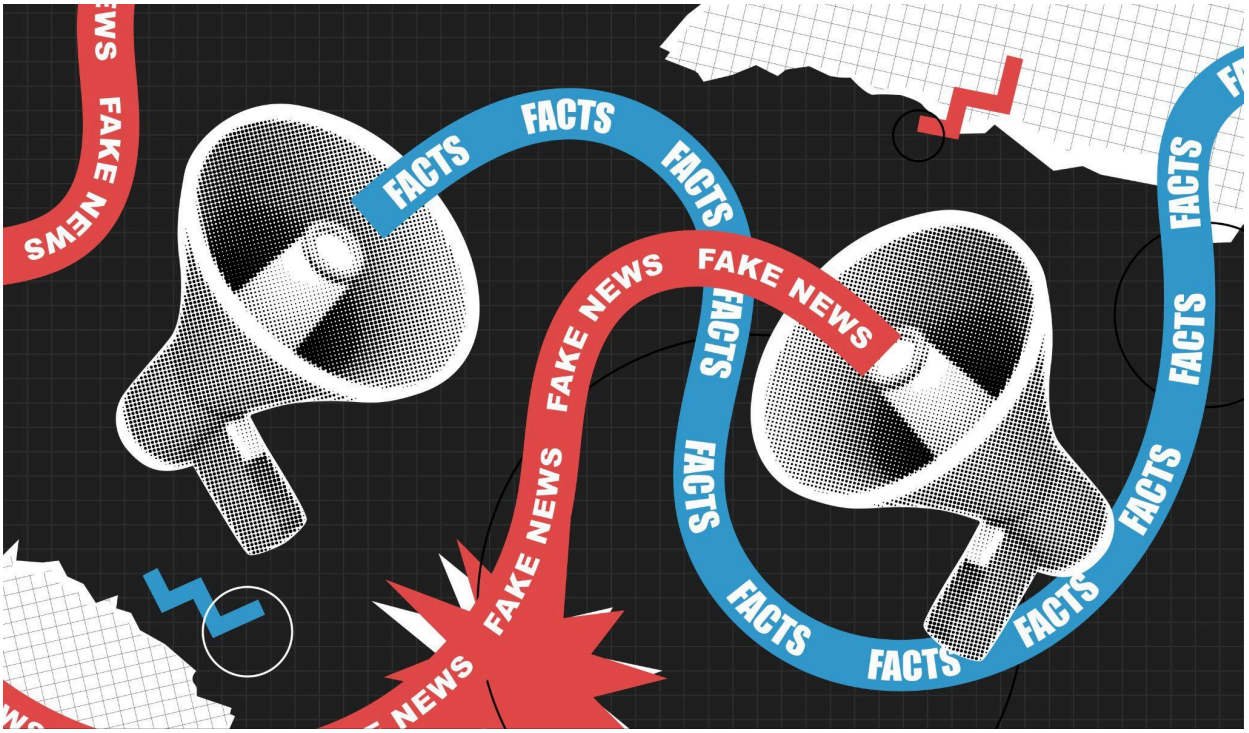
December 2024

Countering Disinformation
With a focus on Fact-Checking and AI



Authors

Etienne Gerster, Student, FHNW

In collaboration with

Prof. Dr. Barbara Eisenbart, Institute of Management, FHNW

Sandrine Denti, digitalswitzerland Foundation

Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Studierendenprojekt

digitalswitzerland FOUNDATION

About

FHNW

FHNW is one of Switzerland's leading universities of applied sciences and arts, actively involved in teaching, research, continuing education and service provision – both innovative and practice-oriented. Its broad range of degree programmes, hands-on concept, innovative, application-oriented research and global network make FHNW a diversified and appealing educational institution, a sought-after partner to industry and an attractive employer in northwestern Switzerland.

digitalswitzerland Foundation

The digitalswitzerland Foundation is a neutral entity that is part of the digitalswitzerland ecosystem aiming to transform Switzerland into a leading digital nation. The Foundation offers an open platform for multi-stakeholder dialogue to address challenging and current digitalisation topics, and to co-create tangible outcomes and innovative solutions by building bridges and breaking silos.

Acknowledgements

I would like to express my sincere gratitude to all those who supported me during the completion of this research paper. A special thanks goes to Prof. Dr. Barbara Eisenbart for her invaluable guidance, and to Sandrine Denti for her insightful and far-sighted perspective, which greatly enriched this work. I am also deeply appreciative of the expert input provided by representatives from the following organisations: Digitale Gesellschaft, AI Business School, Zentrum für Demokratie Aarau, Initiative for Media Innovation (IMI) EPFL, FHNW Digital Trust Center, Swico, Forschungszentrum Informatik, OFCOM, SRF, and others whose contributions were crucial in shaping the outcome of the paper. Lastly, my heartfelt thanks to my family and friends, colleagues, and partner for their unwavering support and encouragement throughout this journey.

Abstract

Summary

This paper delves into the challenges and potentials of leveraging artificial intelligence (AI) and automated fact-checking procedures to counter disinformation. The focus is on how Switzerland can exploit its digital ecosystem and AI expertise to combat disinformation efficiently. This research paper employs a mixed-methods approach, including a comprehensive literature review, expert interviews, and an interactive Digital Xchange Event. Particular emphasis is placed on examining the legal frameworks, societal implications, and existing and planned measures to counter disinformation, including through AI.

Findings

The paper emphasises the importance of a holistic approach, with preventive measures such as training in disinformation detection combined with active disinformation detection by AI, with human oversight playing an essential role. The study identifies successful international approaches to combating disinformation and derives actionable recommendations for Switzerland. A key finding is the necessity for a combined solution involving public education, active detection and correction of disinformation, as well as regulation and enhanced collaboration between government agencies, the private sector, and civil society to maximise the effectiveness of AI in combating disinformation.

Business and Social Implications

The use of explainable and transparent AI solutions to combat disinformation has the potential to strengthen public trust in digital media and significantly improve the quality of public information. This research paves the way for long-term positive economic and social effects by promoting digital literacy and implementing transparent and explainable AI systems.

Summarised recommendations for Switzerland

Recommendations	Description
Enhance multi-stakeholder cooperation and adopt a unified approach to counter disinformation	• Organise regular exchanges among academia, NGOs, civil society, media, private tech companies, government entities as well as educational institutions to **address disinformation and develop joint strategies** for a unified and coordinated response. • Promote and support **collaborative research projects** between universities and private companies to develop innovative solutions to counter disinformation. • Work with policy and decision-makers to create frameworks conditions and incentives for **cross-sector collaborations** and anti-disinformation projects.
Promote Digital Literacy and Education	• Integrate **digital and media literacy** into school curriculums. • Provide **training programs** and resources for teachers. • Develop **interactive learning tools** like educational software, games, and apps for different target audiences. • Organise **campaigns**, **workshops, seminars,** and **online courses** for the broader public. • **Partner with community centres**, **libraries**, and **civil society organisations** for in-person workshops. • Develop **user-friendly online learning platforms** with various formats.
Invest in Explainable AI	• Build **AI systems that provide transparent and clear explanations** for their outputs and decisions to **build trust**. • **Partner with research and science** for **insights** into developer training and **database** utilisation. • Consider **implementing the DeFakts project** to **save time and resources.**
Leverage International Insights and Best Practices	• Participate in **global forums** and **international collaborative efforts** to stay informed of the latest developments and best-practices in the fight against disinformation. • Learn from **successful strategies from neighbouring countries** and assess their relevance for local initiatives.

Table of Contents

Glossary

Term	Definitions	Source
Disinformation	The deliberate creation and dissemination of false information intending to deceive. This type of false information is particularly malicious, crafted to mislead the audience. Disinformation is often presented in a journalistic format, mimicking legitimate news to gain credibility and effectively deceive the public.	(Egelhofer and Lecheler, 2019)
Misinformation	The spread of false information without the intent to deceive. Misinformation can occur by unintentionally sharing inaccurate news, which can still have significant consequences. People might share incorrect information, believing it to be accurate, and thereby contribute to the spread of falsehoods.	(Egelhofer and Lecheler, 2019)
Propaganda	The deliberate, systematic attempt to shape perceptions and manipulate thinking to achieve a desired response. Propagandists aim to influence public opinion by controlling the flow of information and often present distorted information from what appears to be a credible source. Propaganda can include both true and false information and is typically used by state and non-state actors to further specific agendas.	(Egelhofer and Lecheler, 2019)
Artificial Intelligence	The process of automating tasks that typically require human intelligence, involving replicating diverse facets of human thinking and behaviour. AI systems strive to replicate human reasoning, learning, planning, creativity, and numerous other capabilities, often using algorithms crafted to accomplish specific objectives.	(Santos, 2023)
Machine Learning	A subset of AI that allows systems to automatically learn and improve from experience without being explicitly programmed. It involves algorithms that identify patterns in data, enabling predictions and decision-making based on past experiences. ML can be categorised into supervised learning, unsupervised learning, and reinforcement learning.	(Santos, 2023)
Deep Learning	A specialised subset of ML that uses complex algorithms and deep neural networks to train models. It uses multiple layers of processing units to analyse vast amounts of data, learning representations at multiple levels of abstraction. This approach is efficient for tasks such as image and speech recognition.	(Santos, 2023)
Fact-Checking	A structured and thorough approach to verifying the accuracy of statements and information. It involves identifying relevant statements, conducting thorough research using reliable sources, dual-source verification, contextual analysis, internal review, maintaining transparency, feedback mechanisms, and ongoing training for the fact-checking team to ensure accuracy and effectiveness.	(Baker and Fairbank, 2022, FactCheck.org, 2020, Shahzad et al., 2022, and SRF, 2022)
Prebunking	A preventative approach that involves familiarising people with the tactics and methods of disinformation dissemination before they encounter it. This strategy aims to raise awareness, digital literacy, and media literacy by educating audiences about how disinformation works and giving them tools to recognize and resist it.	(Lewandowsky and Van Der Linden, 2021)

Debunking	A reactive process of correcting disinformation after it has already been disseminated. It involves identifying the false claim, providing evidence-based corrections, and explaining why the original information is incorrect, aiming to mitigate the damage caused by disinformation by setting the record straight.	(Humprecht, 2019)
Algorithmic Bias	Systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group over others. This bias can also result from the programmers' own biases, which may unintentionally be incorporated into the system.	(Ünver, 2023)
Human Oversight	The process of involving human judgement and intervention in the development, deployment, and monitoring of AI systems to ensure they operate correctly and ethically.	(Demartini et al., 2020)
Explainable AI	AI that provides understandable and interpretable explanations for its decisions and actions, making it possible for humans to comprehend and trust the system.	(Ünver, 2023)

1. Introduction

The advent of artificial intelligence (AI) has ushered in a transformative era where automated systems increasingly dominate conversations, technological advancements, and actions. AI's potential is widely acknowledged, promising unprecedented efficiency and innovation across multiple domains (Kertysova, 2018). However, this rapid progression also presents significant challenges, particularly in information integrity. A notable concern is the escalating trend of misinformation and disinformation that threatens to undermine the very fabric of democracy and society (Kertysova, 2018).

In Switzerland, as in many parts of the world, there is a growing reliance on digital platforms for news consumption, especially among younger populations. Social media and online channels are becoming the primary sources of information, which amplifies the challenges associated with distinguishing between authentic news and fake news (Bundesamt für Statistik, 2022a). This shift has heightened the need for effective mechanisms to manage and mitigate the risks of false information, ensuring the public has access to reliable and accurate data (Bundesamt für Kommunikation, 2021).

In response to the increasing opportunities and challenges of AI, such as the spread of disinformation and effective measures to combat it, this research paper aims to explore how AI can effectively combat disinformation. Through an in-depth analysis of current global AI practices and potential future global innovations, it aims to provide valuable insights into Switzerland's initiatives. The findings and recommendations presented will help develop a more informed and strategic approach to integrating AI in the fight against disinformation and support Switzerland's approach to promote a resilient and informed digital society.

1.1 Problem Statement

In Switzerland, the rapid spread of digital technologies has led to a significant increase in disinformation, particularly on social media platforms. Widespread internet usage makes the Swiss population more vulnerable to disinformation and hate speech (Bundesamt für Statistik, 2023). Switzerland lacks legal frameworks to counter disinformation effectively and lags behind other countries in disinformation detection and the use of AI to combat disinformation (Thouvenin et al., 2024). This situation presents a substantial challenge for Switzerland in managing the

growing issue of disinformation. The problem tree diagram in Figure 1 highlights the primary focus areas for addressing disinformation. The root issue is disinformation, which branches into three main categories: disinformation creators, disinformation spreaders, and existing disinformation. This thesis concentrates specifically on the problem of 'existing disinformation', with a focus on 'disinformation detection' and 'avoiding the spread of disinformation'. These areas have been explored through a comprehensive literature review of global AI approaches, along with expert interviews and a Digital Xchange Event to gather insights from experts and address the specific needs of the Swiss population.

1.2 Objectives and Research Questions

This report investigates how Switzerland can effectively counter and combat disinformation by fostering collaboration within its ecosystem and by leveraging AI, with an emphasis on fact-checking. The report aims to answer the following two research questions in three phases:

- RQ1: What insights can Switzerland gain from European countries and the USA on addressing disinformation, particularly regarding the use of AI and fact-checking techniques?
- RQ2: How can Switzerland utilise AI practices to enhance fact-checking and counter disinformation effectively?

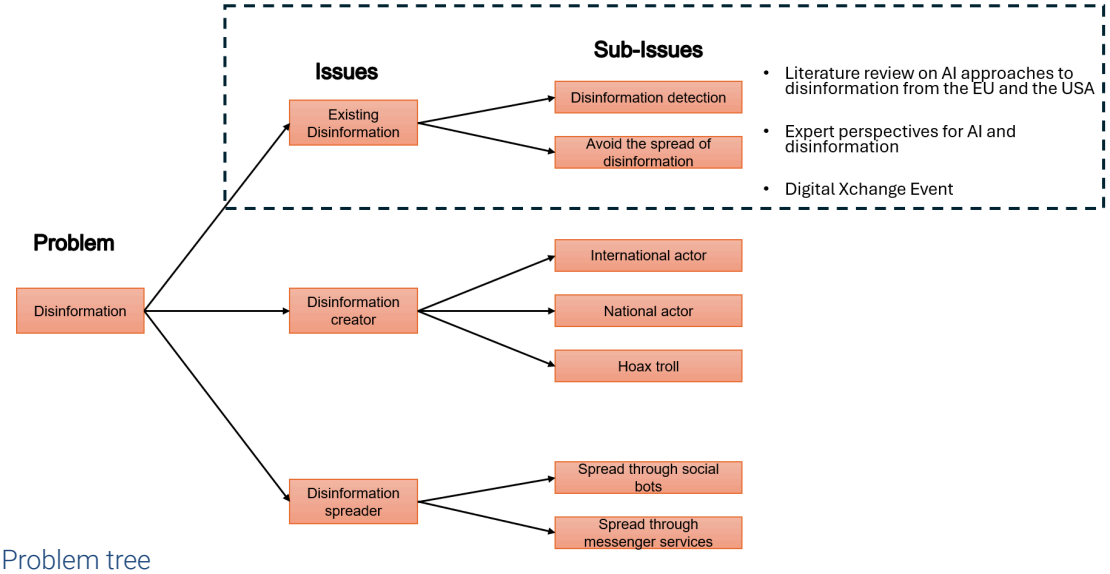


Figure 1: Problem tree
Source: Etienne Gerster

2. Results

2.1 Insights from Literature Review

This section presents a comprehensive analysis of existing literature on the regulation of AI and disinformation, focusing on the legal frameworks, the absence of AI-powered fact-checking platforms in Switzerland, public trust in AI, education, awareness-raising against disinformation, the use of AI methods abroad for information verification, and the role of transparent AI techniques in enhancing trustworthiness and reducing disinformation.

Legal Framework and Regulations

Disinformation

The Swiss legal framework does not include specific laws addressing disinformation, which presents challenges for regulation. Instead, Switzerland relies heavily on self-regulation by social media platforms and other intermediaries to address disinformation (Thouvenin et al., 2024). In addressing the current legal framework for disinformation governance in Switzerland, critical points from Thouvenin et al. (2024) highlight the legal challenges and considerations related to the governance of disinformation in digitalised publics. The focus is on creating effective legal measures while balancing freedom of expression and public discourse, mainly since digital platforms significantly shape public opinion (Thouvenin et al., 2024). The legal framework for disinformation governance in Switzerland is multifaceted and evolving to address challenges posed by digital platforms in areas such as content moderation, disinformation, and hate speech. This framework, which balances the need for regulation and the protection of fundamental rights such as freedom of expression and democratic participation, is at the heart of the Swiss approach to combating disinformation. According to the Federal Office of Communications (OFCOM) (2021), digital platforms have become key players in information dissemination, influencing public opinion and discourse; they face scrutiny for their role in spreading disinformation and their content moderation practices. These practices are often criticised, as they can influence the spread of disinformation and hate speech. OFCOM (2021) discusses the impact of intermediaries on public communication and proposes governance approaches to improve transparency and accountability, including systematic

monitoring of disinformation, similar to that in the EU. These approaches ensure that digital platforms act responsibly while respecting the rights and freedoms of users, and they aim to create a more structured and proactive approach to detecting and combating disinformation. In addition, Switzerland actively participates in international forums on AI and disinformation standards to ensure that its regulations align with global best practices and contribute to international efforts against disinformation.

Artificial intelligence

To provide a comprehensive approach to the regulation and advancement of AI, Switzerland has developed a series of strategic initiatives that form a cohesive framework for digitalisation and AI governance. The Digital Switzerland strategy 2018 sets out the objectives and guidelines for digitalisation across all critical areas of life, forming the foundational framework for AI regulation. The objectives of this strategy include promoting innovation, ensuring secure and reliable digital infrastructures, fostering digital skills among the population, and enhancing the public's trust in digital technologies. The guidelines highlight the importance of collaboration between public and private sectors, ethical standards, and inclusivity in digital transformation. The key areas the strategy addresses encompass education, healthcare, public services, and economic development. This foundational framework aims to create a robust environment for developing and implementing AI technologies, ensuring they align with Switzerland's broader digital goals and societal values (Digitale Schweiz, 2023).

Blarer Marcel Buffat et al. (2019) provide an overview of the framework conditions for AI use, highlighting challenges across various policy areas and discussing the need for federal adaptation. The framework conditions for AI use outlined in the report include the availability of large datasets, advancements in computational power, and the development of Machine Learning methods that enable AI systems to learn autonomously from data. These conditions are essential for the effective deployment and utilisation of AI technologies (Blarer et al., 2019). Additionally, Blarer Marcel Buffat et al. (2019) emphasise the need for a robust legal and ethical framework to ensure that AI is used responsibly and transparently while respecting human rights and avoiding discrimination. It identifies several challenges associated with AI use, including the lack of transparency and explainability of AI systems,

the potential for bias and discrimination, data privacy concerns, and the need for regulatory frameworks that can keep pace with rapid technological advancements. The policy areas affected by these challenges are diverse and encompass sectors such as healthcare, finance, cybersecurity, education, and public administration. Blarer Marcel Buffat et al. (2019) also discuss the need for federal adaptation, which refers to the need for the Swiss government to modify existing regulations and possibly introduce new ones to address the unique challenges posed by AI. This includes ensuring that AI systems are transparent and accountable to protect individuals from potential harm and foster public trust in AI technologies. Blarer Marcel Buffat et al. (2019) suggest that the federal government must continuously evaluate and update its policies to keep up with the evolving landscape of AI and its applications.

The Digital Switzerland strategy guidelines highlight the importance of collaboration between public and private sectors, ethical standards, and inclusivity in digital transformation.

The Federal Council (2021) offers a general orientation framework for federal administrative tasks and aims to ensure a coherent AI policy. These guidelines stem from the report of Blarer Marcel Buffat et al. (2019) and serve as a comprehensive reference for using AI within the federal administration, ensuring a unified policy across various sectors. Key objectives include developing sectoral AI strategies that align with federal policies, introducing or adapting regulations across all sectors affected by AI, and guiding the development and use of AI systems within the federal administration (Der Bundesrat, 2021). The Federal Council (2021) also emphasises Switzerland's role in shaping international AI governance and ensuring global standards reflect national values and interests. AI is seen as a key factor in digital transformation, offering significant potential for innovation and growth but posing challenges such as data-based discrimination, lack of transparency, and privacy concerns. To address these issues, the guidelines include putting people first, establishing favourable regulatory conditions, ensuring transparency and explainability in AI decisions, defining accountability, and ensuring that AI systems are safe and resilient. The Federal Council

(2021) also stresses the importance of Switzerland actively shaping global AI governance and involving all relevant stakeholders in political decision-making processes related to AI. This comprehensive approach balances technological advancements with ethical and legal considerations, fostering innovation while protecting societal values.

The Competence Network for AI (CNAI) 2021 advances digital transformation across federal administrations, promoting AI use and trust through communities of practice and expertise and ensuring transparency in federal AI projects. The further development of CNAI 2023-2025 initiative is set to evaluate the CNAI in 2024, serving as a contact point for AI. Measures for further development will be proposed as needed, based on an exchange of experience and knowledge (CNAI, 2022).

Lastly, the AI layout 2024-2025 task involves the Federal Department of the Environment, Transport, Energy and Communications identifying potential regulatory approaches for AI by the end of 2024; this involves all federal agencies responsible for relevant areas of law (Digitale Schweiz, 2023). This initiative is part of the digital Switzerland strategy and aims to provide a comprehensive framework for AI governance, ensuring that regulatory measures are up to date with technological advancements. The process includes a detailed examination of existing legal frameworks, consultations with various stakeholders, and alignment with international best practices. This effort is critical for maintaining Switzerland's position as a leader in AI innovation while safeguarding ethical standards and public trust (Digitale Schweiz, 2023).

These initiatives are part of Switzerland's broader strategy to regulate and advance AI, ensuring alignment with both national and international standards (Der Bundesrat, 2024). This strategy aims to address various challenges, including ethical considerations, transparency, and the societal impacts of AI. By fostering an informed and discerning public, Switzerland seeks to bolster resilience against disinformation and reinforce its democratic processes (Bieri et al., 2021). The legal framework, though still evolving, is designed to uphold high data protection and privacy standards while preventing discrimination and bias in automated processes. This continuous legislative adaptation is crucial for maintaining public trust and ensuring that AI technologies are developed and deployed responsibly.

Absence of AI-Powered Fact-Checking Platforms

Analyses of the literature and extensive web research revealed that Switzerland currently lacks AI-automated fact-checking solutions (Duke Reporters, 2023). This absence highlights a significant gap in the tools available to counter disinformation effectively using advanced technologies. Similarly, in both the US and Europe, there are no publicly accessible AI-automated fact-checking platforms available to the general population. This indicates a broader challenge in implementing and providing AI-driven solutions for disinformation detection and verification on a wide scale.

In Switzerland, the digitalisation initiative of the Zurich Higher Education Institutions (DIZH) is leading efforts in this area through ClarifAI, a notable project focusing on using AI to detect and highlight unseen patterns in media, thus making it easier to identify and counteract propaganda and disinformation (DIZH, 2024). This innovative approach leverages advanced AI technologies to enhance the accuracy and efficiency of disinformation detection, significantly contributing to digital resilience in Switzerland. The project runs from 2024 to 2025 and represents a pioneering effort within the country towards automated fact checking (DIZH, 2024). However, detailed information about ClarifAI is limited, reflecting Switzerland's nascent stage of AI-powered fact-checking initiatives.

In the German-speaking region, the Research Centre for Information Technology (FZI) in Germany is currently at the forefront with DeFakts, a leading research project that employs AI-driven technologies to detect and counter disinformation across various online communication channels (FZI Forschungszentrum Informatik, 2021). The project focuses on creating explainable AI (XAI) that detects disinformation and explains the reasoning behind its detection.

This initiative aims to integrate a user-friendly application programming interface (API) that alerts users to potentially deceptive content, thus enhancing media literacy and empowering individuals to critically evaluate the information they encounter. Once completed, DeFakts aims to make its tools publicly accessible, providing a robust resource for the population to combat disinformation (FZI Forschungszentrum Informatik, 2021).

Public Trust in AI, Disinformation and Regulation

The Mobiliar Digitalbarometer 2024 reveals a significant need for more trust among the Swiss population regarding the government's ability to appropriately regulate AI-based technologies. The survey indicates that 72% of respondents have low trust in the government's regulatory capabilities, while only 26% express high or remarkably high trust, with 2% unsure (Ramp et al., 2024).

This scepticism is particularly alarming given that the Swiss population generally holds a high level of trust in their government (Ramp et al., 2024). The need for more confidence in AI regulation could be attributed to the inherent uncertainties associated with new risks compared to known ones. Additionally, the rapid advancement of AI technologies may outpace the slow-moving legislative processes in Switzerland, further exacerbating public concerns. As Switzerland prepares to draft its own AI regulatory framework by the end of 2024, following the European Union's enactment of a comprehensive AI act in March 2024, the public's trust in effective and balanced regulation becomes crucial. Ensuring that AI technologies are regulated with adequate consideration of both opportunities and risks is essential for fostering a secure and ethical AI landscape (Ramp et al., 2024).

In light of these findings, recent surveys by the digital society initiative at the University of Zurich reveal a nuanced public sentiment towards AI in Switzerland. The findings indicate a significant increase in scepticism, particularly regarding AI's involvement in critical decision-making processes and its potential misuse in spreading disinformation. Despite these reservations, there is strong interest in leveraging AI's practical benefits, provided robust human oversight and transparent regulatory frameworks are in place (Digital Society Initiative, 2024).

In the realm of disinformation, the situation is different. A survey conducted by Vogler et al. (2021) indicates that the public expects the government, platform operators, and media companies to actively combat disinformation and prevent its spread. This expectation underscores these entities' critical role in maintaining information integrity and public trust. In addition, the findings reveal that the public is aware of the increasing prevalence of disinformation and its potential impact on society. Vogler et al. (2021) emphasise that effective regulation and proactive measures are essential to address this growing

concern, particularly in an election year when the integrity of information is paramount.

Education and Awareness

Education is seen as a proven preventive measure against disinformation, also known as prebunking (Lewandowsky & Van Der Linden, 2021). According to Bieri et al. (2021), Switzerland significantly emphasises educational initiatives to enhance digital literacy and critical thinking skills among its citizens, which is crucial for combating disinformation. Integrating media literacy education into school curricula, providing training for educators, and conducting public awareness campaigns are essential strategies. These measures aim to empower citizens to discern and reject disinformation, thereby reducing its impact on public opinion and democratic decision-making (Bieri et al., 2021). In addition to integrating digital literacy into school curricula, Switzerland focuses on continuing education for adults to ensure that all age groups have the necessary skills to safely navigate the digital landscape (Graf et al., 2022). These efforts include organising workshops, seminars, and online courses that address the specific needs of different demographics. Public libraries and community centres are vital in offering access to digital literacy resources and training programs. By fostering an informed and discerning public, Switzerland seeks to bolster resilience against disinformation and reinforce its democratic processes (Graf et al., 2022).

Integrating media literacy education into school curricula, providing training for educators, and conducting public awareness campaigns are essential strategies.

These educational efforts include partnerships with media organisations and technology companies to provide resources and tools that enhance digital literacy across various age groups. Collaborating with media outlets and tech companies can enhance the effectiveness of educational initiatives by leveraging their expertise and platforms to reach a broader audience. These partnerships also involve developing digital literacy campaigns that address emerging issues, such as deepfakes or algorithmic biases, thus helping keep the public informed about the latest challenges in the digital world (Graf et al., 2022). The role of continuous education is underscored by

initiatives such as the digital Switzerland program, which emphasises lifelong learning and the need for ongoing skill development to keep pace with technological advancements. This program includes targeted initiatives to improve digital literacy among older adults, ensuring that all citizens, regardless of age, can effectively engage with and understand digital content (Bieri et al., 2021).

Swiss educational efforts also highlight the importance of transparency and accountability in AI applications (Graf et al., 2022). Ensuring that AI systems are transparent and their operations are understandable to the public is vital in building trust and preventing misuse. According to Graf et al. (2022), legal frameworks and public policies are being developed to enhance transparency and prevent discrimination and manipulation through AI, further supporting educational initiatives to combat disinformation.

Use of AI Methods

According to Ünver (2023), AI methods are increasingly being employed in various countries to enhance information verification processes. The process of uncovering disinformation is known as debunking (Humprecht, 2019). These methods encompass a range of advanced technologies, including Natural language processing (NLP), ML, and blockchain. Automated fact-checking tools utilise NLP algorithms to analyse vast amounts of textual data in real time, identifying and verifying claims to counter misinformation effectively. These tools streamline fact-checking efforts by providing near-instant results, allowing fact checkers to keep pace with the rapid spread of information online.

Ünver (2023) shows that blockchain technology offers an additional layer of trust and transparency in information verification. By timestamping and securing verified information, blockchain ensures an immutable record of claims and their accuracy, thereby countering information manipulation and creating reliable sources of truth. Deepfake detection technologies, originally developed to combat the rise of manipulated multimedia content, have become integral to fact checking. These technologies help identify and debunk altered audio and video content, which are increasingly used to deceive audiences.

Despite the sophistication of AI tools, human oversight remains crucial. The HITL approach ensures that human experts supervise and

validate AI systems, thus enhancing the accuracy and reliability of automated fact-checking processes (Ünver, 2023). Human intervention is essential for training AI models, interpreting nuanced contexts, and making judgment calls that automated systems might miss. The integration of these technologies highlights the collaborative potential between AI and human fact checkers. AI systems assist by prioritising tasks, summarising information, and offering evidence from knowledge graphs, while human fact checkers provide the critical oversight necessary to navigate complex and context-dependent claims. This HITL synergy ensures a robust and dynamic approach to combating disinformation, fostering a more informed and resilient information ecosystem (Ünver, 2023).

Pioneering organisations like Full Fact in the UK and FactCheck.org in the US are at the forefront of integrating AI with traditional fact-checking methods (Ünver, 2023). Full Fact leverages ML to identify falsehoods in news reporting, while FactCheck.org employs data analytics to track the spread of misinformation on social media. Similarly, the Agence France-Presse uses digital forensics to scrutinise the authenticity of images and videos, combining AI-driven analysis with human expertise to maintain high standards of verification (Ünver, 2023). The integration of these technologies highlights the collaborative potential between AI and human fact checkers.

Despite the sophistication of AI tools, human oversight remains crucial. Human intervention is essential for training AI models, interpreting nuanced contexts, and making judgment calls that automated systems might miss.

Use Cases for Disinformation

Six analysed real-life use cases on the impact of disinformation have demonstrated that disinformation can lead to company financial losses, erosion of public trust in government institutions, and even the endangerment of public safety. Disinformation manifests, among other things, in the form of conspiracy theories, false reports, and propaganda. The analysis of real-world cases of disinformation reveals significant societal, political, and health impacts across different countries. In Switzerland, during the COVID-19 pandemic, conspiracy theories and false claims about the virus and vaccines spread via social media, particularly on WhatsApp. This led to

vaccine hesitancy and public protests against government measures such as mask mandates and lockdowns. In addition, trust in the government and institutions was severely damaged and weakened, leading to groupings and polarisation of people who were hostile to each other. The spread of disinformation undermined public health efforts and forced the Swiss government to consider new legislative measures to combat online disinformation (Swissinfo, 2020).

Disinformation can intensify during crises, amplifying fear and exacerbating societal tensions. This has been evident in various scenarios where false information spreads rapidly, capturing and reinforcing people's anxieties. Algorithmic biases can promote sensational and one-sided content, which often fuels disinformation. This was highlighted in numerous use cases where such biases led to real-world violence and societal disruptions.

While not directly mentioned in the provided scenarios, there is an implication that the effectiveness of disinformation could be mitigated with the help of AI solutions. Transparent AI techniques could play a crucial role in detecting and curbing the spread of false information. Credible and trustworthy data, supported by transparent AI, could enhance trust in institutions and governments. Although this was not explicitly stated, implementing effective disinformation detection strategies is an underlying benefit. If disinformation were effectively reduced, the adverse impacts observed in the analysed use cases would not have been as severe. This would lead to a more informed public and a less polarised society.

Conceptual Model on AI for Disinformation

According to Bontridder and Poullet (2021), Ünver (2023), Shahzad et al. (2022) and Iqbal et al. (2023), the effective use of AI in combating disinformation revolves around three primary challenges: ethical and societal considerations; algorithmic bias and missing transparency; and the impact of disinformation coupled with a lack of awareness. Addressing these challenges requires the deployment of transparent AI techniques that can significantly enhance the trustworthiness and credibility of AI systems and data and reduce the spread of disinformation.

Challenges

Ethical and societal considerations

The integration of AI into disinformation detection raises ethical concerns. These

concerns necessitate the development of robust ethical frameworks that guide the deployment of AI technologies, ensuring they are used responsibly and for the benefit of society. According to Bontridder & Poullet (2021), AI systems facilitate the creation and dissemination of disinformation and bring about multiple ethical and human rights concerns. These concerns include threats to human dignity, autonomy, democracy, and peace.

Algorithmic bias and missing transparency

AI systems are susceptible to biases arising from training data or algorithmic design. These biases can lead to unfair outcomes and diminish the effectiveness of disinformation detection. Ensuring transparency in AI algorithms is crucial to mitigate these biases (Ünver, 2023). This involves making the decision-making processes of AI systems more understandable and interpretable to users and stakeholders. Ünver (2023) highlights the importance of transparency in AI systems to address biases and improve the reliability of disinformation detection technologies.

Ensuring transparency in AI algorithms is crucial to mitigate these biases. This involves making the decision-making processes of AI systems more understandable and interpretable to users and stakeholders.

Impact of disinformation and missing awareness

The pervasive nature of disinformation can affect public opinion and societal trust (Iqbal et al., 2023). The public should be aware of the sophisticated techniques used to disseminate false information. Raising awareness about disinformation tactics and enhancing digital literacy are essential in empowering individuals to critically evaluate the information they encounter. This can be illustrated by the 'infodemic' during the COVID-19 pandemic, where the World Health Organisation pointed out the spread of excessive and false information, complicating public health responses (Bontridder & Poullet, 2021).

Solution

Transparent AI techniques for disinformation detection

According to Bontridder & Poullet (2021) and Ünver (2023), transparent AI techniques are a powerful tool in the fight against disinformation. One of these techniques is XAI, which ensures that AI systems provide clear and understandable explanations for their outputs. By making the decision-making process transparent, XAI helps identify the reasons behind disinformation detection, thereby fostering trust among users (Ünver, 2023). Human oversight helps mitigate biases and addresses the limitations of fully automated systems, ensuring more nuanced and context-aware analysis (Bontridder & Poullet, 2021).

Additionally, cross-lingual fact-checking utilises multilingual AI models to verify claims across different languages, helping to address the global nature of disinformation. This method ensures that disinformation campaigns do not exploit language barriers to spread false information widely (Ünver, 2023). Lastly, real-time detection and debunking involves implementing real-time monitoring systems to detect and debunk disinformation as it arises, preventing its rapid spread. These systems use APIs and web scraping tools to gather data and promptly alert users about potentially false claims (Bontridder & Poullet, 2021; Ünver, 2023).

Outcome

Increase of Trustworthiness and Credibility of AI and Data

By employing transparent AI techniques, the credibility and trustworthiness of AI systems and the data they analyse are significantly enhanced. This credibility is crucial for the acceptance and effectiveness of AI in disinformation detection. Trustworthy AI fosters public confidence and encourages broader adoption of these technologies in various sectors (Shahzad et al., 2022).

Reduce of Disinformation

Effective disinformation detection and mitigation strategies reduce the spread of false information. Transparent AI techniques play a crucial role by enabling precise and reliable identification of disinformation, thereby preventing its proliferation and minimising its impact on society (Shahzad et al., 2022; Ünver, 2023). Accurate detection and flagging of false content, along with public education, significantly mitigate the impact of disinformation. Ünver (2023)

underscores the importance of integrating AI with human oversight to ensure system effectiveness and reliability.

2.2 Insights from Interviews

The study includes interviews with ten organisations from the academic, the private/tech sector, the NGO/civil society sector, the media and the government. The diversity of the interviewees ensured a broad range of perspectives. Additionally, a summary of the most important results of the interviews and the workshop can be found in Table 3.

To further clarify the findings, the key themes and characteristics derived from the literature and supported by the interviews are summarised in Table 1. The table highlights the themes that were frequently mentioned across multiple interviews, as well as those that individual interviewees particularly emphasised. For instance, Educational Efforts under the theme Impact of Disinformation and Missing Awareness was mentioned by all the interviewees, highlighting its critical importance. Conversely, Trust in Democratic Processes was a unique point, only emphasised by a single interviewee, indicating a less widespread but still significant concern.

Own model based on literature	Based on interviews								Sorted by the highest total number		
Key Theme	**Characteristics**	**IN1**	**IN2**	**IN3**	**IN4**	**IN5**	**IN6**	**IN7**	**IN8**	**IN9**	**IN10**	**Total**
Impact of Disinformation and Missing Awareness	Educational Efforts	X	X	X	X	X	X	X	X	X	X	10
Reduce (spread) of Disinformation	Public Media Literacy	X	X	X	X	X	X	X	X	X	X	10
Impact of Disinformation and Missing Awareness	Defining Disinformation	X			X	X	X	X	X	X	X	8
Algorithmic Bias and Missing Transparency	Missing Transparency	X	X	X	X	X	X				X	7
Transparent AI Techniques for Disinformation Detection	Human Oversight	X		X		X		X	X	X	X	7
Increase of Trustworthiness and Credibility of AI and Data	Public Awareness Campaigns	X	X	X	X	X		X			X	7
Ethical and Societal Consideration	AI Ethics Solutions and Bias	X	X	X				X	X	X		6
Reduce (spread) of Disinformation	Transparency and Human-in-the-Loop	X		X			X	X		X	X	6
Increase of Trustworthiness and Credibility of AI and Data	Research	X		X		X		X			X	5
Algorithmic Bias and Missing Transparency	Training	X								X	X	3
Transparent AI Techniques for Disinformation Detection	Explainable AI Approaches			X				X		X		3
Ethical and Societal Consideration	Trust in Democratic Processes				X							1

Table 1: Key insights from interviews

Source: Etienne Gerster

Ethical and Societal Considerations

Many interviewees expressed concerns that AI ethics solutions might inadvertently create more bias and restrict freedom of expression. One interviewee said, 'The reality is that at the moment these fantastic AI ethics solutions are mostly used to censor and filter and unintentionally create more bias' (expert 1). The expert also highlighted the risk of powerful gatekeepers filtering information: 'The biggest systematic risk is that powerful gatekeepers filter information that they manipulate. This is about freedom of expression and freedom of opinion' (expert 1). Another expert spoke of the importance of involving the entire population in the ethical journey: 'I think it is crucial that we take people on this journey, from schoolchildren to pensioners, and in my view, it is crucial that we strengthen transparency in principle. So that people know what they are facing' (expert 4). This emphasis on education and transparency forms a core part of addressing the societal impacts of AI.

Enhancing media literacy and critical thinking from an early age is seen as crucial to addressing AI-related risks and ensuring the responsible use of AI in society. One expert said, 'I believe what's really important is the media literacy of the population and where I see a big problem is precisely in the AI creation of visual content […] That's one thing, but the other is that the credibility of images generally suffers because ultimately everything can be a fake' (expert 2). Another expert supported this by stating, 'Therefore, I always believe that we should not seek technological solutions for human problems […] We should educate people in critical thinking from an early age' (expert 5). Another interviewee noted the importance of adult education: 'There also needs to be offers in adult education and adult education centres, perhaps presentations also in places where older people come into contact with digitality anyway, that there are still' (expert 3).

Maintaining high levels of trust in democratic processes and institutions is essential, especially in direct democracies like Switzerland. One interviewee stated, 'From a Swiss perspective, what characterises us is an extremely high level, relative to other countries, for example, of trust in processes and institutions. In my view, it is crucial [...] that we can maintain trust in these institutions and processes, and disinformation is aimed at eliminating this trust' (expert 4). This trust is linked to the transparency and credibility of the

information shared with the public. The challenge of balancing security and freedom of expression was another recurring theme. One expert noted, 'Disinformation poses a complex challenge, requiring careful regulation that balances freedom of expression with societal safety and democratic integrity' (expert 6). This sentiment was echoed by another interviewee, who stated, 'We realise that when the federal ministry posted on Twitter about our project, there was a lot of negative feedback from people who felt like they would be censored' (expert 9).

There is a consensus on the need for ethical guidelines and transparency in AI technologies. One interviewee stated, 'I think it's very important that such ethical guidelines exist […] On the one hand, it provides answers as to how employees should deal with it. But it also makes it transparent that we as a company deal responsibly with the technologies' (expert 2). Another interviewee supported this view, stating, 'As with everything, we should first look for the solution in people and then perhaps technology can support us where it is useful or not' (expert 5).

Raising awareness about AI-related risks and educating the public is seen as a fundamental solution. One interviewee said, 'We should train people as we did in the past in the physical world about physical dangers, be it fire, be it criminality etcetera, and make people understand that there are also digital risks, cyber risks, AI-related risks' (expert 1). Another stated, 'On the one hand, I think we need to further strengthen our education system in the sense that people need to improve their understanding of digital technologies, but also that young people in particular, who are also extremely exposed to disinformation, are better able to categorise media information, for example, or information' (expert 4). Another expert stated, 'The crucial question is always, Does it have an effect or does it not have an effect? And when we talk more about potentials than we really talk about effects, that also means for us that we now want to look much more at this effect side and at the same time want to see, for example, how strongly artificially generated content has the same effect or not, whether you can recognise it or not' (expert 7).

While there is general agreement on the importance of ethical considerations and transparency, some interviewees expressed differing views on how to implement these solutions. For instance, one interviewee

suggested a balanced approach between regulation and freedom of expression: 'We need a balanced approach between risk-based and harm-based regulation, where applications that are unproblematic can be used freely, and those that are fundamentally incompatible with our basic rights, like social scoring or biometric surveillance in public places, are banned' (expert 6). However, another interviewee raised concerns about the potential overreach of such regulations: 'I therefore warn against regulating everywhere, and I strongly caution against regulating any technology [...] In my view, we don't need AI regulation with regard to disinformation, but we need to define as a society how we want to protect our system and our trust in institutional processes' (expert 3).

Algorithmic Bias and Missing Transparency

One of the key issues raised was the use of AI for algorithmic censoring and filtering, which can lead to unintended biases. An interviewee stated, 'AI is mostly used today to do algorithmic censoring and filtering. That's the reality' (expert 1). This points to a broader concern about the inherent risks involved in AI technology, as another interviewee stated: 'Then there's the big issue here, which biases also come into play, because AI technology development is not simply broadly supported' (expert 3).

A critical aspect that emerged was the need for transparency in AI development. An interviewee questioned the training data used for AI, noting that it often reflects societal biases: 'How was this tool trained? It generally works well when searching for Caucasian males; it becomes more challenging with women of colour; you can tell it was trained differently' (expert 2). This necessity for transparency was a recurring theme, highlighting challenges in accessing and understanding proprietary algorithms. As one interviewee pointed out, 'We will never get the algorithm from Google because it is a trade secret. Even if we did get it, we wouldn't be able to understand its complexity. This means it remains a black box' (expert 7). Another noted, 'I think it is crucial that we take people on this journey, from schoolchildren to pensioners. And in my view, it is crucial that we strengthen transparency, which is fundamental' (expert 4).

Efforts to mitigate bias include comprehensive training and implementing explainable AI approaches. One interviewee mentioned, 'The programmers also let their bias flow into the algorithms. They all have biases, and that's why I always believe that

we should not seek technological solutions for human problems' (expert 5). Another stated that the question is 'whether we can distinguish an algorithmic bias from a social bias. Because we know that we have that; it's clear that women are treated differently and that they are portrayed differently. But the question is, Is there a surplus, an algorithmic surplus, that also plays a role here?' (expert 7).

Despite these efforts, challenges remain. An interviewee pointed out the difficulty in regulating AI effectively due to rapid advancements: 'Regulation is a major challenge because the field of AI develops so quickly. By the time a regulation is established, the field might have advanced significantly or adapted dynamically' (expert 8).

Contrasting views were also present. Some interviewees believed that bias could not be entirely eliminated. One expert remarked, 'I don't think we can fully exclude bias at any time. Once you train a model, you have some sort of human input' (expert 9). An interviewee from a different sector discussed the complexity of balancing transparency and practicality: 'I believe it is indeed more a question of algorithms, distribution, and data that we generally lack because we lack the interfaces, be it the interfaces to Facebook, TikTok, X, and all platforms' (expert 6).

Additional contrasting opinions emerged regarding the role of education and regulation. One interviewee suggested that a focus on education and awareness is critical: '[W]e should not look for the solution in a technology that replaces our thinking, but we should look for the solution primarily in people by training people, i.e., first of all ethical guidelines, i.e., first of all people should of course have values and not want to manipulate others, but that is also the reality. We can't prevent that. Secondly, we should train all people to deal with all information and with themselves in a reflective and critical way. And these are skills that all people need, and then the danger of disinformation is not so great. In other words, the only solution to AI and disinformation is to educate people and then use AI because we can't stop it' (expert 5). In contrast, another interviewee emphasised the importance of combining technological solutions with human oversight: 'So we need not just an AI approach, but an explainable AI approach so that it really provides users with more information and hopefully thereby also kind of eventually tries to make them more independent of such tools because they will

be provided with information that helps them critically reflect on their own' (expert 9).

Impact of Disinformation and Awareness

One of the key concerns raised was the potential harm to freedom of expression. An interviewee stated, 'If I say you know freedom of expression is super important, I don't want information to be filtered. I want people to get the chance to get access to all kinds of information as long as it's not a clear lie, a proven lie, etcetera' (expert 1). This concern is closely tied to the erosion of trust caused by disinformation, as another interviewee highlighted: 'Ultimately, I understand disinformation as false information being spread through wrong channels with the intention to cause harm, thus usually aiming to damage the trust in an organisation' (expert 4).

To address these issues, the need for enhanced digital literacy and awareness was also emphasised. All the experts interviewed considered educational endeavours to be crucial. One interviewee stressed the importance of fostering critical evaluation skills: 'It's also important that the population critically questions content, and I think that needs to be communicated more—that you can't just believe everything that circulates on the net' (expert 2). Another interviewee highlighted the role of education in this context: 'The debate must be pursued, and it needs continuity and recurring formats. And on the other hand, we certainly also have the question of what the education system needs to do so that young users, children, adults, [...] are also led to start asking questions or demanding transparency' (expert 3).

However, there were differing views regarding the understanding of disinformation. One interviewee stated, 'I find the term "disinformation" itself dangerous because it implies that someone or a group of people defines what correct information is and what false information is. And that itself is an abuse of power' (expert 5). Another interviewee pointed out the complexity of defining disinformation in general: 'There are three different possibilities. You can say the

first option: Disinformation is when information that is shared can be proven to be wrong, a lie, and often combined with a bad intention, a negative intention [...] The second way to look at it is to say that disinformation is also censored if information is filtered and not displayed. Reduced information is also disinformation [...] The third option: Disinformation also contains information that I don't like. It doesn't need to be wrong, but I don't like it' (expert 1) This complexity was confirmed by another expert, who also pointed out the manipulative effect of disinformation: 'So disinformation can generally have the most drastic effects. Disinformation as a whole is manipulative [...] The greatest danger is that it is manipulative; in whatever direction, it depends on the information, on the context, and on the objective of the disinformation. It is always intended to manipulate, to form an opinion in [...] a direction that someone who spreads it, someone or something that spreads it, is aiming for' (expert 5).

Transparent AI Techniques for Disinformation Detection

One key aspect mentioned was the importance of explainable AI approaches. An interviewee noted, 'So it's not a binary classification that the user receives, but it gives more background information on the different indicators that the AI itself detects as standing for disinformation. It gives the users more background information on that so the user can easily understand why the AI actually came up with this decision. And it's also what you said to provide more transparency about the whole system, the model, and to generate trust thereby, but also to kind of generate a better understanding among the users for the concept of disinformation beyond this tool that we're creating. So that ideally, people would actually be able to identify those factors by themselves eventually without using the app' (expert 9).

This sentiment underscores the need for AI systems that not only detect disinformation but also provide users with background information on how these conclusions were reached, thus fostering greater understanding and trust. Despite the potential of AI in detecting disinformation, the reliance on human oversight remains crucial. One interviewee stated, 'I simply realised that ultimately every picture that is somehow floating around on the net has been changed, and you can't see what exactly has been changed. So perhaps only the shading was

adjusted somehow. It's lightened up a bit or a person has been removed or added and so at the moment it still needs the human eye to go along with it and somehow think with it, so I wouldn't trust it blindly' (expert 2). This highlights the importance of combining technological solutions with human judgment to enhance the credibility and reliability of AI tools. Another critical issue discussed was the complexity and proprietary nature of AI algorithms, which create challenges in ensuring transparency and accountability. An interviewee pointed out, 'The lack of interfaces and transparency from major platforms like Facebook, TikTok, X makes it difficult for science and civil society to understand how these platforms work and how information is distributed' (expert 6). This lack of transparency needs to be addressed to obtain more reliable information and effectively manage disinformation.

There were also differing opinions on the role of AI in disinformation detection. While some interviewees were optimistic about the potential of AI, others expressed caution. One expert remarked, 'AI is a driver of our economy; it is part of the structural change, and I know that Switzerland is extremely well positioned to benefit from this change and seise opportunities. And here, of course, we are extremely well positioned, as we represent not only the ICT [information and communication technology] and tech sectors but also the financial sector, the healthcare sector, etc. So we are in a super good position for AI. So the starting position is super good for AI, because the technology is still in its infancy but has great potential' (expert 4). In contrast, another interviewee was more sceptical about the current capabilities of AI, stating, 'We can do so much innovation with AI automatic detection, but if the awareness is not there, it's like not a good position' (expert 1).

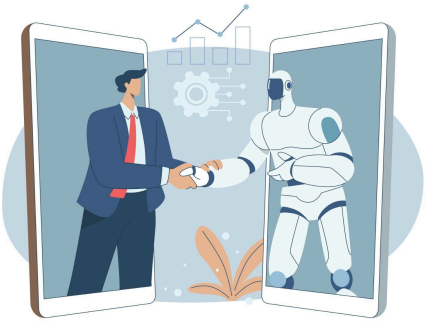
Additionally, an insight from the interview with FZI highlights the development of an XAI solution specifically designed to detect disinformation and provide explanations as to why the information is considered disinformation. This approach aims to increase transparency and user understanding, helping users critically reflect on the AI's decisions. The interviewee explained, 'We use methods of computer science [...] to kind of build an AI-based tool that is able to detect disinformation on digital platforms and provide users on those platforms with specific warnings about the piece of information that it detected. And not

only provide users with a specific warning that could be like […] "Attention! This piece of information was detected as disinformation". But it also kind of explains why the AI actually came up with this classification in the first place, to kind of give users more information about the whole concept that is behind these different factors that may stand for disinformation' (expert 9).

Increase of Trustworthiness and Credibility of AI and Data

One significant concern is the power wielded by large tech companies as gatekeepers of information. This dominance raises issues about the manipulation and filtering of data, which in turn affects the credibility and trustworthiness of the information disseminated. As expert 1 pointed out, 'In most of the cases, regulation is in favour of the big players, of the big boys, and is making it difficult for small players to attack the big players. So it is in fostering these oligopolistic, monopolistic structures. That's the reality' (expert 1). This manipulation by powerful gatekeepers is a significant risk, as another statement by expert 1 underscores: 'The biggest systematic risk is that powerful gatekeepers filter information that they manipulate' (expert 1).

In addition to the issues related to data manipulation, there is the danger of over-relying on AI tools for verifying information. This over-reliance can lead to significant risks, including political damage, if the tools provide inaccurate results. One interviewee emphasised the importance of human reflection and training in dealing with AI: 'I believe AI should be used by having people utilise and reflect on it and be trained in dealing with it, but not by focusing on building AI that controls disinformation. Again, we must ask who builds this AI, who defines disinformation, and who determines what is true and false' (expert 5). This highlights the need for transparency and a Human-in-the-Loop (HITL) approach to enhance trust in AI systems.



Ensuring accountability and mitigating the negative impacts of automated decisions on information distribution are crucial steps in this direction. As another expert mentioned, '[W]e are still confronted with these hallucinating technologies. In some cases, disinformation is still being produced by technology, for example, by making the wrong combination, in which it is not entirely clear who the author of the information is, even in these language models. Dealing with citations and sources. All of this is actually still underdeveloped, and at the same time, we have to wait for this to happen very quickly for significant gaps to be closed' (expert 3). This statement emphasises the need for transparency and education to address these challenges.

Despite the general consensus on the importance of trustworthiness and credibility, there are differing views on the best approaches to achieving this. For instance, while some experts advocate for stronger regulatory frameworks to control the influence of large tech companies and ensure transparency, others emphasise the role of education and awareness in fostering a critical understanding of AI and its limitations. One expert stated, 'I would not trust just one institution or just one element of that system that is tracking something and saying it's true or not. So, if only one institution has the legitimacy to say it's fake or not, I think that's quite critical' (expert 4). Another expert highlighted the role of education: 'It needs […] a lot in educating and enlightening people, because we end up using that, and if we don't learn to think even more critically, then we'll unlearn that at some point. And that doesn't just apply to Switzerland, but to all countries in general' (expert 5).

Despite the general consensus on the importance of trustworthiness and credibility, there are differing views on the best approaches to achieving this. For instance, while some experts advocate for stronger regulatory frameworks to control the influence of large tech companies and ensure transparency, others emphasise the role of education and awareness in fostering a critical understanding of AI and its limitations.

Reducing Disinformation

One of the primary strategies identified is enhancing digital literacy and critical thinking through education from an early age. This approach empowers individuals to recognise and counter disinformation. As expert 1 stated, 'The first very important tool would be to educate people; in German we would say *Aufklärung*. So, it's like in the normal world,

it's not different in the digital or AI-driven world. I have to make people aware of certain dangers' (expert 1). Another expert added, 'It is crucial, from my point of view, that we take people along on this journey from school to retirement. Transparency needs to be strengthened fundamentally so that people understand what they encounter' (expert 4).

Implementing public awareness campaigns and fostering media competence across all age groups is another crucial measure. These initiatives can significantly reduce the impact of disinformation by promoting critical engagement with digital content. One expert stated, 'I wouldn't just limit it to education or schools, but also we as a media company, report on the topic, raise awareness, and I think there are many other public companies or public institutions that can support this' (expert 2). Another expert noted, 'On the one hand, we certainly need to invest in education; we also have projects that show that media literacy in the area of digital media is much worse among the Swiss population than we previously assumed. We certainly need to do something about this. But at the same time, we also need to do something about those who are largely responsible for disseminating this content, i.e., the platforms. So it's not an either-or, but for us, the approach is, of course' (expert 8).

Legal frameworks and regulations are necessary to address illegal actions related to disinformation; these are supported by law enforcement and judicial systems to ensure accountability. Expert 1 highlighted this necessity: 'We need laws that cover these illegal actions, and I think for most of them there are already laws; there might be some white spots, and third, we need to be police and function in a legal system which can

sanction people and punish people who are obviously doing illegal things' (expert 1). Another expert expressed a similar view, stating, 'So I also think it's very important to understand how the mechanisms of the platforms work, how information is disseminated, i.e., according to which priorities, according to which aspects, how do these algorithms work?' (expert 6). Moreover, there was a consensus on the importance of public media literacy. All interviews highlighted the necessity of public media literacy as a fundamental component in the fight against disinformation.

2.3 Insights from Workshop

Own model based on literature	Based on Digital Xchange event	3 discussion rounds			
Key Theme	**Characteristics**	**Table 1**	**Table 2**	**Table 3**	**Total**
Ethical and Societal Consideration	Importance of Transparency in AI Processes	X	X	X	3
Algorithmic Bias and Missing Transparency	Ensuring Transparency in Decision-Making Processes	X	X	X	3
Transparent AI Techniques for Disinformation Detection	Explainable AI Approaches	X	X	X	3
Increase of Trustworthiness and Credibility of AI and Data	Public Awareness Campaigns to Build Trust in AI	X	X	X	3
Reduce (spread) of Disinformation	Increasing Public Media Literacy to Combat Disinformation	X	X	X	3
Ethical and Societal Consideration	Fact-Checking and Verification Methods	X		X	2
Impact of Disinformation and Missing Awareness	Educational Programs to Raise Awareness of Disinformation	X		X	2
Transparent AI Techniques for Disinformation Detection	Necessity of Human Oversight in AI Applications	X		X	2
Increase of Trustworthiness and Credibility of AI and Data	Promoting Research and Education on AI Credibility		X	X	2
Reduce (spread) of Disinformation	Implementing Human-in-the-Loop Systems for Better Control	X		X	2
Algorithmic Bias and Missing Transparency	Training and Education on Using AI Tools	X			1
Impact of Disinformation and Missing Awareness	Defining and Highlighting the Importance of Accurate Information	X			1

Table 2: Key insights from workshop
Source: Etienne Gerster

The Digital Xchange Event was organised in collaboration with digitalswitzerland Foundation. The aim was to understand the population's opinions, hopes, fears, and needs regarding disinformation, AI, fact-checking, and digital literacy. No specific target group was addressed, so no demographic data, such as age, gender, or occupation, is available. The target number of participants was between 20 and 30 people.

To further clarify the findings, the key themes and characteristics derived from the literature and supported by the event discussions are summarised in Table 2. The table highlights the themes that were frequently mentioned across multiple discussion tables, as well as those that were particularly emphasised during the event. For instance, the importance of transparency in AI Processes under the theme Ethical and Societal Consideration was mentioned by all discussion tables, highlighting its critical importance. Similarly, Explainable AI Approaches and Public Awareness Campaigns to Build Trust in AI were also emphasised by all groups, underscoring their significance. Conversely, Training and Education on Using AI Tools was only emphasised by a single table, indicating a less widespread but still significant concern. Additionally, a summary of the most important results of the interviews and the workshop can be found in Table 3.

Ethical and Societal Considerations

The participants emphasised the necessity of fact-checking tools that provide reliable information and clear explanations for their decisions. Fact-Checking should not solely rely on AI but also involve human oversight to ensure accuracy and trust . The need to cross-check information to trust verified information more effectively was also expressed.

There was a strong emphasis on the need for education and awareness campaigns to inform the public about the techniques used to spread disinformation. This includes integrating digital literacy into school curricula and providing resources for ongoing education. Continuous education and engagement with diverse information sources were raised as being critical to building resilience against disinformation. The development of ethical frameworks to guide the use of AI in disinformation detection was highlighted. These frameworks should address concerns about privacy, data security, and the potential misuse of AI tools. Transparency in how AI systems are trained and operated is essential to ensuring they respect human dignity, autonomy, and democratic values. Integrating human expertise into AI processes was deemed crucial to address biases and improve the reliability of disinformation detection systems. Human oversight can help ensure more accurate and context-aware analyses. It was highlighted that building public trust through transparent and explainable AI systems is essential to ensure credibiity and uphold democratic values. The role of public institutions and private companies in maintaining ethical standards and accountability in deploying AI technologies was emphasised.

Building public trust through transparent and explainable AI systems is essential to ensure credibility and uphold democratic values

Algorithmic Bias and Missing Transparency

Participants highlighted the importance of transparency in AI systems, ensuring users can understand how decisions are made. They highlighted the role of human expertise in addressing AI limitations and the need to educate the public on algorithmic biases. Transparency and awareness of potential biases are seen as key to mitigating their impact. Ethical frameworks were discussed as necessary to ensure responsible AI use,

which would help maintain public trust. Clear communication about AI's limitations and decision-making processes is crucial for building user confidence.

Impact of Disinformation and Awareness

Participants stressed the serious impact of disinformation and the public's lack of awareness, calling for educational campaigns to address these issues. They suggested integrating digital literacy into school curricula and promoting ongoing public education to build resilience against disinformation. Media literacy programs were highlighted as essential in teaching the public to critically evaluate information. Collaboration among governments, tech companies, and civil society was seen as key to combating disinformation, with clear guidelines and best practices for information verification. Critical thinking skills were emphasised as vital, with workshops and online courses suggested to help develop them.

Transparent AI Techniques for Disinformation Detection

Participants stressed the need to develop and use user-friendly tools that help individuals verify information themselves. These tools should be accessible and easy for the general public. Participants discussed the need for real-time detection and debunking systems to prevent the rapid spread of false information. Participants also discussed the need for AI systems that provide clear and understandable explanations for their outputs. Integrating human expertise into AI processes was essential to enhance accuracy and reliability. The desire for automated fact-checking with human oversight was considered essential.

Reduction of Disinformation

Participants stressed the need for real-time monitoring systems to detect and debunk disinformation, providing timely alerts to prevent its spread. Education and public awareness campaigns are crucial for helping individuals recognise false information and verify its accuracy. User-friendly verification tools should be developed to help the public combat disinformation. Additionally, improving digital literacy through workshops and online courses is key to building resilience against disinformation tactics. Making AI tools accessible was also emphasised to aid in this effort.

	Key Theme	Interviews
Challenges	Ethical and Societal Considerations	- Filtering and censoring by big tech companies - AI ethics solutions can inadvertently create more bias and restrict freedom of expression - Need for ethical guidelines to ensure responsible use of AI technologies
	Algorithmic Bias and missing Transparency	- Algorithmic bias and lack of transparency in AI systems pose significant risks - Efforts to mitigate bias include training annotators comprehensively and implementing explainable AI approaches
	Impact of Disinformation and missing awareness	- Disinformation poses a significant threat by spreading false information that can harm individuals - There is a critical need for enhanced digital literacy and awareness - The definition of disinformation is not clear and a balance
Solution	Transparent AI Techniques for Disinformation Detection	- Transparent AI techniques should involve explainable AI approaches that provide users with background information - Human oversight remains crucial - Transparency of Large platforms is necessary to increase the effectiveness and credibility of AI tools
Outcome	Rise Trustworthiness and credibility of AI and Data	- Transparency in AI solutions and a human-in-the-loop approach are crucial to enhance trust in AI systems - Regulation should aim to create transparency and access to the platforms' data in order to better understand them - Over-reliance on AI tools to verify information can lead to significant risks
	Reduction (the spread of Disinformation)	- Enhancing digital literacy and critical thinking through education from an early age is essential - Implementing public awareness campaigns and fostering media competence across all age groups - Legal frameworks and regulations are necessary to address illegal actions related to disinformation
	Key Theme	**Workshop**
Challenges	Ethical and Societal Considerations	- The importance of developing an ethical framework for the use of AI in detecting disinformation
	Algorithmic Bias and missing Transparency	- Knowing how an AI has been trained and how the concept behind it works - Does it tend to have bias in the algorithm by programmers? - AI systems must be transparent in their decision-making processes
	Impact of Disinformation and missing awareness	- Promoting critical thinking among the public in recognizing disinformation was seen as crucial - Participants emphasized the role of media literacy programs in educating the public
Solution	Transparent AI Techniques for Disinformation Detection	- Participants emphasized the need for AI systems that provide clear and understandable explanations for their outputs - The integration of human expertise into AI processes was seen as essential to enhance accuracy and reliability - The desire for automated fact checking with human oversight was considered essential
Outcome	Rise Trustworthiness and credibility of AI and Data	- Clear and simple explanations provided by AI systems were seen as essential for gaining user trust - Human involvement can help ensure the accuracy and reliability of AI systems, thereby increasing trust in these technologies
	Reduction (the spread of Disinformation)	- Education and awareness campaigns were seen as vital to reducing the spread of disinformation - Collaboration between governments, tech companies, and civil society was highlighted as essential to effectively combat disinformation - Teaching individuals how to navigate the digital landscape and critically assess information can significantly reduce the spread of disinformation

Table 3: Summarised findings from interviews and workshop
Source: Etienne Gerster

3. Discussion

The main findings of this research underscore the important role AI and fact-checking techniques can play in the fight against disinformation. However, it is crucial to note that these tools cannot operate in isolation. The key to effectively combating disinformation lies in the collaborative efforts of governments, technology companies, and civil society. This collective action is essential for establishing clear guidelines and best practices for information verification (Science Media Center, 2024).

Addressing the first research question:

> What insights can Switzerland gain from European countries and the USA on addressing disinformation, particularly regarding the use of AI and fact-checking techniques?

The study found that European countries and the US have implemented various AI-powered fact-checking systems. These systems, while not accessible to the public, have shown promise in increasing trust and transparency in verifying information. However, their effectiveness is currently limited by challenges such as language nuances and the interpretation of cultural context. This underscores the need for human oversight alongside automated processes, as human judgement is crucial in interpreting complex information.

The second research question:

> How can Switzerland utilise AI practices to enhance fact-checking and counter disinformation effectively?

found that Switzerland can rely on XAI to detect disinformation. With a transparent and explainable solution behind the results, trust among users can be strengthened. However, integrating human overview into AI processes must also be considered to improve and strengthen accuracy, reliability, and trust. The results showed that relying on AI practices alone is insufficient to combat disinformation effectively. A comprehensive strategy must include a combination of (i) prebunking, (ii) debunking, (iii) regulation, and (iv) collaboration. Prebunking (i.e., improving public awareness of disinformation detection before it spreads through training and awareness campaigns) is a proactive approach that can significantly reduce the impact of misinformation. Debunking (i.e., uncovering and correcting false claims)

remains crucial as a reactive measure. Regulations are needed to set standards and guidelines for verifying information and hold those spreading disinformation accountable. Cross-sector collaboration is essential to pool resources, share knowledge, and form a united front against disinformation.

3.1 Fact-Checking in EU Countries

Recent advancements (Duke Reporters' Lab, 2023; Ünver, 2023) have shown how countries effectively deploy automated fact-checking systems, albeit with limited accessibility to the general public. These efforts are part of a broader movement towards integrating evolving technologies such as blockchain, which has been key in enhancing trust and transparency in information verification (Buțincu & Alexandrescu, 2023). This move towards more reliable and secure systems reflects significant technological progress and collaborative initiatives that are shaping the future of automated fact-checking (Ünver, 2023). However, as combating disinformation with AI is a relatively new field, few use cases and statistics demonstrate that fact-checking is effective.

However, this technological progression is not devoid of challenges. Automated systems often struggle with the complexity of language nuances and the subtleties of cultural contexts crucial for accurate interpretation — a limitation echoed by researchers such as Trokhymovych and Saez-Trumper (2021). These challenges underscore the current limitations of automated systems and emphasise the ongoing need for human oversight. Experts in interviews and the participants of the Digital Xchange Event underline the critical need for human oversight in fact-checking solutions to increase trust. Human judgement remains indispensable in fact-checking, particularly in interpreting the context and significance of claims—factors that automated systems often overlook (Demartini et al., 2020). This insight is vital, especially when considering disinformation's political or social implications, where a more profound understanding can significantly influence the accuracy and impact of fact-checked claims (Bieri et al., 2021).

Furthermore, according to Bieri et al. (2021), these technologies have far-reaching implications for public discourse and policy, especially in democratic societies where access to accurate information is crucial. Automated fact-checking can influence public opinion and policy by quickly clearing

up disinformation. However, automated disinformation detection raises important ethical questions about suppressing information and content regulation. These considerations must be carefully weighed to ensure that automated fact-checking supports healthy public discourse without compromising freedom of expression oraccess to information (Thouvenin et al., 2024). One expert in the interview supported this statement, explaining that people in a survey said there was a fear that automated fact-checking solutions would deliberately censor content. By contrast, participants at the Digital Xchange event said they were not afraid of censorship but would favour only seeing content checked by AI. However, the human control aspect of fact-checking was again found to be very important.

In order to address these concerns, various regulations are being considered and implemented. For example, the EU adopted the Digital Services Act whereby digital platforms are responsible for combating disinformation (European Commission, 2023). In Switzerland, OFCOM is preparing a consultation on a draft legislation for early 2025: the New Federal Law on Communication Platforms and Search Engines (LPCom). The law seeks to give the Swiss population has more rights vis-à-vis the significant communication platforms and enables them to demand transparency (Der Bundesrat, 2023).

Participants of the Digital Xchange Event emphasised the need for human oversight in automated AI fact-checking solutions to build trust, with a heavy reliance on human resources. Experts from the interviews also highlighted that automated fact-checking solutions are resource-intensive, as supported by Ünver (2023). Nevertheless, the demand for automated fact-checking solutions is omnipresent among Digital Xchange participants. This is particularly relevant for the FZI's "DeFakts" project, which develops an XAI system that enables Germany to identify disinformation in real time. DeFakts can be a smartphone mobile application and will be available to the public.

After the interview, FZI provided a video of the beta test, which shows how accurately disinformation can be discovered within every application on the smartphone, with an additional explanation of why the content was identified as disinformation. DeFakts, which can analyse German language content, could be interesting for Switzerland as, according to the Federal Statistical Office (2022b), 62% of the Swiss populace speaks

German. "DeFakts" also aims to minimise the bias of programmers and project members through workshops — a critical step acknowledged by experts and event participants alike. They concur on the importance of reducing this bias, as discussed in the literature review (Thompson et al., 2022).

Automated fact-checking can influence public opinion and policy by quickly clearing up disinformation. However, automated disinformation detection raises important ethical questions about suppressing information and content regulation.

An expert further highlighted the communication gap between programmers and the public, noting that the complexity of algorithms often renders them a “black box” that is difficult for people to understand. This transparency issue complicates the public's ability to trust and effectively use these AI systems, underscoring the complex psychosocial dynamics of disinformation that cannot be addressed solely through technology (Graf et al., 2022). According to Bateman and Jackson (2024), the integration of generative AI holds promise beyond mere information generation. When well-designed and supervised by humans, these systems can significantly expedite the fact-checking process, though the long-term impact of generative AI on disinformation has yet to be fully understood.

3.2 Effective Measures against Disinformation

The experts interviewed agree that fact-checking alone is not a solution for combating disinformation. To do so requires education, awareness campaigns (prebunking), and the verification of information, leading to the discovery of disinformation (debunking). Regulations and collaboration influence both prebunking and debunking methods (Science Media Center, 2024). Disinformation can be discovered through professional fact-checkers or AI solutions, as underlined by Graves (2018). Both require significant resources and time. Aside from the resource problem, another issue is the lack of definition of disinformation. Various participants highlighted the issue that the term has countless definitions, pointing out that content disliked by an individual can also be viewed as disinformation (Vogler et al., 2021). Additionally, some interviewed experts mentioned that it often falls to a single group

or organisation to define what constitutes disinformation, which can introduce biases. Thus, the line between disinformation, correct information, and freedom of opinion is narrow (Vogler et al., 2021).

FZI has confronted this issue with its DeFakts project. The company has invested significant time in training its employees to determine what disinformation is and what it is not. That means that training and education are not only relegated to the general public but also to the developers of such fact-checking solutions. Several participants noted that they have already conducted training and presentations in schools on disinformation and its detection. Participants of the Digital Xchange event emphasised the necessity of such training, not only in schools but also for older, less tech-savvy people. According to the Federal Statistical Office (2023), 97% of respondents aged 15 to 88 accessed the Internet in the past three months, and 50% encountered false or questionable information. Correctiv (2024) launched a prebunking campaign with various videos and posts on disinformation and its detection to raise awareness among citizens of all ages. In Switzerland, similar initiatives exist, such as “ch.ch” and “Jugend und Media”, which also strive to raise awareness about disinformation detection. Interview experts and participants of the event saw the promotion of digital and media literacy as an essential prebunking method. The Science Media Center (2024) highlights the importance of these competencies, noting the need for long-term and in-depth development of these skills. Long-term media literacy promotion is crucial to enabling people to evaluate information critically. Transparency and accountability of online platforms and legal frameworks are also necessary to combat disinformation effectively (Science Media Center, 2024). This underscores the need for a holistic approach integrating education, regulation, and technological solutions.

According to some interview experts, debunking requires continuous development of fact-checking and AI-based solutions to effectively counter the ever-evolving techniques of disinformation. This approach, emphasised by Humprecht (2019) and FZI with their DeFakts project, is crucial. It also emphasises that online platforms should be obliged to identify, remove, and flag disinformation, regardless of whether it has been identified by automated AI fact-checking. Participants from the event agreed that disinformation should be

labelled, as it would make it easier for them to recognise disinformation themselves. An analysis of a not-yet-published survey by OFCOM suggests that initial analysis of the results showed that most respondents would trust content labelled "checked by an AI fact-checking tool" less. This statement aligns with the opinions of other interview experts, who unanimously said they would not trust AI's isolated identification of disinformation. In addition to the label of disinformation, there also needs to be an explanation of why the AI identified something as potential disinformation—like an XAI, as in DeFakts. Additionally, interview experts and event participants repeatedly referred to the Human-in-the-Loop solution, where the positive attributes such as efficiency and scalability of AI solutions are combined with the final control of humans. Demartini et al. (2020) and Ünver (2023) also see this method as one that can significantly increase public trust in fact-checking.

National and international collaboration plays a crucial role in the fight against disinformation. Switzerland actively works with national and international partners to develop global standards and best practices that address the ethical and legal challenges surrounding AI-driven disinformation identification (Thouvenin et al., 2024). The interview experts and Science Media Center (2024) emphasise that the combination of prebunking and debunking should be supplemented by regulation. Such measures can be supported by guidelines and laws promoting preventive and reactive measures. Platforms need to be more transparent about their algorithms and sources to prevent the spread of disinformation (Science Media Center, 2024).

Additionally, legal regulations should ensure that platforms correct and rectify disinformation (Bateman & Jackson, 2024). According to Thouvenin et al. (2024), regulation of online platforms is an option, but messaging services such as WhatsApp, Signal, and Telegram are neglected because they can establish almost no active disinformation detection within chats. However, a recent survey by Statista (2024a) shows that only 19% of respondents encountered disinformation on these messaging services, while 59% reported encountering disinformation on social media. In addition to prebunking, debunking, and regulation, collaboration within an ecosystem is also helpful in countering disinformation (Crowley, 2023; World Health Organization, 2022). The WHO proposes a three-way

partnership where authorities of member states, industry/platform owners, and civil society work together to combat disinformation effectively. The EDMO also supports collaboration and has created a network of fact-checking organisations based in the EU. Interview experts agree that no single actor can effectively combat disinformation alone; it requires a community from various sectors of the economy.

Key Summary

The model developed and enhanced by theory, was confirmed by interviews and participants at the event.
Transparent AI Techniques for disinformation detection was modified to debunking.
In addition, the solution area was extended with prebunking, regulation and collaboration.

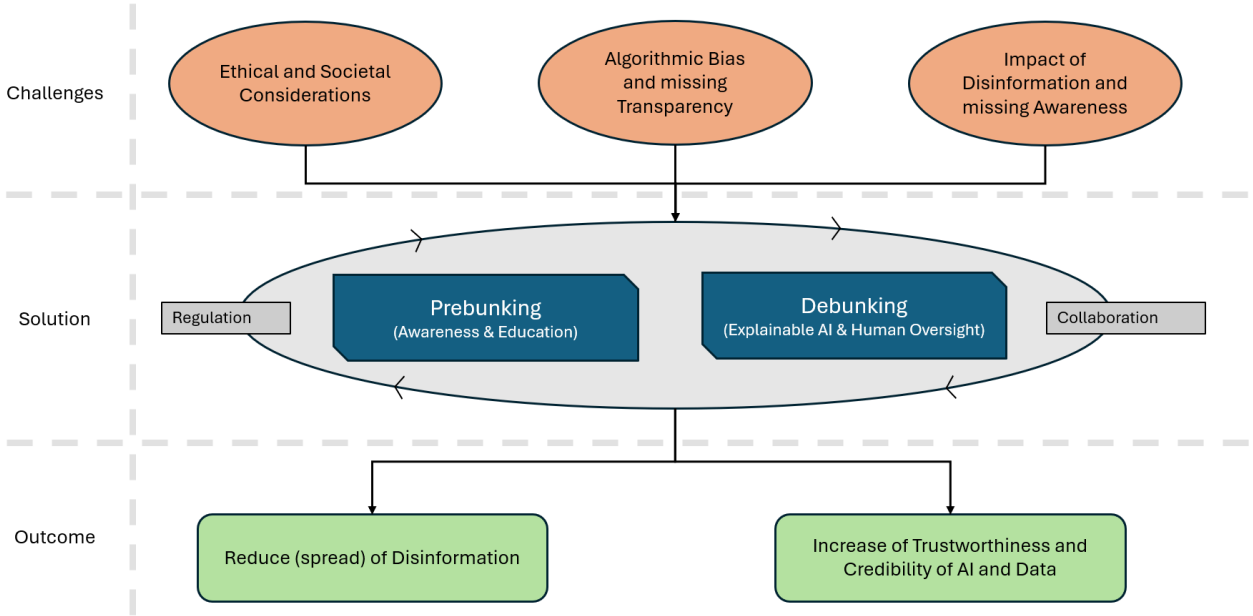


Figure 4: Adjusted theoretical model from literature review after data collection

Source: Etienne Gerster

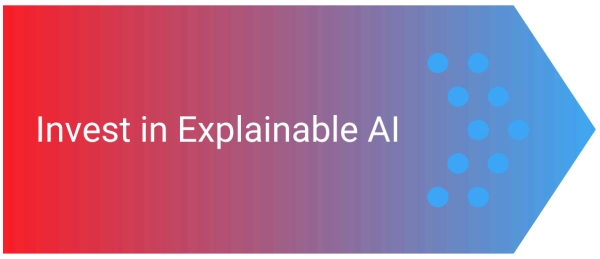
4. Recommendations

In order to effectively address the challenges associated with disinformation, Switzerland must adopt a comprehensive and holistic approach. This section outlines several critical recommendations to foster collaboration, advance technological solutions, and promote education and literacy. The summarised recommendations can be found in Table 4.

To effectively combat disinformation, fostering collaboration and dialogue among various stakeholders is crucial. Organising regular workshops, conferences, working groups that bring together representatives from academia, NGOs, civil society, the media, private technology companies, and government entities. These events should serve as platforms for exchanging ideas, discussing recent advancements in AI and fact-checking technologies, and formulating unified approaches to tackling disinformation. The focus will be sharing practical insights and research findings, fostering partnerships, and developing actionable strategies to address disinformation. Additionally, these exchanges should lead to the development of joint strategies to address specific disinformation-related challenges that combine technological solutions with regulatory and educational measures to ensure a coordinated and multi-faceted approach to disinformation.

Collaborative research projects between universities and private technology companies to develop and test innovative solutions for countering disinformation are also important. For instance, launching a research initiative between Swiss Universities and private technology companies to develop AI solutions for detecting disinformation can foster innovation by leveraging the expertise and resources of academic institutions and the tech industry. Working closely with political decision-makers to create frameworks and incentives that foster cooperation between sectors is also very important. This includes advocating for policies that promote cross-sector collaborations, such as tax incentives for companies participating in anti-disinformation projects. These policies will incentivize private companies to invest in and contribute to multi-stakeholder projects

that address disinformation by creating an enabling environment that encourages collaborative efforts, reducing financial barriers, and providing regulatory support.



Switzerland's digital ecosystem should focus on building an XAI system that creates AI solutions that provide clear and understandable explanations for their outputs and decisions, fostering greater user transparency and trust.

In order to achieve this, a partnership with FZI to gain insights into the training of developers and the utilisation of their German-language database. This collaboration can also provide a model for project development that digitalswitzerland, for example, can emulate for similar initiatives.

Alternatively, instead of developing a new system from scratch, actors can consider implementing the DeFakts project. Leveraging the existing DeFakts project can save time and resources while benefiting from an already proven and effective XAI solution.

Integration of Digital Literacy into Curriculums

Awareness raising campaigns are crucial in raising public awareness and educating the public. By promoting the initiatives undertaken by various stakeholders, these campaigns will build public trust and help identify and counter disinformation. Collaboration with educational authorities to integrate digital and media literacy into school curricula is of utmost importance. This involves cooperation with educators to create comprehensive curriculum modules covering the critical evaluation of information, recognition of disinformation, and responsible use of digital tools. These modules should be age-appropriate and integrated into existing subjects such as social studies, language arts, and technology.

Additionally, training programs for teachers to equip them with the knowledge and resources needed to teach digital literacy

effectively. This could include workshops, online courses, and access to a repository of teaching materials and best practices. Developing and distributing interactive learning tools, such as educational software, games, and mobile apps, will further engage students in learning about digital literacy and critical thinking.

Organise Public Workshops and Courses

Workshops, seminars, and online courses should be tailored to different demographics, especially older and less tech-savvy populations. Partnerships with community centres, libraries, and senior organisations can facilitate in-person sessions on recognizing fake news and using digital tools. User-friendly online platforms with various formats—videos, interactive modules, guides—should be developed. Targeted outreach for older adults might include informational campaigns, printed materials, and hands-on assistance. Collaboration with media and tech companies to create and distribute educational content can further promote digital literacy through public service announcements and social media campaigns.

Leverage International Insights and Best Practices

Monitoring of and participation in global forums to stay informed about the latest developments and best practices in combating disinformation through AI. Engaging in these international platforms ensures that Switzerland remains at the forefront of innovation and policy development in this field.

5. Conclusion

This study's primary focus is investigating how AI can combat disinformation, with a particular emphasis on automated fact-checking. This research is driven by the increasing spread of disinformation in digital media and its associated challenges to democratic processes and public trust.

The research involved a comprehensive literature review, expert interviews, and a Digital Xchange Event with the public. The findings suggest that, when integrated with effective fact-checking mechanisms, AI can

significantly curb the spread of disinformation. However, to effectively combat disinformation, a combination of prebunking (educating and training the public in disinformation detection), debunking (identifying and correcting disinformation), regulation (legislative measures), and, most importantly, collaboration within ecosystems is necessary. The research examined both current practices in Switzerland and international approaches, providing a comparative analysis that informed recommendations for Switzerland.

This study confirmed several findings from the existing literature, such as the potential efficient impact of automated and transparent fact-checking solutions in combating disinformation. The regulatory and legal frameworks for AI in Switzerland are still developing, and there are challenges related to public trust and transparency in AI applications. A notable new finding is the potential of transparent AI-driven fact-checking systems to curb the spread of disinformation and enhance public trust in information, provided these systems are transparent and explainable. Furthermore, the study identified a significant gap in Switzerland's absence of AI-powered fact-checking platforms. Unexpectedly, there is a high distrust from the public toward fully automated fact-checking, which can be mitigated by involving humans in the verification process. The results have significant practical implications for Switzerland and stakeholders committed to combating disinformation. The proposed effective measures for countering disinformation can be applied across various sectors, including media, education, and public administration. Additionally, the findings are transferable to other countries facing similar challenges with disinformation, especially those with high digital media consumption.

Further studies could also explore the development and implementation of AI-based fact-checking platforms in Switzerland and assess their impact on reducing disinformation. Additionally, collaborative projects between academia, government, and industry could promote innovation in AI applications for disinformation management.

6. References

Baker, A., & Fairbank, V. (2022). *How to fact-check*. The Truth in Journalism Fact-Checking Guide. https://thetijproject.ca/guide/how-to-fact-check/

Bateman, J., & Jackson, D. (2024). *An evidence-based policy guide Countering disinformation effectively.* He Whenua Taurikura. https://hwt.ac.nz/wp-content/uploads/2024/05/Carnegie_Countering_Disinformation_Effectively.pdf

Bieri, U., Weber, E., Braun Binder, N., Salerno, S., Keller, T., & Kälin, M. (2021). *Digitalisierung der Schweizer Demokratie Technologische Revolution trifft auf traditionelles Meinungsbildungssystem*. TA-SWISS. https://edoc.unibas.ch/84421/1/20210907115152_6137363884ca6.pdf

Blarer, A., Buffat, M., Busch, C., Egloff, D., Fanzun, J., Haefliger, G., Langer, P., Loison, B., Luder, T., Malz, A., Scheidegger, E., Schneider, T., Schöll, M., Schwaar, P., Stämpfli, M., & Weber, V. (2019). *Herausforderungen der künstlichen Intelligenz Mitglieder der interdepartementalen Arbeitsgruppe künstliche Intelligenz*. KMU-Portal. https://www.kmu.admin.ch/dam/kmu/de/dokumente/FaktenundTrends/herausforderungen-der-kuenstlichen-intelligenz.pdf.

Bontridder, N., & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data and Policy*, *3*, Article e32. https://doi.org/10.1017/dap.2021.20

Bundesamt für Kommunikation. (2021). *Intermediäre und Kommunikationsplattformen*. Bundesamt für Kommunikation. https://www.bakom.admin.ch/dam/bakom/de/dokumente/bakom/elektronische_medien/Zahlen%20und%20Fakten/Studien/bericht-kommunikationsplattformen-und-intermediaere-2021.pdf.

Bundesamt für Statistik. (2022a). *Desinformation im Internet – Wahrnehmung und Massnahmen*. Bundesamt für Statistik. https://www.bfs.admin.ch/asset/de/22624872

Bundesamt für Statistik. (2022b). *Sprachenlandschaft in der Schweiz*. www.statistik.ch

Bundesamt für Statistik. (2023). *Die weit verbreitete Internetnutzung macht die Schweizer Bevölkerung anfälliger für Desinformation und Hassreden*. Bundesamt für Statistik. https://www.bfs.admin.ch/asset/de/28465185

Buțincu, C. N., & Alexandrescu, A. (2023). Blockchain-based platform to fight disinformation using crowd wisdom and artificial intelligence. *Applied Sciences*, *13*(10), Article 6088. https://doi.org/10.3390/app13106088

CNAI. (2022). *CNAI - Kompetenznetzwerk für künstliche Intelligenz*. https://cnai.swiss/

Correctiv. (2024). *Lass dich nicht manipulieren*. https://correctiv.org/falschinformationen-erkennen/

Crowley, B. (2023, July 20). *Collaboration is key to fighting online misinformation*. Friends of Europe. https://www.friendsofeurope.org/insights/critical-thinking-collaboration-is-key-to-fighting-online-misinformation/

Demartini, G., Mizzaro, S., & Spina, D. (2020). *Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities*. ResearchGate. https://www.researchgate.net/publication/345264315_Human-in-the-loop_Artificial_Intelligence_for_Fighting_Online_Misinformation_Challenges_and_Opportunities

Der Bundesrat. (2021). *Leitlinien «Künstliche Intelligenz» für den Bund*. Der Bundesrat. https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-81319.html

Der Bundesrat. (2023, April 25). *Federal Council seeks to regulate large communication platforms*. Schweizerische Eidgenossenschaft. https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-94116.html

Digital Society Initiative. (2024*). DSI Insights: ChatGPT erhöht die Skepsis gegenüber KI – darauf verzichten will man aber doch nicht*. Digital Society Initiative UZH. https://www.dsi.uzh.ch/de/current/news/2024/dsi-insights-ki-schweiz-skepsis.html

Digitale Schweiz. (2023). *Strategie Digitale Schweiz*. Schweizerische Eidgenossenschaft. https://digital.swiss/de/strategie/strategie.html#vision

digitalswitzerland. (2023, May 3). *Switzerland to become a leading digital nation – our "*

Strategy 2025."

https://digitalswitzerland.com/switzerland-to-become-a-leading-digital-nation-our-strategy-2025/

DIZH. (2024). *ClarifAI: Das Unsichtbare sichtbar machen mit KI-gestützter Propaganda-Erkennung und Faktenüberprüfung*. https://dizh.ch/2023/12/21/clarifai-das-unsichtbare-sichtbar-machen-mit-ki-gestuetzter-propaganda-erkennung-und-faktenueberpruefung/

Duke Reporters' Lab. (2023). *Fact-checking*. https://reporterslab.org/fact-checking/

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association*, *43*(2), 97–116. https://doi.org/10.1080/23808985.2019.1602782

European Commission. (2023). *The EU's Digital Services Act*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

FactCheck.org. (2020, August 12). *Our process*. https://www.factcheck.org/our-process/

FZI Forschungszentrum Informatik. (2021). *Forschungsprojekte*. https://www.fzi.de/project/defakts/

Graf, F., Obrecht, L., & Weiner, S. (2022). *Erste Erkenntnisse zu Transparenz, Diskriminierung und Manipulation Rechtliche Rahmenbedingungen für Künstliche Intelligenz in der Schweiz*. ITSL UZH. https://www.itsl.uzh.ch/dam/jcr:9224497c-ce45-43a5-9b7f-1c3eb106bce3/Graf%2520Fabienne,%2520Obrecht%2520Liliane,%2520Weiner%2520Soraya,%2520Erste%2520Erkenntnisse%2520zu%2520Transparenz,%2520Diskriminierung%2520und%2520Manipulation,%25202022.pdf

Graves, L. (2018). *Understanding the promise and limits of automated fact-checking*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-nqnx-bg89

Humprecht, E. (2019). How do they debunk "fake news"? A cross-national comparison of transparency in fact checks. *Digital Journalism*, *8*(3), 310–327. https://doi.org/10.1080/21670811.2019.1691031

Iqbal, A., Shahzad, K., Khan, S. A., & Chaudhry, M. S. (2023). The relationship of artificial intelligence (AI) with fake news detection (FND): A systematic literature review. *Global Knowledge, Memory and Communication*. Advance online publication. https://doi.org/10.1108/GKMC-07-2023-0264

Kertysova, K. (2018). *Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered*. Security and Human Rights Monitor. https://www.shrmonitor.org/assets/uploads/2019/11/SHRM-Kertysova.pdf

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Ramp, D., Köng, A.-L., Holenstein, M., & Angst, L. (2024). *Mobiliar #Digital Barometer 2024 Die Stimme der Schweizer Bevölkerung Initiiert und durchgeführt durch*. Risiko_Dialog. https://www.digitalbarometer.ch/uploads/digitalbarometer_2024_de.pdf Rest of World. (n.d.). *Rest of World*. Retrieved July 19, 2024, from https://restofworld.org/

Rüegg-Stürm, J., & Grand, S. (2020). *Das St. Galler Management-Modell: Management in einer komplexen Welt*.

Santos, F. C. C. (2023). Artificial intelligence in automated detection of disinformation: A thematic analysis. *Journalism and Media*, *4*(2), 679–687. https://doi.org/10.3390/journalmedia4020043

Saunders, M. N. K., Lewis, P., & Thornhill, Adrian. (2007). *Research methods for business students*. Financial Times/Prentice Hall.

Science Media Center. (2024). *Welche Maßnahmen gegen Desinformation helfen*. https://www.sciencemediacenter.de/alle-angebote/science-response/details/news/welche-massnahmen-gegen-desinformation-helfen/

Shahzad, K., Khan, S. A., Ahmad, S., & Iqbal, A. (2022). A scoping review of the relationship of big data analytics with context-based fake news detection on digital media in data age. *Sustainability*, *14*(21), Article 14365. https://doi.org/10.3390/su142114365

Shruti, M. (2023). *Differences between AI vs. machine learning vs. deep learning*. Simplilearn. https://www.simplilearn.com/tutorials/artificial-intelligence-tutorial/ai-vs-machine-learning-vs-deep-learning

SRF. (2022, April 1). *Tag des Faktenchecks: So funktioniert der Faktencheck bei SRF*. https://www.srf.ch/news/international/tag-des-faktenchecks-so-funktioniert-der-faktencheck-bei-srf

Statista. (2024a). *Desinformation: Verbreitung über Plattformen*. https://de.statista.com/statistik/daten/studie/1453891/umfrage/umfrage-zur-verbreitung-von-desinformation-ueber-plattformen/

Statista. (2024b). *Wahljahr 2024: Wahlen weltweit im Überblick*. https://de.statista.com/themen/11879/wahlen-weltweit-2024/

Swissinfo. (2020, November 2). *How Switzerland has responded to online disinformation*. https://www.swissinfo.ch/eng/swiss-politics/how-switzerland-has-responded-to-online-disinformation/46135098

Thompson, R. C., Joseph, S., & Adeliyi, T. T. (2022). A systematic literature review and meta-analysis of studies on online fake news detection. *Information*, *13*(11), Article 527. https://doi.org/10.3390/info13110527

Thouvenin, F., Volz, S., Eisenegger, M., Vogler, D., & Jaffé, M. (2024). *Governance von Desinformation in digitalisierten Öffentlichkeiten*. Europäische Kommission. https://digital-strategy.ec.europa.eu/de/policies/online-disinformation

Tong, A., Sainsbury, P., & Craig, J. (2007). *Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups*. *International Journal for Quality in Health Care*, *19*(6), 349–357. https://doi.org/10.1093/intqhc/mzm042

Trokhymovych, M., & Saez-Trumper, D. (2021). *WikiCheck: An end-to-end open source Automatic Fact-Checking API based on Wikipedia*. ArXiv. https://doi.org/10.48550/arXiv.2109.00835

Ünver, A. (2023). *Emerging technologies and automated fact-checking: Tools, techniques and algorithms*. EDAM. https://edam.org.tr/Uploads/Yukleme_Resim/pdf-28-08-2023-23-40-14.pdf

Vogler, D., Schwaiger, L., Schneider, J., Udris, L., Siegen, D., Marschlich, S., Rauchfleisch, A., & Eisenegger, M. (2021). *Falschinformationen, Alternativmedien und Verschwörungstheorien-Wie die Schweizer Bevölkerung mit Desinformation umgeht. Bericht für das Bundesamt für Kommunikation*. Zurich Open Repository and Archive. https://doi.org/10.5167/uzh-219206

World Health Organization. (2022, October 20). *Collaboration is key to countering online misinformation about noncommunicable diseases*. https://www.who.int/azerbaijan/news/item/20-10-2022-collaboration-is-key-to-countering-online-misinformation-about-noncommunicable-diseases--new-who-europe-toolkit-shows-how

Zürcher Hochschule der Künste. (2022). *Jahrestagung 2022*. https://www.zhdk.ch/doktorat/epistemologien-aesthetischer-praktiken-9145/veranstaltungen-19505/jahrestagung-2022-1951

6.1 List of Abbreviations

Abbreviations	Full text	First-time entry
AI	Artificial Intelligence	Section 1
ML	Machine Learning	Section 1.3.2
DL	Deep Learning	Section 1.3.2
COREQ	Consolidated Criteria for Reporting Qualitative Research	Section 2.2
NGO	Nongovernmental organisations	Section 2.2
FHNW	Fachhochschule Nordwestschweiz	Section 2.3
CNAI	Competence Network for AI	Section 3.1.1
XAI	Explainable Artificial Intelligence	Section 3.1.2
FZI	Research Center for Information Technology	Section 3.1.2
DIZH	Digitalisation Initiative of the Zurich Higher Education Institutions	Section 3.1.2
API	Application Programming Interface	Section 3.1.2
HITL	Human-in-the-loop	Section 3.1.5
OFCOM	Federal Office of Communication	Section 4.1
EDMO	European Digital Media Observatory	Section 4.2

Overview of some Swiss initiatives on disinformation (non- exhaustive)