Stanford | Stanford Existential Risks Initiative
Center for International Security and Cooperation
Freeman Spogli Institute

Psychological and Psychosocial Consequences of Super Disruptive A.I.: Public Health Implications and Recommendations

David D. Luxton [1]*, Eleanor Watson [2]

Citation: Luxton, D. D., Watson., E; Psychological and Psychosocial Consequences of Super Disruptive A.I.: Public Health Implications and Recommendations. Proceedings of the Stanford Existential Risk Conference 2023, 60-74. https://doi.org/10.25740/mg941vt9619

Academic Editor: Trond Undheim, Dan Zimmer

Funding: Not applicable.

Conflict of Interest Statement: The authors declare no conflict of interest.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors wish to extend their gratitude to the SERI Team.

Author Contributions: Conceptualization, D.L., and E.W.; writing—original draft preparation, D.L. and E.W.; writing—review and editing, D.L. and E.W. All authors have read and agreed to the published version of the manuscript.

Abstract: A moral panic is burgeoning now that the disruptive impacts of A.I. are becoming unmistakable, and further advancements shall have increasingly profound ramifications on how we interact with one another, how we experience our lives, and how we perceive the future and the world around us. With these profound, existential changes, there will also be implications for the psychological and psychosocial well-being of people around the world. The psychological and psychosocial impacts of A.I. from a public health perspective deserve attention. The purpose of this paper is, therefore, to begin to elucidate these issues and provide recommendations. Topics discussed include the mechanisms of psychological and social impacts, including threats to interpersonal and societal trust, problems of deception, overreliance on machines in decision-making, and the public health risks caused by the displacement and disenfranchisement of persons in work sectors most impacted by A.I. Recommendations are presented for addressing these emerging issues to be considered by technology developers, policy-makers, ethicists, healthcare clinicians, politicians, and the general public.

Keywords: artificial intelligence, psychology, well-being, mental health, disruption, shock

[1] University of Washington, School of Medicine, Department of Psychiatry and Behavioral Sciences, 1959 NE Pacific St, Seattle, WA 98195, United States; ddluxton@uw.edu (D.L)

[2] University of Gloucestershire, School of Computing and Engineering, The Park, Cheltenham, GL50 2RH, United Kingdom; eleanorwatson@connect.glos.ac.uk (E.W)

* Correspondence: ddluxton@uw.edu

1. Introduction

The advent of new technology has often brought about significant societal changes and a range of positive and negative impacts. In some cases, the introduction of new technology has resulted in the creation of unanticipated social problems. Unfortunately, as history has all too often demonstrated, awareness of the potential negative consequences of any new technological innovation, and thus opportunities to mitigate those consequences, is often lacking, and it is only after a new technology is put to use or brought to market that the problems become evident. For example, the rise of the Industrial Revolution in the 18th and 19th centuries led to the growth of cities and the development of new forms of work and leisure, as well as new social problems such as pollution, overcrowding, and the exploitation of workers. Similarly, the widespread adoption of the Internet and the rise of social media in the 21st Century have significantly improved how we communicate and access information but have also contributed to new social problems impacting public health, such as Internet addiction, cyberbullying, online harassment, self-directed violence, and social isolation (Luxton et al., 2012; Schou Andreassen et al., 2016). While technological impacts on societal change are often gradual and incremental, such as with the examples mentioned above, history has shown that sudden technological disruption can cause intense reactive fear and anxiety, such as following the detonation of the first atomic weapons and the launch of Sputnik in the mid-20th Century.

The rise of artificial intelligence (A.I.) has also been incremental and with leaps, and the topic of both enthusiasm and consternation. The last few years have been especially exciting given the significant advancements in machine learning (ML), natural language processing, and computer vision, resulting in a major impact on the global economy (Dean, 2019; The White House, 2022). The development of powerful prompt-driven multimodal generative A.I. technologies, such as ChatGPT, has the potential to significantly transform the way in which we provide, receive, and share information in social interactions. One key impact of these technologies is their ability to generate a wide range of content, such as images, videos, and audio, based on a given prompt. This can enable the creation of new forms of communication and expression that may not have been possible before and can also lead to the automation of certain tasks that are currently performed by humans, such as content creation and moderation. Another potential impact of these technologies is their ability to interact with users in a dynamic and personalized way. This can enable the creation of immersive and interactive social experiences that can adapt to the interests and preferences of individual users.

Just like previous technological advancements, the latest in A.I. developments are having a range of impacts on society, both positive and negative. On the positive side, A.I. has the potential to increase efficiency and productivity by automating tasks and making better decisions than humans. This can lead to cost savings and increased competitiveness for businesses and governments that adopt A.I. technologies. A.I. can also create new opportunities for innovation and growth, as it can enable the development of new products and services across domains, such as art and entertainment, business, transportation, and healthcare, that were not previously possible.

There are also potential negative impacts of A.I. on society, such as the potential for job displacement as A.I. becomes more prevalent in various industries. This could lead to workers losing their jobs or requiring retraining for new roles, which could negatively affect their financial well-being. In addition, there are also concerns about the potential for A.I. to perpetuate and amplify biases and prejudices, which could lead to unfair and discriminatory outcomes that disproportionately affect certain groups of people. The increasing use of A.I. in various areas of society also carries with it the potential for psychological harm to individuals due to the potential loss of traditional humanitarian

values. These values may include the principle of putting people first, the value of autonomy and self-determination, the right to privacy, and the importance of pride in one's work.

Given the rapid emergence of generative A.I. and leaps in associated enabling technologies, such as cloud-based super-computing and virtual reality, sudden technological super-disruption is a possibility. We define super-disruptive A.I. as an advance in A.I. that profoundly and unexpectedly alters societal functioning at a large scale, whether caused by an incredible technological leap or incremental progression with cascading effects on society. We recognize that the impact of A.I. on society is complex and multifaceted, with both negative and positive outcomes as possibilities. It is therefore essential to consider the potential benefits and risks of A.I. carefully and to take steps to mitigate any negative consequences, such as investing in education and training programs to help workers adapt to the changing job market, as well as guidelines, regulations, and policies to ensure that the development and deployment of A.I. are responsible and ethical.

Our goal with this paper is to illuminate the psychosocial and psychological risks and implications of super-disruptive A.I., particularly those that have thus far received a lesser amount of discussion, such as the risk of supernormal stimuli experiences creating new social dysfunction and disparities in access to A.I. that have psychosocial implications. We highlight current risks and speculative near-future threats across varied societal domains and economic sectors, including entertainment, healthcare, business, criminal justice, public safety, and defense. We also posit an intersection between A.I. and the emergence of a new theater of war based upon demoralization, which seems likely to intersect with moral panic (with the risk of widespread paranoia or mass hysteria). Finally, we set an expectation for an imminent shock of an "AI Sputnik moment" and provide high-level recommendations to guide the public through this adjustment.

2. Current and Emerging Threats

2.1 Algorithmic Distortions of Reality and Behavioral Manipulation

Social media is one of the most pervasive and influential modern technologies underpinned by A.I. TikTok, and other social media platforms use AI-driven algorithms to create personalized content feeds designed to keep users engaged for as long as possible within filter bubbles of comfortably familiar ideas. These algorithms personalize content and user recommendations based on their past activity and behavior. While these algorithms can be beneficial in helping users discover new content and connections, they can also have negative consequences. Social media algorithms can contribute to the polarization of political views and the spread of misinformation. This issue can occur when algorithms prioritize content that aligns with a user's preexisting beliefs and values, creating echo chambers in which users are only exposed to information that confirms their views. This can lead to a narrowing of perspectives and a lack of exposure to diverse viewpoints and information, which can contribute to the polarization of opinions and the strengthening of extremist views (Wojcieszak, 2010). In addition, social media algorithms can also contribute to negative perceptions of self-worth by promoting a distorted view of reality (Luxton et al., 2012). For example, social media algorithms may prioritize content that is designed to be attention-grabbing or sensational, which can create a skewed perception of what is normal or desirable. This can lead to feelings of inadequacy or anxiety among users who may compare themselves to others based on their online profiles or posts (Braghieri et al., 2022). Social media use is also linked to depressive disorders and suicidal behavior, especially among young people (Sedgwick et al., 2019).

A.I. is also increasingly used to intentionally manipulate and control others, which can harm people while undermining or violating individuals' rights to freedom of expression, privacy, and non-discrimination. This can occur when tech platforms act as de facto gatekeepers of online speech and content and use their power to enforce their policies and regulations in an arbitrary, biased, or inconsistent manner. One example is the use of algorithms and automated systems to enforce content moderation policies, which can lead to the removal of legitimate speech or the failure to remove harmful or defamatory content. Moreover, using automated bots to prop up opinions and influence others via social media is also a threat. The use of A.I. to suppress speech in social media received much public attention with Elon Musk's purchase of the social media site Twitter in 2022. The subsequent release of the "Twitter Files" and Congressional hearing revealed deep gatekeeping and biased censorship, not only by the company but also by the influence of the government and powerful political individuals (Sardarizadeh & Schraer, 2022; U.S. House of Representatives, 2023).

The psychological and psychosocial impact of using A.I. to censor and manipulate online content can create a chilling effect on free expression, as individuals may self-censor or refrain from sharing particular ideas or perspectives for fear of being censored or punished by the platform. Psychological studies have shown that the suppression of free speech and expression contributes to negative emotional states (Krøvel & Thowsen, 2019; Maitra & McGowan, 2012; Parekh, 2017). It's firmly established those verbalizing thoughts and feelings increases a person's sense of control, increases understanding of the motivations of others, and thus reduces anxiety (Ayers, 2009; Niles et al., 2015; Lepore et al., 2000).

2.2 Algorithmic Prejudice and Unfairness

A risk of prejudice is presented by A.I. when it overfits on characteristics that are not germane to the question or which may be secondary indicators of protected characteristics (e.g., whether one buys certain beauty products). The use of A.I. in decision-making processes carries with it the risk of perpetuating and amplifying biases and prejudices that are present in the data and algorithms used to train and operate the A.I. system. This can occur when the A.I. system overfits on characteristics that are not relevant to the task at hand or which may be correlated with protected characteristics such as race, gender, age, or sexual orientation. Algorithmic prejudice is especially troublesome when human oversight is no longer present or has become complacent due to trusting a system too much. Examples such as the Horizon Post Office Scandal illustrate that prejudicial systems that wrongly label and harass people as having committed illegal actions can lead to enormous miscarriages of justice (Peachey, 2022).

Disparities in outcomes that result from algorithmic bias have the potential to cause harm to the health and well-being of people (Luxton & Poulin, 2020). For example, if an A.I. system is trained to identify health risks or to assist in making decisions about medical treatments, but the source data is biased due to under-representative learning samples, such as by race or socio-economic status, then some groups may suffer from less-than-optimal treatment options that lead to degraded health outcomes, or in some situations, no treatment at all. Another example is the use of algorithmic risk prediction in criminal justice systems. Algorithmic bias is a problem for systems that are used to determine level-of-risk for incarcerated persons who are considered for release into the community or when algorithms are used in the sentencing process (McKay, 2020). Actual bias or the perception of bias due to lack of transparency seeds distrust in people, which can lead to feelings of resentment and discontent among those who are negatively affected and thus exacerbate social tensions.

2.3 Supernormal Stimuli

There is a risk that people may become lost within endless fascinating generative worlds created by artificial intelligence (A.I.) and other technologies, leading to a phenomenon known as "irresistible supernormal stimuli." Having roots in animal behavior and evolutionary psychology (Barrett, 2010), this refers to the idea that people may become excessively drawn to and engaged with artificial stimuli designed to be more appealing or rewarding than their real-life counterparts. The risk of irresistible supernormal stimuli is particularly concerning in the context of A.I. technologies that can generate vast amounts of content, such as virtual reality environments or chatbots that can engage in endless conversations. These technologies can provide an endless source of stimulation and engagement, leading people to become excessively drawn to and reliant on them. The over-engagement with irresistible supernormal stimuli can have negative consequences for individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. It can also have negative impacts on mental health, as it may lead to feelings of isolation and disconnection from the real world.

2.4 Demoralization and Anomie

While A.I. can bring many benefits, including increased efficiency and productivity, it also carries with it the potential for negative consequences, including the risk of demoralization, humiliation, and discontent among certain segments of the population. One potential risk is that the increasing use of A.I. in various fields may lead to job displacement, as machines and algorithms are able to perform tasks more efficiently than humans. This can lead to feelings of anxiety and insecurity, as well as a loss of pride in one's work and a sense of meaninglessness. Another issue is the risk of traditional middle management being increasingly given to algorithms for expediency's sake, resulting in workplaces with potentially less employee autonomy and omnipresent scrutiny of the smallest details, including unfair recording of infractions. It is ironic that as machine autonomy advances, humans are increasingly finding themselves robbed of their own autonomy.

2.5 Alteration of Human Interactions and Trust

Disparities in access to A.I. and perceptions about another's use of it will undoubtedly alter everyday social interactions, the trust between people, and ultimately psychological well-being (Luxton, 2014). At an individual level, distrust may result when someone comes into contact with another person who's suspected of using A.I. to give them an unfair advantage.

Imagine a scenario when you meet a stranger who may or may not have a Neuralink™ implant that gives them instant access to cloud-based A.I., allowing them to have detailed knowledge about you. This issue has come into the public spotlight in past years with the development of smart glasses and other smart wearable technologies (Due, 2015; Nonan, 2013; Schuster, 2014). Another scenario is when a colleague may have access to A.I. that provides them with a superior edge over you in particular tasks or domains, from leisure (e.g., a poker game) to professional settings (e.g., a court proceeding). It is now feasible for people to access the capabilities of ChatGPT remotely, or use technology such as A.I. 'Cyrano de Bergeracs,' which can whisper answers into Bluetooth earbuds.

Perception of unfair access to and use of A.I. does not need to be in real-time during interpersonal interaction but at any time before or after an interaction. It is probable the uncertainty about whether another person is using A.I. to their advantage during an interaction; however, that may be most problematic from a psychological health

perspective. Knowing there is an advantage allows a person to adjust to and adapt, reducing psychological unease, whereas not knowing is likely to contribute to maladaptive stress, anxiety, and paranoia, consistent with scientific studies documenting the link between uncertainty and problems with mental health (Massazza et al., 2022).

The disparity in access to A.I. technology could also lead to shared distrust and antagonization towards groups perceived to have unfair advantages and privileges because of their access to A.I. If only wealthy elites or governments have access to advanced A.I. and solely possess authoritative decision-making regarding it, then a perception of unfair inequality and distrust among the population would ensue, as do disparities of any other resource (Mirza, et al., 2019; Bénabou, 2003; Hargittai, 2011; Van Dijk, 2006). However, a disparity in access or understanding of A.I. may be exacerbated by distrust in the use of technology by the elite to surveil and control others (Wee & Findlay, 2020; Feldstein, 2019; Matthews, 2021; Manheim & Kaplan, 2018).

Competition between governments and corporations to advance and deploy A.I. for strategic and tactical advantages may also have deleterious consequences for psychosocial well-being. This is similar to what was experienced during the nuclear arms race between the United States and the Soviet Union during the Cold War – each side continued to build their nuclear arsenals, promising peace through deterrence and assured mutual destruction. Indeed, the "A.I. arms race" has begun, and while development and strategy may remain clandestine and under secrecy, what results are increasingly aggressive and Machiavellian behaviors and too, fear and distrust among the general population.

The use of A.I. and associated technologies in criminal justice and policing also has implications for social distrust. For example, the increasing use of mass facial scanning and surveillance at public places, such as airports, has the potential to increase psychological distress in the forms of anxiety and distrust among citizens regarding how data about them is used. Surveillance is also linked to behavioral changes, including self-censorship and avoidance of assembly, and the right to protest for fear of retribution (Starr et al., 2008; O'Connor & Jahan, 2014; Kaminski & Witnov, 2015). And as we noted earlier, algorithmic biases that result in disparities in criminal justice outcomes have the potential to increase distrust and increase social unrest.

2.6 AI-enabled Deception

Artificial intelligence (A.I.)-enabled forms of deception, such as deepfakes, refer to the use of A.I. technologies to create and disseminate convincing, yet fake or misleading, images, videos, and other media. These technologies can be used to manipulate the appearance or content of media in ways that are difficult to detect, and that can potentially deceive or mislead viewers.

The use of AI-enabled deception can have a range of negative impacts on social trust and cohesion at the individual, organizational, and societal levels. For example, deepfakes can be used to create and disseminate non-consensual or malicious content, such as revenge porn or harassment, which can have serious consequences for the victims. This can lead to feelings of betrayal, mistrust, and fear, which can have deleterious effects on mental health while further undermining social cohesion. Another concern is that deepfakes and other forms of AI-enabled deception can be used to spread misinformation or propaganda, which can erode trust in information and institutions (Luxton, 2022). This can lead to the erosion of social norms and values, as well as to social polarization and division.

It is essential to be aware of these risks and to take steps to mitigate them. This can include measures such as regulations and policies to prevent the use of AI-enabled deception for harmful purposes, as well as efforts to promote media literacy and critical thinking skills.

2.7 Breakdown of Human Relationships

There is a risk that an over-affinity with A.I. waifus' (digital companions designed to be attractive and submissive) could lead to a breakdown in romantic and friendship relationships. This risk is particularly concerning in the context of incel identity, where individuals may feel isolated and rejected from romantic opportunities and turn to A.I. waifus' as a substitute for real-life relationships. This may have negative consequences for individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. This can also contribute to a breakdown in traditional dating styles, as individuals may rely more on digital interactions with virtual partners rather than seeking out real-life relationships. In addition, the use of "hook-up apps", which enable dating outside of one's social circle, can also contribute to a breakdown in traditional dating styles and may lead to a concentration of attention on a few algorithmic-selected matches. This can perpetuate a cycle of social isolation and frustration, as individuals may feel that they are not able to compete for attention and connection with others.

As a society, we remain particularly underprepared for the potential impacts of A.I. on certain vulnerable groups, such as lost youth and hikikomori. 'Lost youth' refers to young people who may feel disconnected from society and may lack a sense of purpose or direction in life. Hikkikomori refers to a social phenomenon in Japan where individuals, typically young men, become isolated and withdraw from society, often spending most of their time alone in their rooms (Teo & Gaw, 2010). This risk is particularly concerning in the context of A.I. technologies that are able to generate vast amounts of content, such as virtual reality environments or chatbots that can engage in endless conversations and consistent virtual relationships. The over-engagement with irresistible supernormal stimuli can have negative consequences for individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. Where on one hand, A.I. technologies such as virtual companions may provide social support and reduce social isolation; they can also have negative impacts on mental health, as they may lead to feelings of isolation and disconnection from the real world.

The use of AI-enabled virtual care providers in place of human care providers is another potential threat to human relationships (Luxton, 2014a; Luxton, 2014b). Virtual therapists in the form of chatbots apps and virtual human therapists are becoming increasingly capable and popular, and while they provide numerous benefits, such as 24/7 availability and customization, they also come with some drawbacks. A lack of human connection between a care provider, such as between a psychotherapist and patient, may minimize the humanness in the therapeutic process and potentially diminish the therapeutic relationship, which has been shown to be the strongest predictor of positive therapeutic outcomes (Luxton, 2014b).

2.8 Loss of Human Creativity, Inspiration, and Drive

Emerging generative A.I. is beginning to outdo all the creative achievements and potentials of humans, to include the visual arts, music, and literature. This new AI-generated art will rapidly devalue the creative expression of human artists, not just because it is technically or thematically impressive, but also because it is "cheap" to produce (Luxton, 2022). Many artists and writers have faced significant fiscal and emotional impacts from the present wave of prompt-driven A.I. systems. This economic

reality is depressing for creative types who will no longer be able to eke out a living making art, which is already a famously precarious vocation. Even those whose creative pursuits are merely a hobby may discover that they find their joy and sense of mastery eroded.

However, the future may not be all doom and gloom for human creativity. Generative A.I. can potentially augment human creativity (Dwivedi et al., 2023), perhaps inspiring it while eliminating the requirements of more mundane tasks that burden creativity. This can allow people to explore and experience creative expression in beneficial ways.

But perhaps most troubling moving forward is the devaluing and loss of the human experience recorded and communicated through art since time immemorial. Art communicates the joy, suffering, passion, and beauty of human experience, moving us toward an enlightened state of being. The human race will lose this expression with the ubiquity of AI-generated art. AI-generated art is devoid of human experience and inspiration, and it is soulless, just like an AI-generated virtual human (Luxton, 2022). Why should we care about what is depicted in AI-generated art? The human response to AI-generated art may soon transfer from, "Wow, how did it do that?" to, "Who cares?"

2.9 Dependence on Machines for Problem-Solving and Decision Making

The use of artificial intelligence (A.I.) in healthcare has the potential to transform the way in which healthcare is delivered and experienced by patients. One area in which A.I. could have a significant impact is in the dynamic personal tracking of health, which refers to the use of technologies such as the Internet of Things (IoT) and non-contact means, such as voice or movement analysis, to continuously monitor and track patients' health (Hilty et al., 2021). Dynamic personal tracking of health has the potential to produce temporal health metrics, which are data points that are collected over time and can provide insight into patterns and trends in patients' health. This can help to identify potential health issues early on, and can enable healthcare providers to intervene more effectively to prevent or mitigate negative health outcomes. Dynamic personal tracking of health may also have implications for patient autonomy and privacy, as it may involve the collection and analysis of sensitive personal data. It is important to ensure that appropriate safeguards are in place to protect patients' privacy and to ensure that the data collected is used in a responsible and ethical manner.

The use of artificial intelligence (A.I.) in sciences has the potential to transform the way in which knowledge is generated and transferred. One key aspect of this is the shift towards a bottom-up paradigm of science, in which data is analyzed to identify patterns, rather than following the traditional approach of formulating a hypothesis and testing it. This shift towards a bottom-up paradigm is enabled by the increasing availability of large amounts of data, as well as the development of machine learning algorithms that are able to identify patterns and make predictions based on this data. One potential benefit of this approach is that it can allow scientists to identify patterns and relationships that may not have been evident using traditional methods, and can enable the discovery of new knowledge. However, there is also a risk that this approach may lead to a reliance on machines to find answers to problems rather than on human expertise and understanding. Furthermore, there is a risk that machines may be able to find answers to problems but may not necessarily be able to explain why these answers are correct. This can make it difficult for humans to understand and interpret the results and can limit the transferability of knowledge.

In the context of labor and business, the use of A.I. has the potential to both enhance and disrupt various aspects of the employment relationship. One potential impact of A.I. is

the rise of algorithmic management, which refers to the use of algorithms to monitor and evaluate employee performance, and to make decisions about tasks and responsibilities (Lee et al., 2015). Algorithmic management can lead to the emergence of petty bureaucracies in which employees are subject to an increasing level of micromanagement and control. This can have negative consequences for employee autonomy and dignity, as it may reduce the ability of employees to make their own decisions and exercise their own judgment.

Another potential impact of A.I. in work and business contexts is the return of *Taylorism*, which refers to a management philosophy based on the idea of breaking down work into smaller tasks and optimizing efficiency through the use of scientific principles. The use of A.I. in tasks such as nudging, which refers to the use of subtle cues or incentives to influence behavior, may lead to a return of this type of management philosophy. The use of A.I. in work and business contexts may also disrupt middle management, as it may lead to the automation of specific tasks and responsibilities that are currently performed by middle managers. This can have negative consequences for the employment prospects and job security of middle managers, who are typically tasked with personnel management but with little leeway for strategic decision-making. There may be a perception amongst upper management of the opportunity to squeeze further performance out of staff whilst reducing middle management headcount through the use of automated management systems (Roberts & Shaw, 2022).

2.10 Psychological Warfare

The defense industry has a long history of Artificial intelligence (A.I.) innovation for various purposes, from weapon guidance systems, cyberwarfare deterrence, command and control of unmanned vehicles, robots, and more. These technological advancements have also benefited other industries through technology transfer. The psychosocial and psychological effects of the use of these technological advancements on persons, however, including adversarial combatants, operators of these technologies, and collateral persons, must not be overlooked. In particular, A.I. can enable a range of risks to psychosecurity, which refers to the protection of individuals' mental health and well-being.

The psychological effects of the use of unmanned aerial vehicles (UAVs) are one example that has emerged in the last twenty years. Mass traumatic stress experienced by collateral citizens during U.S. military operations in Afghanistan, Iraq, and Pakistan has been reported (Litz, 2007). The inability to observe the presence of circling aerial drones and uncertainty about imminent threats exacerbates the experience of anxiety and reinforces psychological distress (Ullah, 2016).

Another emerging risk is the use of A.I. to conduct Automated Zersetzung attacks, which are designed to demoralize and undermine targeted individuals (Dennis & LaPorte, 2011). Zersetzung (German for "decomposition" or "corrosion") refers to a range of psychological warfare tactics that are designed to undermine the mental health and well-being of targeted individuals or groups. These tactics may include harassment, intimidation, manipulation, and other forms of psychological abuse and can be conducted through a variety of means, including social media, email, and other online platforms. The goal of Zersetzung tactics is to demoralize and undermine the targeted individuals or groups and to create a sense of uncertainty, vulnerability, and fear. These tactics can be used to interfere with the ability of the targeted individuals or groups to function effectively and to disrupt social and political movements or other forms of collective action. Zersetzung tactics have a long history of use by various states and non-state actors and have been documented in a range of contexts, including political repression, espionage, and counter-terrorism. In recent years, the increasing use of A.I. and other

technologies has enabled the automation and scaling of Zersetzung tactics, leading to concerns about the potential impacts on psychosecurity and social cohesion (Averkin et al., 2019).

The use of A.I. to conduct Automated Zersetzung attacks can have serious consequences for the targeted individuals, including negative impacts on mental health, social connections, and reputation. It can also undermine trust and cohesion within society, as it may lead to feelings of fear and vulnerability among individuals who may feel that they are at risk of being targeted. In the context of security and defense, the use of A.I. for Automated Zersetzung attacks can be particularly concerning, as it allows for the use of psychological warfare tactics in a plausibly deniable manner. This means that it can be difficult to trace the source of the attacks and to hold those responsible accountable, which can further undermine trust and stability within society.

4. Recommendations

There is a need for increased public awareness of the potential impacts of artificial intelligence (A.I.) on society, as well as guidance for governments and regulatory bodies, professional organizations, and individuals working in fields related to A.I., such as clinicians and researchers. One challenge is that A.I. is a rapidly evolving field, and it can be difficult for the public and even for experts to keep up with the latest developments and their potential impacts. And history informs us that new technologies are often rushed to market and deployed before all of the real-world risks are known. This can make it difficult for governments and regulatory bodies to develop appropriate policies and regulations to address potential risks and impacts. Professional healthcare organizations, such as the American Medical Association (AMA) and the American Psychological Association (APA), can play a role in providing guidance and resources to their members on how to responsibly use and interact with A.I. (Luxton, 2014b). This can include guidance on ethical considerations, best practices, and potential risks and impacts. Training is also an essential component in helping individuals working in fields related to A.I. to understand the potential implications of A.I. on society, as well as to develop the skills and knowledge to use A.I. responsibly. This can include training on ethical considerations, as well as training on technical skills related to A.I.

Public health research on the psychological impacts of artificial intelligence (A.I.) can help to inform our understanding of the potential risks and benefits of this technology and can inform the development of standards, certifications, and governance frameworks. There are several professional organizations that have developed standards and certifications related to A.I., such as the Institute of Electrical and Electronics Engineers (IEEE). These standards and certifications can help to ensure that A.I. is developed and used in a responsible and ethical manner and can help to build trust in this technology and its implementation. In addition to standards and certifications, effective governance frameworks for A.I. can help to build trust and confidence in this technology. One key element of effective governance is the inclusion of stakeholders from diverse backgrounds and perspectives, such as the community itself. This can help to ensure that the interests and concerns of all stakeholders are considered and can help to build trust in the governance process.

As we noted earlier in this paper, certain vulnerable groups and populations may be especially at risk for harm by disruptive A.I. Isolated and alienated youth, persons who've not had an opportunity to receive education about A.I. technology, persons who do not have access to these technologies, and persons with certain mental health conditions may be especially vulnerable to harm. It is, therefore, essential that the threats to these populations are adequately considered and research on this topic is supported. The

involvement of underrepresented persons and the most vulnerable in decision-making about developing and deploying A.I. technologies has been recommended (Luxton, 2020).

The increasing use of A.I. in our daily lives, primarily when A.I. mediates or supplants human relationships, is especially important, given how this may lead to changes in public morality and social norms. For example, the use of A.I. companions or friends may become more common, leading to situations where people may have dinner or engage in other social activities "alone" with A.I. entities. The development of these types of relationships with A.I. entities may have significant impacts on individuals' social and emotional well-being, as it may lead to a lack of connection and fulfillment in real-life relationships. It is also possible that these relationships could have an influence on people's beliefs and values, as it is often said that we become like the people with whom we spend significant time.

One of the most challenging threats to individuals, nation-states, and society we've discussed here is Automated Zersetzung attacks. The use of A.I. in fifth-generation plausibly deniable warfare can enable a range of tactical capabilities, such as the ability to conduct complex and coordinated military operations without the need for large, visible military forces. It can also enable the use of covert or indirect means to achieve strategic objectives, such as through the use of drones or cyber operations. One concern with the use of A.I. in fifth-generation plausibly deniable warfare is the potential for it to be used in ways that are difficult to trace or attribute to a specific state or actor. This can create uncertainty and instability within the international system, as it may be challenging to know the motivations and intentions of those responsible for such attacks. It can also make it difficult to hold those responsible accountable, undermining the rule of law and the international order. Investments in systems that can detect and thwart these attacks are needed.

Moving forward, intended or unintended manipulation of human behavior with A.I. will increasingly become a public health concern as recognition has become easier (Luxton, 2022). While behavioral influence and modification, such as with the use of virtual coaches and companions, may be adaptive and oriented to improving health and well-being, this same technology can be used to influence people in ways that cause harm. As we noted earlier, the use of social media bots to persuade and manipulate the public is already evident. The revelations from the release of the "Twitter Files" have undoubtedly damaged trust in governmental institutions that have a duty to protect the health and safety of the public. And while some persons may be most vulnerable and harmed by maladaptive or misuse of A.I. technologies, all of society can suffer from harm when A.I. is used by the powerful to achieve ends that are not in the public's best interest. Transparency, trust, and freedom from psychological manipulation are paramount in a healthy society, and we must strive to assure public trust in how A.I. technologies are used.

6. Final Considerations

There is a range of potential risks and opportunities for the future of artificial intelligence (A.I.) and society. One opportunity is that A.I. is used to enhance human welfare by improving healthcare, education, and other essential services. For example, A.I. could be used to analyze vast amounts of data to identify patterns and trends that professionals could use to improve public health or to develop personalized learning experiences that help individuals achieve their full potential. Another opportunity is that A.I. can be used to address global challenges, such as climate change, poverty, and inequality. For example, A.I. could analyze data and develop solutions to help mitigate climate change's impacts or identify and address the root causes of poverty and inequality. On the other

hand, a significant risk is that A.I. is used to harm or exploit individuals or society, either intentionally or unintentionally.

Based on historical events and analysis from an STS frame, it is likely that the reality of A.I. and society will be a blend of positive and negative effects on human well-being. To mitigate risks to public health, it is essential that a Geneva Conventions-style agreement prohibits the usage of demoralization tools against civilians as a crime against humanity. It is also very important that strong investment is made in improving the transparency and auditability of models, along with in-built encryption techniques such as models partly trained on-device, which better respect user privacy and security.

References

Averkin, A., Bazarkina, D., Pantserev, K., & Pashentsev, E. (2019). Artificial Intelligence in the Context of Psychological Security: Theoretical and Practical Implications. Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019). https://www.atlantis-press.com/proceedings/eusflat-19/125914786

Ayers, J. (2009). Coping with speech anxiety: The power of positive thinking. *Communication Education. 37*(4) 289-296. https://doi.org/10.1080/03634528809378730

Barrett, D. (2010). Supernormal Stimuli: How Primal Urges Overran Their Evolutionary Purpose. W. W. Norton & Company.

Bénabou, R. (December 2003). Inequality, Technology, and the Social Contract. Woodrow Wilson Economics Discussion Paper No. 226. https://doi.org/10.2139/ssrn.525043

Braghieri, L. Levy, R. & Makarin, A. (July 28, 2022). Social Media and Mental Health. Available at SSRN: https://ssrn.com/abstract=3919760 or http://dx.doi.org/10.2139/ssrn.3919760

Dean, J. (2019). The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design. https://arxiv.org/ftp/arxiv/papers/1911/1911.05289.pdf

Dennis, M. & LaPorte, N. (2011). "The Stasi and Operational Subversion". State and Minorities in Communist East Germany. Berghahn Books.

Due, B. L. (2015). The social construction of a Glasshole: Google Glass and multiactivity in social interaction. *PsychNology Journal., 13*(2-3), 149-178. http://www.psychnology.org/File/PNJ13%282-3%29/PSYCHNOLOGY_JOURNAL_13_2_DUE.pdf

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K.,… Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational A.I. for research, practice and policy, *International Journal of Information Management, 71*. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Feldstein, S. (April 18, 2019). Artificial Intelligence and Digital Repression: Global Challenges to Governance. http://dx.doi.org/10.2139/ssrn.3374575

Hargittai, E. (2011). The Digital Reproduction of Inequality. In D. Grusky & S. Szelenyi (Eds). The Inequality Reader: Contemporary and Foundational Readings in Race, Class, and Gender (2nd ed.). Routledge. https://doi.org/10.4324/9780429494468-69

Hilty, D. M., Armstrong, C. M., Luxton, D. D., Gentry, M., & Krupinski, E. (2021). A Scoping Review of Sensors, Wearables, and Remote Monitoring For Behavioral Health: Uses, Outcomes, Clinical Competencies, and Research Directions. Journal of Technology in Behavioral Science. 1-36. https://doi.org/10.1007/s41347-021-00199-2

Kaminski, M. E., & Witnov, S. (2015). The Conforming Effect: First Amendment Implications of Surveillance, Beyond Chilling Speech. *University of Richmond Law Review, 49*, 465-518. https://lawreview.richmond.edu/files/2015/01/Kaminski-492.pdf

Krøvel, R. & Thowsen M. (eds). (2019). Making Transparency Possible: An Interdisciplinary Dialogue. Cappelen Damm Akademisk/NOASP. https://doi.org/10.23865/noasp.64.ch16

Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015: 160-1612. https://doi.org/10.1145/2702123.2702548

Lepore, S. J., Ragan, J. D., & Jones, S. (2000). Talking facilitates cognitive–emotional processes of adaptation to an acute stressor. *Journal of Personality and Social Psychology, 78*(3), 499–508. https://doi.org/10.1037/0022-3514.78.3.499

Litz, B. T. (2007). Research on the Impact of Military Trauma: Current Status and Future Directions. *Military Psychology, 19*(3), 217-238. https://doi.org/10.1080/08995600701386358

Luxton, D. D. (2014a). Artificial Intelligence in Psychological Practice: Current and Future Applications and Implications. *Professional Psychology: Research & Practice. 45*(5), 332-339. https://doi.org/10.1037/a0034559

Luxton, D. D. (2014b). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine, 62*(1), 1-10. https://doi.org/10.1016/j.artmed.2014.06.004

Luxton, D. D. (2020). Ethical Challenges of Conversational Agents in Global Public Health. *Bulletin of the World Health Organization, 98*(4),285-287. doi: http://dx.doi.org/10.2471/BLT.19.237636

Luxton, D. D. (2022). Deepfake AI and Mass Virtual Reality Are a Public Health Risk: https://daviddluxton.substack.com/p/deepfake-ai-and-mass-virtual-reality

Luxton, D. D. June, J. D. & Fairall, J. M. (2012). Social Media and Suicide: A Public Health Perspective. *American Journal of Public Health, 102*(2), 195-200. doi: 10.2105/AJPH.2011.300608

Luxton, D. D. & Poulin, C. (2020). Advancing Public Health in the Age of Big Data: Methods, Ethics, and Recommendations. In L. Goldschmidt & R. M. Relova (Eds.). Patient-Centered Healthcare Technology: The Way to Better Health. IET. https://doi.org/10.1049/PBHE017E_ch3

Maitra, I. & McGowan M. K. (Eds.) (2012). Speech and Harm: Controversies Over Free Speech, Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199236282.001.0001

Massazza, A., Kienzler, H., Al-Mitwalli, S., Tamimi, N. & Giacaman, R. (2022). The association between uncertainty and mental health: a scoping review of the quantitative literature, *Journal of Mental Health, 32*(2), 480–491. https://doi.org/10.1080/09638237.2021.2022620

Manheim, K. & Kaplan, L. (2018). Artificial Intelligence: Risks to Privacy and Democracy. *Yale Journal of Law and Technology, 21*, 106.

Matthews, J. (2021). Keynote: Surveillance, Power and Accountable A.I. Systems: Can We Craft an A.I. Future that Works for Everyone?, 2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, 2021, pp. 5-5. https://doi.org/10.1109/icedeg52154.2021.9530871

McKay, C. (2020). Predicting risk in criminal procedure: actuarial tools, algorithms, A.I. and judicial decision-making,

Current Issues in Criminal Justice, 32(1), 22-39. https://doi.org/10.1080/10345329.2019.1658694

Mirza, M. U., Richter, A., van Nes, E. H, & Scheffer, M. (2019). Technology driven inequality leads to poverty and resource depletion. *Ecological Economics, 160*, 215-226. https://doi.org/10.1016/j.ecolecon.2019.02.015

Niles, A. N., Craske, M. G., Lieberman, M. D. & Hur. C. (2015). Affect labeling enhances exposure effectiveness for public speaking anxiety. *Behaviour Research and Therapy, 68*, 27-36. https://doi.org/10.1016/j.brat.2015.03.004

Nonan, M. (2013). I, Glasshole: My Year With Google Glass. *Wired*. https://www.wired.com/2013/12/glasshole

O'Connor A. J. & Jahan F. (2014). Under Surveillance and Overwrought: American Muslims' Emotional and Behavioral Responses to Government Surveillance. *Journal of Muslim Mental Health. 8*(1), 95-106. https://doi.org/10.3998/JMMH.10381607.0008.106

Parekh, L. B. (2017). Limits of Free Speech. *Philosophia 45*, 931–935. https://doi.org/10.1007/s11406-016-9752-5

Peachey, K. (March, 2022). Post Office scandal: What the Horizon saga is all about. *BBC News*. https://www.bbc.co.uk/news/business-56718036

Roberts, J. & Shaw, K. L. (2022). Managers and the Management of Organizations. NBER Working Paper Series. National Bureau of Economic Research. https://www.nber.org/papers/w30730

Sardarizadeh, S. & Schraer, R. (2022). BBC News. Twitter Files spark debate about 'blacklisting'. https://www.bbc.com/news/technology-63963779

Schou Andreassen, C., Billieux, J., Griffiths, M. D., Kuss, D. J., Demetrovics, Z., Mazzoni, E., & Pallesen, S. (2016). The relationship between addictive use of social media and video games and symptoms of psychiatric disorders: a large-scale cross-sectional study. *Psychology of Addictive Behaviors, 30*(2), 252. https://doi.org/10.1037/adb0000160

Schuster, D., (2014, July 14). The revolt against Google 'Glassholes'. *New York Post*. https://nypost.com/2014/07/14/is-google-glass-cool-or-just-plain-creepy/

Sedgwick, R., Epstein, S., Dutta, R., & Ougrin, D. (2019). Social media, internet use and suicide attempts in adolescents. *Current Opinion in Psychiatry 32*(6), 534-541. doi: 10.1097/YCO.0000000000000547

Starr, A., Fernandez, L.A., Amster, R., Wood, L. J., & Caro, M. J. (2008). The Impacts of State Surveillance on Political Assembly and Association: A Socio-Legal Analysis. *Qualitative Sociology, 31*, 251–270. https://link.springer.com/article/10.1007/s11133-008-9107-z

The White House (2022). The Impact Of Artificial Intelligence on The Future of Workforces in the European Union and the United States of America. https://www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf

Teo, A. R., & Gaw, A. C. (2010). Hikikomori, a Japanese culture-bound syndrome of social withdrawal?: A proposal for DSM-5. *The Journal of Nervous and Mental Disease*, *198*(6), 444–449. https://doi.org/10.1097/NMD.0b013e3181e086b1

Ullah, M. (2016). Impact of Drone Attacks Anxiety on Students at Secondary Level in North Waziristan Agency. *Public Policy and Administration Research*, *6*, 1-7. https://www.iiste.org/Journals/index.php/PPAR/article/view/28550

U.S. House of Representatives (March 2023). Hearing on the Weaponization of the Federal Government on the Twitter Files. Available at: https://judiciary.house.gov/committee-activity/hearings/hearing-weaponization-federal-government-twitter-files

Van Dijk, J. A. (2006). Digital divide research, achievements and shortcomings. *Poetics, 34*(4-5), 221-235.

Verma, P. (2022). Humans vs. Robots: The Battle Reaches a 'Turning Point'. The Washington Post. https://www.washingtonpost.com/technology/2022/12/10/warehouse-robots-amazon-sparrow

Wee, A. & Findlay, M. J. (September 14, 2020). A.I. and Data Use: Surveillance Technology and Community Disquiet in the Age of COVID-19. SMU Centre for A.I. & Data Governance Research Paper No. 2020/10. http://dx.doi.org/10.2139/ssrn.3715993

Wojcieszak, M. (2010). 'Don't talk to me': effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*, *12*(4), 637-55. https://doi.org/10.1177/1461444809342775