected folder, with backups stored on an encrypted, password protected backup drive. No data will be stored online in cloud-based storage services, in order to protect from data theft. De-identified data from this study may be shared with the research community at large to advance understanding of AI and trust. I will remove or code any personal information that could identify you before files are shared with other researchers to ensure that, by current scientific standards and known methods, no one will be able to identify you from the information we share. Despite these measures, we cannot guarantee the anonymity of your personal data.

If You Have Questions

The main researcher conducting this study is Riley Lankes, a graduate student at University College Dublin. If you have questions, contact Riley at LankesRileyD@gmail.com or at +1 (315) 751-8702.

Anonymity Decision

I wish to remain anonymous in the final publication. ☐

I wish to be credited in the final publication. ☐

Please detail any specific requests about how you wish to be credited/remain anonymous below:

Statement of Consent

I have read and understood the information above. I consent to take part in the study.

Your Signature_____ Date_____

Your Name (printed)_____

Signature of person obtaining consent_____ Date_____

Printed name of person obtaining consent_____

This consent form will be kept by the researcher for two years beyond the end of the study.

Item 2. Table of Interview Participants (*Table 1)*

Participant Name	Participant Title	Organizational Affiliation	Area(s) of Expertise
Dr. Andrea Hickerson	Dean, School of Journalism and New Media	University of Mississippi	Journalism, Disinformation, Deepfakes
Dr. Kenneth Fleischmann	Professor and Director of Undergraduate Studies	University of Texas Austin	Information Science, AI Ethics
Dr. Christopher Schwartz	Postdoctoral Research Associate, Department of Cybersecurity	Rochester Institute of Technology	Journalism, Cybersecurity, Disinformation
Quentin Miller	Principal Program Manager, AI	Microsoft	AI Development, AI Ethics
Dr. Nitin Verma	Postdoctoral Fellow	School for Future Innovation in Society	Content Trust, Deepfakes, Information Science
Professor Chris Johnson	Faculty Pro-Vice-Chancellor, School of Electronics, Electrical Engineering and Computer Science	Queen's University Belfast	Cybersecurity, Tech Development

Table 1: Expert Interview Participants