Master Thesis

Designing for youth resilience against AI-generated extremist disinformation

TU Delft

Colophon

Master Thesis

MSc. Integrated Product Design

Faculty of Industrial Design Engineering

Date and Place

6th of May, 2025, Delft - Wim Crouwel Hall

Author

O.I. (Odine) de Bruijn

Supervisory team

Dr. rer. nat. T.D. Dingler (chair)	SDE - KInD
Prof. Dr. P.A. Lloyd (mentor)	DOS - CP

TU Delft

Acknowledgements

This thesis marks the conclusion of my master's journey at TU Delft. I would like to take this opportunity to thank the people who supported and inspired me throughout this process.

My sincere thanks go to my supervisors, Tilman Dingler and Peter Lloyd, for their guidance and critical feedback. Their advice to focus on small, concrete steps rather than grand ambitions shaped the direction of this project and helped turn it into a practical and implementable solution. It is a lesson I will carry with me beyond this thesis.

I am grateful to Gwenda Nielen, whose talk at a symposium a few years ago first sparked my interest in the topic of digital resilience. At the start of this thesis, I reached out to her again. She kindly responded and took the time for a good conversation, which helped guide the focus of this project. Instead of trying to eliminate disinformation entirely, the aim shifted towards building resilience in individuals.

Special thanks go to my old high-school, Stedelijk Gymnasium Breda, and Loes Rullens for the opportunity to give a guest lesson and gather essential user feedback. The students' experiences with generative AI and social media, and the reflections of their teachers, played an important role in shaping the final educational toolkit.

I also thank my friends, basketball teammates, and housemates for their patience, encouragement, and support during the busy final months. Special thanks to Elies for the many coffee breaks and thesis-writing sessions that helped keep me focused, even if some breaks were longer than planned.

I am deeply grateful to my family for their unwavering support throughout my studies. Thank you Mom, for sending me every article that even remotely had something to do with AI. I am sure some of them inspired decisions in this thesis ;). Thank you Dad, for always helping me find a direction when I felt stuck, often understanding where I wanted to go before I realised it myself. And thank you to my brother for checking in, believing in me, and supporting me.

Finally, thank you Isabelle for your unconditional support, endless patience, and for always helping me find a solution when I needed it most. Your presence truly made this journey not only successful but also fun.

Delft, May 2025,
Odine de Bruijn

Abstract

This thesis presents the design and development of an Educational Toolkit aimed at building the resilience of 15–16-year-olds against AI-generated extremist disinformation. With the rise of generative AI, AI-powered tools are rapidly increasing in sophistication and accessibility. Malicious actors, mainly extremist groups, leverage these technologies to manipulate, recruit, and radicalise youth through spreading disinformation by various tactics. Examples are the generation of deepfakes, memetic warfare, and AI-enhanced grooming. Existing media literacy interventions have not yet answered to these developments and a digital media literacy tool for this specific target group is yet to be made. This thesis aims to respond to this need.

To address this gap, an educational toolkit was designed, combining Inoculation theory with an interactive, gamified experience. The contents are structured through various frameworks. The toolkit will consist of multiple lessons, where each lesson is paired with a serious game that simulates real-world disinformation tactics in a controlled and ethical environment. These lessons will focus on key manipulation techniques such as meme-based manipulation (memetic warfare), source impersonation through AI, and emotionally charged misinformation campaigns.

The thesis uses Bloom's Taxonomy to define cognitive learning objectives and Gagné's Nine Events of Instruction to guide lesson structure. Integrating the theory with the game, and connecting multiple lessons to form a reflective experience, is done by using Kolb's Experiential Learning Cycle. This theoretical foundation is operationalised in the game through the Mechanics-Dynamics-Aesthetics (MDA) framework. One of the toolkit's most notable features is a genAI sandbox, a simulated chatbot and image generation interface that allows students to experiment with AI prompt creation in a fictionalised and safeguarded setting. This component encourages hands-on learning and enables learning by doing.

Two prototypes were developed: one using Twine and one via ChatGPT's custom GPT function. A guest lesson with a live classroom gave insights on the target group and shifted the project focus to educational environments instead of standalone serious games. During the lesson, students created AI-generated memes containing disinformation. This activity was followed by reflection on narrative techniques and emotional impact. A general survey and classroom observation further supported the need for not only gamified approaches, but also teacher-led theory and reflection.

While the educational toolkit lays a strong foundation, further steps are needed to bring it into practice. These include further developing of the lesson theory and corresponding serious games and validating the toolkit in high-schools. Completing these steps will turn the educational toolkit from a high-level research-based solution into a digital media literacy solution for strengthening resilience among the target group.

This thesis contributes to the field of design for digital resilience by offering a proposed, research-based educational toolkit that acknowledges the current and evolving threat landscape as enabled by generative AI. It also provides a replicable framework for integrating serious games into (digital) media literacy education.

Table of Contents

Introduction

1. Introduction

With the rise of generative AI, extremists are finding novel ways of creating disinformation and spreading their, often violent, ideologies. Younger audiences are vulnerable targets to these narratives. Traditional media literacy programs often focus on spotting misinformation but rarely address the growing threat of AI-generated extremist content. This thesis examines how AI-generated extremist content is created and how 15-16-year-olds can build resilience against these threats. This thesis poses a solution in the form of an educational toolkit for teachers integrated with serious game elements.

1.1 Project Goal

Building resilience among teenagers can be achieved through multiple ways. To make it interactive, we will work towards making a lesson with serious gaming elements. Through this tactic teenagers will learn about the mechanisms behind extremist disinformation, its consequences and generative AI (genAI) literacy. Therefore the project goal of this thesis is defined as follows:

"Design a gamified tool that educates young users on the creation and impact of AI-generated disinformation, encouraging awareness and resilience against extremist-driven disinformation through simulated experiences."

1.2 Key Topics

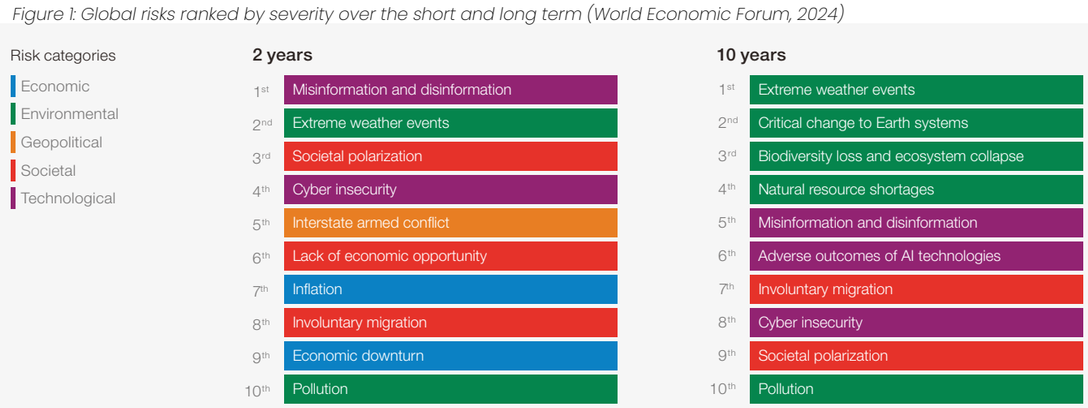
The subject of this thesis can be divided into three key topics: disinformation, extremism, and generative artificial intelligence. At their intersection lies the main challenge of this thesis. How do extremists use generative AI to manipulate and recruit young people, and how can teenagers develop the skills needed to recognise and build resilience against these tactics?
The topics will shortly be defined in this introduction. A more in-depth look will be given in Chapter 2 *Literature Review.* Furthermore, this introduction will also outline the target group and the focus on building resilience.

Disinformation

All three key topics are highly relevant in todays world, starting with the risk of disinformation on society. Disinformation distinguishes itself from misinformation, as disinformation is intended to deceive or manipulate.
The relevance of this topic can be seen in Figure 1. Here recent findings from the World Economic Forum's Global Risks Report 2024 are portrayed. The report highlights that misinformation and disinformation are perceived as the greatest short-term (2 years) global risk.

Dis- and misinformation can not only undermine democratic processes but also fuel societal polarisation and strengthen extremist narratives. This risk is even greater now, as in 2024 and 2025 nearly three billion people worldwide will have voted or will vote in national elections. After 2024, political polarisation became clearer with voters

Figure 1: Global risks ranked by severity over the short and long term (World Economic Forum, 2024)



Source
World Economic Forum Global Risks Perception Survey 2023-2024.

in over sixty countries expressing their frustration with the status quo, leading to shifts in leadership, protests, and heightened societal divisions (Wike et al., 2024). False narratives, especially those that are spread with malicious intent, can deepen ideological divides, legitimise hate speech, and incite violence. The recent events in the UK, where misinformation led to anti-immigration riots, illustrates this divide well (GNET Research, 2024).

Next to this, research shows that disinformation erodes trust in institutions and the government. In turn, low trust in these institutions correlates with a higher likelihood of engaging with and spreading disinformation. Thus, creating a downward spiral (Humprecht, 2023). As traditional media outlets and political institutions become delegitimised, individuals increasingly turn to alternative sources. Some of these sources promote extremist ideologies. This further deepens polarisation in society (Kutiyski et al., 2021).

Extremism

This brings up the next key topic: extremism. As described above, a polarised society is more susceptible to extremism, while extremism in turn deepens polarisation by reinforcing 'us vs. them' narratives. In this thesis, the focus is not on a specific extremist group or ideology, but rather on the generalised tactics used to spread disinformation. Lately, these tactics have been enabled by generative AI technologies.
In this thesis we will mainly look at the use of generative AI to recruit youth and to weaponise memes, called memetic warfare.

Generative Artificial Intelligence

The third key topic is Generative Artificial Intelligence. GenAI is a field of Artificial Intelligence that can generate, seemingly, new content such as text, images, audio, or video.

Since 2022, the release of OpenAI's ChatGPT[1], has democratised genAI and brought its capabilities to the public. An easy-to-use chat engine made it possible for users to generate new content by simply giving instructions to the chatbot. Many models followed suit and the technology is quickly revolutionising the way we work and communicate.

1 https://chatgpt.com/

Relevance to the Target Group

The target group for this thesis are 15-to-16-year-olds. This group is chosen because they are prone to fall for disinformation. This vulnerability is due to their curious nature, susceptibility to emotional language, and their strong presence online. At their age, they start being curious for the world around them and begin forming political opinions. For them, it is also important to find out where they are positioned in the world and develop a sense of belonging. This purpose is nowadays often found in online spaces.

It is through these platforms that extremists take prey on teenagers to lure in new supporters of their ideology. Through grooming, memetic warfare and targeted disinformation campaigns, they try to manipulate the minds of these teenagers to fit their worldview.

Focus on Building Resilience

The motivation for this thesis started when I was talking to Gwenda Nielen. She is the founder of TILT, an agency that builds innovative products and services to build information resilience. In this conversation I realised that building resilience and media literacy is one of the most effective ways to counter the effects of online disinformation. On the website of TILT the following quote can be found:

"The solution to disinformation is not so much about fact-checking, but about building peoples' skills and resilience." (Tilt, 2024)

I agree with this statement and believe in the value of the Bad News Game, as developed by TILT and Drog[1]. The game builds upon inoculation theory. The website of Inoculation Science formulates it as follows: "inoculation theory is a social psychological communication theory that explains how an attitude or belief can be protected against persuasion or influence in much the same way a body can be protected against disease". This theory will be used extensively in the final design and is an important enabler of the proposed solution.

2 https://drog.group/

1.3 Project Vision and Research Questions

Vision

The project vision is as follows:

"Helping young users recognise and critically assess AI-generated extremist disinformation. Theory and serious gameplay build resilience by unveiling the tactics of how disinformation is created and spread. Integrating gameplay into digital literacy education strengthens long-term awareness and promotes critical thinking about online content."

This is visualised in Figure 2.

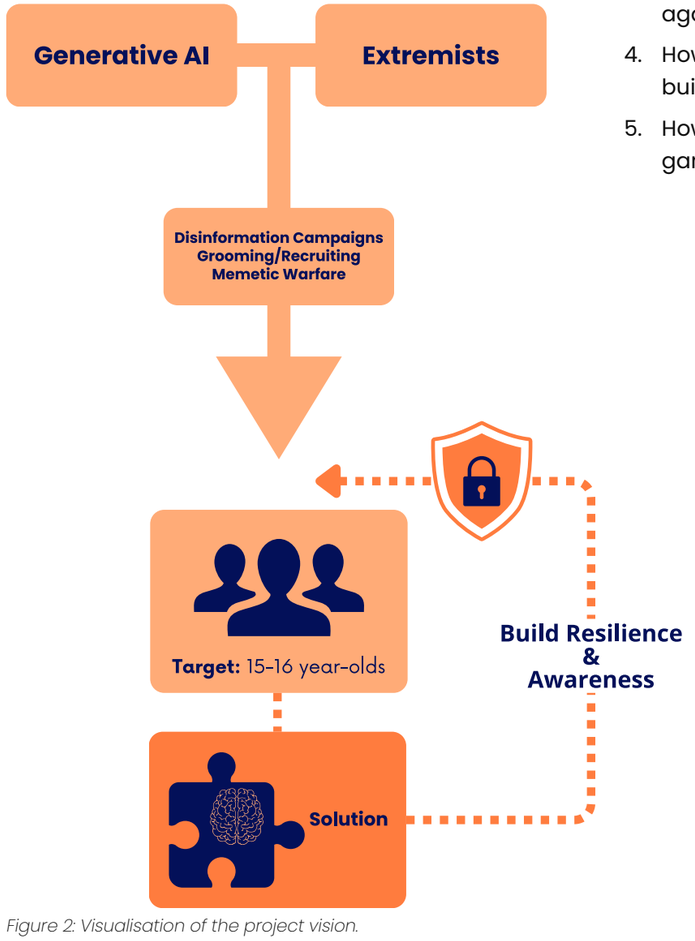


Figure 2: Visualisation of the project vision.

Research Questions

The following research questions guide this thesis, as well as the literature research unfolding in the next chapter. The questions ensure a structured and methodical approach to research and designing.

Main Question:

How can a serious game build the resilience of 15-16-year-olds against AI-generated extremist disinformation?

Sub questions:

1. Who are the extremists using generative AI maliciously, and how do they create disinformation to manipulate teens?
2. What are the societal impacts of AI-generated (extremist) disinformation?
3. What narratives and techniques are most effective in building resilience among teens against AI-generated disinformation?
4. How can a serious game be designed to build resilience among teens?
5. How can the effectiveness of a serious game be measured and validated?

1.4 Advancing the Field

Although many media literacy toolkits, games, and educational resources address disinformation, the rapid rise of generative AI has left the field without a solution specifically designed to respond to this new technological context. Some media literacy initiatives for generative AI exist, but they are rarely focused on younger audiences and do not address the specific threat of AI-generated extremist disinformation. In a time where both generative AI and polarisation through extremism are becoming increasingly widespread, this leaves a critical gap in the education of teenagers.

This thesis contributes to the field of digital media literacy and educational design for 15 to 16-year-olds by addressing that gap. Disinformation is a highly researched topic, and several tools exist to detect it. However, this thesis will focus on building resilience among people, not eliminating the presence of disinformation.

This thesis builds on the ideas from existing solutions such as the Bad News Game, which uses inoculation theory to help people recognise manipulation techniques. It extends on that approach by incorporating AI-generated content, extremist tactics, and a gamified learning experience that fits within an educational toolkit. By embedding the game within an educational toolkit, the project aims to reach more teenagers than a standalone online game. It also offers a scalable, responsible, and engaging way to help them recognise and build resistance against extremist disinformation, both today and in the future.

1.5 Thesis Structure & Design Process

Thesis Structure

This report reflects the journey of designing a solution that makes 15-16-year-olds more resilient to AI-generated extremist disinformation. It is structured as follows:

Starting with Chapter 2, the literature review includes an overview of generative AI, disinformation tactics, and the ways extremist groups exploit both. The target group is also defined more explicitly. It covers educational theories and existing media literacy solutions, which shaped the direction of the project. These insights are used to define the design requirements for the toolkit.

Based on that foundation, Chapter 3 introduces initial concept prototypes and outlines how the early design choices aimed to translate those theoretical insights into a serious game.

Chapters 4 and 5 describe user tests to validate or challenge assumptions. One user test was a guest lesson that took place in a Dutch high school. These chapters walk through the test setup, observations, and outcomes, highlighting what worked, what did not, and how students engaged with the topic. A second user test consisted of a general survey that was distributed among various age groups. This user feedback directly informed the second design iteration.

As a result, Chapter 6 presents a new design that slightly diverges from the original ideas presented in Chapter 3. In this chapter, the educational toolkit and final design are outlined.

With the final design in place, Chapter 7 looks at how the toolkit could be disseminated and maintained. It also explores practical considerations such as scalability and responsibilities for future updates.

Chapter 8 evaluates the overall concept based on feasibility, desirability, viability, ethics, and sustainability.

This is followed by Chapter 9, which outlines the project's main limitations. Chapters 10 and 11 provide a discussion and conclusion, along with key contributions to the field and recommendations for future development.

Finally, Chapter 12 contains a personal reflection on the process and what was learned along the way.

All supporting material is included in the appendices.

Design Process

The design process in this project mostly follows the Basic Design Cycle (BDC) by Roozenburg and Eekels (1998). This model is quite linear and consists of core reasoning steps that designers can go through. The process is iterative, which means the cycle will often be repeated multiple times through the project.

The process of this thesis is shown in Figure 3. As seen in the figure, the process is written down chronologically. It starts with the project brief and the literature review. In this phase, the problem space becomes clear and a direction for the project is chosen. The introduction was written later in the process. However, it was also added to this phase to provide the reader with the required background knowledge.

The first integration followed, which mainly involved simulating possible solutions. These simulations were then evaluated, and conclusions were drawn. For me, the decision phase is a moment where the next design steps are defined, by leaning on the conclusions made in the evaluation.

After this first BDC cycle, it was necessary to get valuable insights on the target group itself. This was done in two ways: through a general survey, and a guest lesson. Preparing for the guest lesson made me rethink how the serious game concept could be tested. Since there was no fully developed game yet, I designed a lesson that allowed students to experience the core ideas in a different way. A full lesson package, including theory and a slide deck, had to be created. I also made a simulation of the game concept. All of

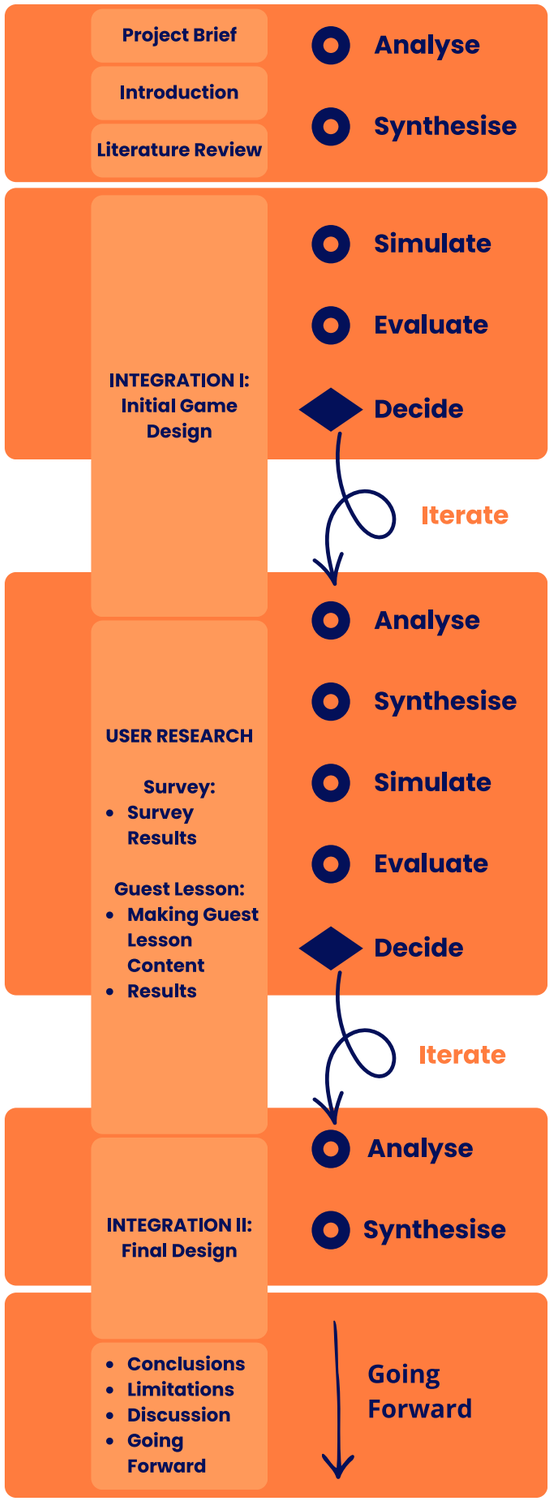


Figure 3: The design proces that was followed during this project. It follows the Basic Design Cycle, as presented by Roozenburg and Eekels (1998).

the insights gathered were evaluated afterwards, from my own observations and experience to what the students wrote down during the lesson.

This part of the process gave me valuable insights on the target group and a much clearer idea of what a good solution might look like. These findings were analysed, and a final synthesis step followed, which resulted in the final direction and concept design for this thesis. This design can be found in the second integration part, in chapter 6.

After Chapter 6, additional steps were taken to further support and refine the design. Earlier decisions were evaluated and strengthened, but the design was not finalised or tested again in a full validation process. Plans for future testing are included. What follows are the conclusions and a reflection on the project.

Literature Review

2. Literature Review

To start this thesis off with a solid foundation, this literature review is written. The exploration of the subject, as introduced in the introduction, will be further broken down here using supporting sources. This exploration leads to a first concept design, which will be explored in *Section 3: Initial Game Design.*

This literature review starts off with a breakdown of generative AI, including a short introduction of the technology behind it, its potential, and its growing place in society. In the second subchapter Disinformation and Extremism will be highlighted, examining how disinformation is used by extremists to shape narratives and influence public perception. It will also give a general exploration of the identities of these extremists groups and which tactics they use to attempt to reshape society. Subsequently, the next chapter will examine how extremists utilise AI for spreading disinformation. Followed by the broader societal impact of the growing availability of genAI. Highlighting the relevance and importance of the subject. Next, the target group will be looked at, examining what makes this target group susceptible for disinformation. It will also explore how a solution designed to build resilience can be made engaging and effective for this target group. The next chapter will show examples of this with an overview of media literacy approaches. The serious game will be highlighted along with other media literacy approaches and educational tools. The second-to-last chapter will explore some existing media literacy approaches. Finally, conclusions will be drawn, providing insights that will be used in the next section of the thesis.

2.1 Generative Artificial Intelligence

Innovation of Artificial Intelligence

Even though Artificial Intelligence (AI) has been studied for a considerable period, widespread recognition and public attention mostly started with the release of OpenAI's ChatGPT (IBM, 2025). This is a generative AI (GenAI) Large Language Model (LLM) designed as a chatbot that engages in natural language conversations (ChatGPT, 2025). The development of accessible LLM's gained traction and transformed AI into a prominent buzzword. Numerous companies are searching for ways to integrate AI into their business strategies. Investors are putting billions into innovating better (generative) AI. Educational institutions are struggling to implement it in teaching or to set up guidelines for its use. The amount of searches for "generative AI" shows the buzz around the topic, as can be seen in the Google search trends, as pictured in Figure 4.

Although this might appear to be a sudden emergence, AI is a field that has a substantial history of research and development and expands further than the, publically well-known, field of generative AI.

Artificial Intelligence, in the broader definition, encompasses technology that enables computers and machines to simulate human learning, comprehension, problem-solving, decision-making, creativity and autonomy (IBM, 2025). Applications using this technology might be able to process visual information, understand and respond to human language, do data analysis and even act independently.

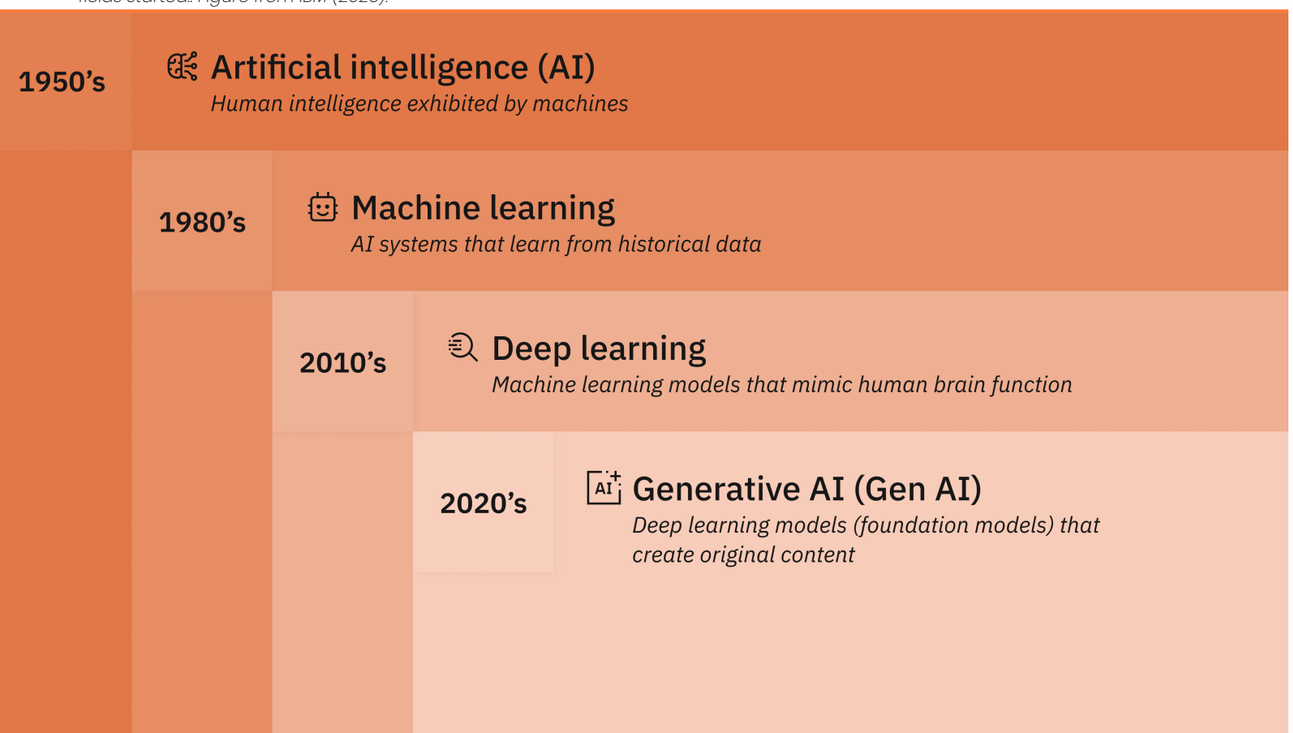
Figure 4: Worldwide Google Search Trends for the search "generative AI" during the timespan of Dec 2021 - Mar 2025 (Google, 2025). The Y-axis represents the amount of searches, the X-axis is the timeline.

As underlined in Figure 5 on the next page, the field of AI has its origins in the 1950's. One of the key questions during this period was the question if machines are capable of (human-like) thinking. This was explored by Alan Turing, who was considered to be the 'father of theoretical computer science' (Beavers, 2013). From the 1980's, Machine Learning (ML) emerged as subfield of AI. This subfield focuses on enabling computers to learn from data without explicit programming (IBM, 2025). Algorithms could then be trained to make predictions or decisions based on data. From the 2010's onwards, a rapid advancement of deep learning started. Deep learning, being a subfield of ML, utilises neural networks with multiple layers to analyse complex data. A well-known example of this was Google's AlphaGo (Google DeepMind, 2025), which demonstrated the possibilities of deep learning by defeating a world champion in a complex Chinese game called Go (Alalaq, 2025). The developments in these subfields paved the way for the breakthroughs in Generative AI in the 2020's. Generative AI is the main subject of AI that will be talked about in this thesis.

Generative AI

Generative AI (GenAI) involves deep learning models that have the capability to create original content, such as long-form text, images, realistic video, or audio, based on an instruction that is given by the user (IBM, 2025). These instructions are called prompts. This seemingly new content is generated by altering simplified representations of their training data, which can sometimes resemble the original data (Feuerriegel, 2024). Well-known examples include ChatGPT as a chatbot and Dall-E for generating Images. However, countless models exist that are integrated into existing tools. For example: Adobe Photoshop, Notion, and AI integration in Microsoft apps, such as Excel. Additionally, open-source models are available for public use. Often, these models provide the user with great adaptability to set up for specific use. This enables innovation, as well as making it easier for malicious users to access and modify those technologies. This will be highlighted and further explained in *chapter 2.3* of this literature review.

Figure 5: How artificial intelligence, machine learning, deep learning and generative AI are related and when the research in these fields started.. Figure from IBM (2025).



2.2 Disinformation

What is Disinformation?

Disinformation distinguishes itself from misinformation by the intent of its maker. The Cybersecurity & Infrastructure Security Agency (CISA), defines disinformation as "information deliberately created to mislead, harm, or manipulate a person, social group, organisation, or country (Vlachos, 2022). Where misinformation is defined as "merely" false information.

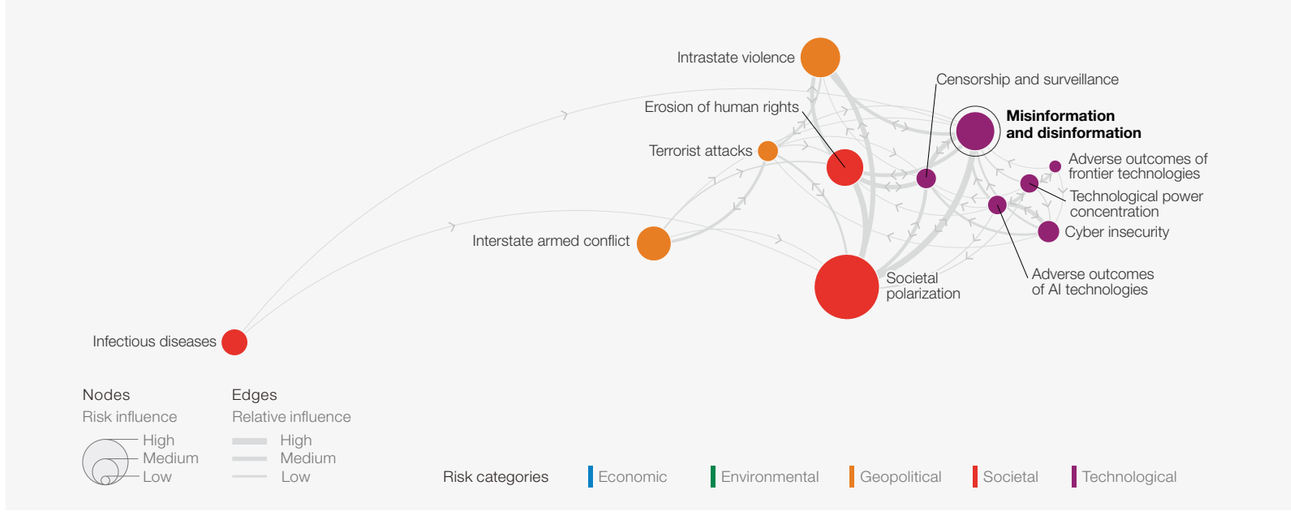


Figure 6: This figure shows how mis- and disinformation are linked to other societal risks. It highlights how disinformation can both induce and be induced by societal polarisation and violence. An important connection for this thesis is the causal link between adverse outcomes of AI technologies and the spread of mis- and disinformation, showing how emerging technologies can worsen information threats. The figure is taken from the WEF Global Risk Report (2024).

The harmful intent makes that this information is used to manipulate groups in alignment with the goals of the maker (McKay & Tenove, 2021). For this thesis the targeted group would be manipulated by Extremists using genAI.

Although disinformation has existed throughout history, it is particularly relevant today. Online Spaces, such as social networks, can spread information rapidly. Making them ideal for malicious users wanting to disseminate disinformation. Examples are the manipulation of elections, anti-immigration narratives or campaigns during the COVID-19 pandemic (McKay & Tenove, 2021)(Humprecht, 2023). The latter leading to lowering social compliance to public health guidelines, such as mask wearing and social distancing.

This makes that online disinformation is being seen as "the defining political communication topic of our time (Freelon and Wells, 2020)", because of its threat that it poses to democracy. The World Economic Forum (2024) identified mis- and disinformation as the most significant short-term global risk for the next two years. In Figure 6 an image from this report can be seen, showing interconnectedness with other risks. A notable risk is societal polarisation, and how it influences the use and outcomes of AI-technologies. Both key subjects in this thesis.

How does disinformation work?

Disinformation works through various techniques and mechanisms, often leveraging mixtures of fact and fabricated content (Wardle, 2022). Manipulated images, videos, and false sources help convey misleading messages. Additionally, the place where disinformation is spread plays a crucial role (Marwick & Lewis, 2017).

Combining the work from Claire Wardle (2017, 2022) and Marwick & Lewis (2017), we can identify techniques and guidelines for fabricating successful disinformation messages. The info in these reports is narrowed down into three key points:

- *Blend truth and falsehoods:* Genuine sources can be used or content can be cherry-picked and framed miseleadingly. Impersonation of credible sources is another common tactic. Blending truths and falsehoods makes for a more subtle message that is far more successful in terms of persuading and engaging people. Fabricating falsehoods can be done by creating new, original content as well. In this thesis the central topic is the creation of false content through genAI.
- *Attention Hacking:* Using sensational or emotionally charged content, memes or strong-language to capture attention and increase (online) visibility.
- *Exploit vulnerabilities in (Social) Media:* Targeting weaknesses in the media

ecosystem such as low public trust and algorithms feeding on sensationalism. Bots can be used to gather online attention, or people can be targeted based on their online activity.

Disinformation and Social Media

The rise and use of social media platforms in society nowadays, has disrupted traditional media gatekeeping (McKay & Tenove, 2021). It has partially replaced journalists and information dissemination has changed. Algorithms determine content visibility and reach. Often, engagement is prioritised above accuracy. This allows for the amplification of disinformation and other polarising content. Disinformation campaigns can exploit the algorithms and make use of the widespread reach that social media enables.

Use of Disinformation

The type of disinformation is dependent on the goal and intention of the maker. Often, groups, institutions and governments work with disinformation campaigns (Wardle & Derakhshan, 2017). These have a predetermined goal, often manipulating public opinion, and work methodically. However, a lone actor can also make use of disinformation. The scale and use of disinformation can vary from a spam email to a global campaign.

As stated before, we can distinguish three motivations for creating and spreading disinformation (Marwick & Lewis, 2017):

- Ideology (or politics)
- Money
- Status and/or attention

Disinformation can therefore be used by anyone. State actors use it for political gain. Black hat hackers use it for personal gain, money, or status. Influencers can use it for attention or earning money. Extremists use it to spread their narrative and ideology. Wardle & Derakhshan (2017) visualised the different elements of 'information disorder'. A remake of the figure, as they use it, can be seen in figure 7. Defining the "agent", "message", and "interpreter" can give a better understanding of disinformation and why it is made.

Extremism and Disinformation

In this thesis disinformation is looked at through the lens of extremist fabrication. Extremists strategically use disinformation to spread their ideological narratives. The definition of extremism, and how they use disinformation will be looked at in the following subchapter.

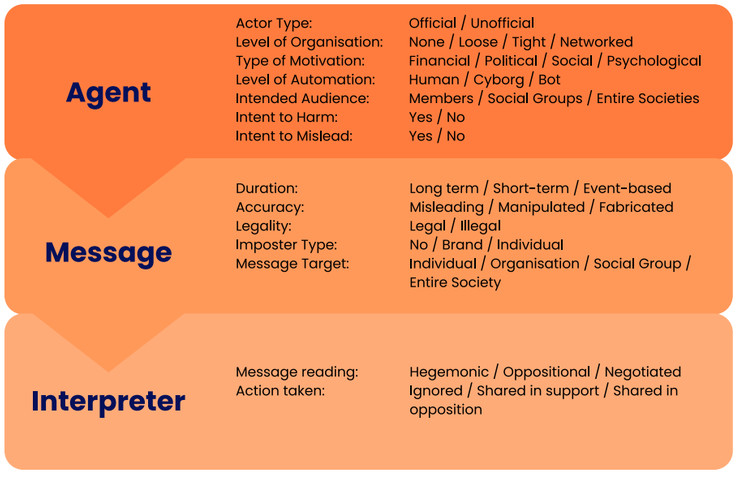


Figure 7: Model of information disorder as proposed by Wardle and Derakhshan (2017). The figure shows three elements involved in the creation and spread of mis-, dis-, and malinformation: the agent (the creator and distributor of the information), the message (the characteristics and framing of the content), and the interpreter (the audience and their reaction). This model helps to systematically analyse information by considering who created it, how it is framed, and how it may influence its audience.

2.3 Extremism

What is Extremism?

Extremism has many different definitions. One definition by Save the Children Finland (2021) is as follows: "Extremism is broadly defined as a way of thinking that is stark, absolute, and black and white. Typically involving stiff and uncompromising views on right and wrong, leading to generalising and a strong separation of groups into an "us" and "them".

This leads to overly simplifying complex problems or constructs in the world to easily understandable explanations. Extremists are frequently unhappy with the world as it is and search for radical changes. These can be changes in the political sphere, religion, or society. Extremist beliefs are often substantially different from the 'mainstream' beliefs or standard practices.

Types of Extremism

Different forms of extremism exist. First, it is important to differentiate between violent and non-violent extremism. It is entirely possible for a group or movement to go from a non-violent group to a violent group or vice versa. Violent extremism involves the use or threat of violent acts to pursue change in the world according to their ideology. Non-violent extremism, on the other hand, holds these extreme beliefs and advocates for radical changes, without the use, or intent to use, violence (NCTV, 2024). Although the intent is to be non-violent, their actions may induce violence.

Since extremism is a general term describing various radical worldviews, it is hard to categorise. To simplify different groups into broad categories, the following distinctions can be made:

- **Right-wing extremism:** Typically focused on racial, ethnic, or nationalist narratives. Groups in this category include neo-nazis, white supremacists, anti-government militias and individuals driven by a strong bias regarding religion, gender, sexuality or immigration (Williams & Evans, 2021)(Doering et al., 2023) (Sterkenburg, 2021).

- **Left-wing extremism:** Less frequently highlighted. Representing radical views and actions emerging from far left political ideologies. Examples include anti-capitalist and communist movements (Bundesamt für Verfassungsschutz, n.d.).

- **Religious extremism:** Stemming from religious teachings. Often resulting in intolerance or violence towards individuals with different beliefs (McNeil et al., 2019).

- **Single-issue extremism:** Putting one central topic or issue as focus (Ackerman & Kouloganes, 2019). An example could be an anti-abortion extremist.

Extremist use of Online Spaces

Extremism operates both in online and offline spaces. Although in-person recruitment and gathering remains important, the internet provides a low-cost mechanism for extremists to extend their network and gather (financial) aid. Online, they can recruit new members, share knowledge and coordinate actions (Save the Children Finland, 2021). Social media and online networks also make it possible to target specific groups that might be more susceptible to extremist ideas. The possibility to remain anonymous in online spaces leads to the proliferation of illegal or shocking content (Williams & Evans, 2021). Extremist content can be found on multiple platforms but is mainly prevalent on messaging forums, bulletin boards, and social networking platforms. There is even the possibility to set up dedicated platforms

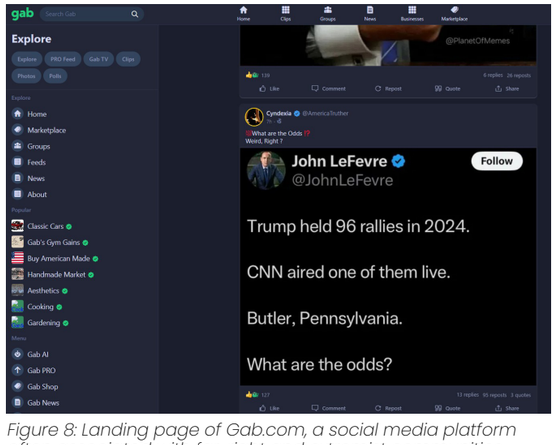


Figure 8: Landing page of Gab.com, a social media platform often associated with far-right and extremist communities. The highlighted post spreads disinformation about the assassination attempt on Donald Trump in Butler, Pennsylvania, aiming to create an "us versus them" narrative.

to talk to like-minded individuals. Gab is one of these far-right, and often extremist, social media networks, that was set up under the realm of 'free-speech'. In Figure 8, the dashboard of Gab can be seen, resembling any other social media platform. Within 1 minute of exploring this dashboard, multiple posts containing misinformation, were found. The claim on figure 8 is disinformation framed to evoke suspicion about the assassination attempt on Donald Trump in Butler, Pennsylvania. It attempts to form an "us" vs "them" narrative.

A news article titled "Gab's Racist AI Chatbots Have Been Instructed to Deny the Holocaust" from Wired (2024) was one of the motivations for this thesis. It clearly demonstrates how extremists use new technologies to spread disinformation to radicalise individuals.

Extremism and Disinformation

The spread of disinformation can create a polarised society and fuel hate speech, pushing public rhetoric towards extremes. Disinformation serves as a tool to gain group support, justify their (violent) actions, and undermine opposing views.

As mentioned in the previous chapter, disinformation plays a significant role in extremist tactics. Manipulative narratives are used to recruit new members. Often combined with carrying out a strong "us" vs "them" narrative. This is done by spreading disinformation about opponents or events happening in the world. Frequently, exploiting existing social divisions and targeting vulnerable groups (Szakacs & Bognar, 2021). Young people, searching for their identity and a sense of belonging, are one of those vulnerable targets. The extremist way of phrasing complex issues to simple truths or stark categorisations is appealing to many, children and adults alike (Save The Children Finland, 2021).

Just like other malicious actors in online spaces, extremist groups also utilise new technologies to spread their narrative more effectively, using them to micro-target individuals or broaden their influence in society (Colomina et al., 2021).

The next chapter will explore their use of new technologies to spread disinformation, focusing particularly on Generative AI.

2.4 GenAI and Extremist Disinformation

Just as with the adoption of internet, extremists make good use of upcoming technologies. One of the big innovations of this decennia is generative AI. Extremists leverage this technology to reach a bigger audience and to automate the spreading of their ideology, as well as generating high quality content.

In this chapter, an explanation will be given about the content extremists create using generative AI, as well as how they leverage the technology to spread disinformation.

What is (generative) AI used for?

Extremist groups use AI in multiple ways, from disinformation to building websites or apps. For this thesis, the focus will lie on the different strategies used by extremists to spread disinformation and widen their network through recruitment. To simplify the scope, extremist use of AI can be broken down into three components:

Recruitment and Radicalisation: LLM's can generate highly personalised and persuasive content. AI powered chatbots could also automate one-to-one interactions with potential targets for manipulation/recruitment (Janjeva et al., 2024)(Weimann et al., 2024).

Micro-targeting disinformation: AI driven algorithms excel at analysing large datasets to identify who might be susceptible to certain narratives. Using this, extremists can target disinformation campaigns effectively, based on demographics, beliefs or vulnerabilities (Stefan, 2024)(Janjeva et al., 2024)

Disinformation Campaigns content: Many forms of content make up a (successful) disinformation campaign. The following are a three forms relevant to this thesis:

- Deepfakes: Highly realistic images or videos impersonating public figures, that do things they have never actually done. Extremists can spread disinformation that seems authentic to the untrained eye (Stefan, 2024). This links closely with 'Impersonating Sources'. However, deepfakes can also be used to give a public face to fabricated news shows or interviews.
- Memetic Warfare: This is the creation and dissemination of emotionally charged or viral content through memes or visuals. GenAI tools such as image and videos generators are used for imagery, and LLM's are used for captions (NCTV, 2024)
- Impersonating Sources: AI can impersonate sources by copying writing styles or the look and feel of certain content. AI generated deepfakes can impersonate experts or journalists to benefit from the trust they have with the public. It can also create fabricated documents or texts to mislead audiences (Ferrara, 2024).

For this thesis, the third component, focusing on disinformation campaigns, is most relevant because generative AI significantly simplifies the creation of convincing disinformation content, making it easily accessible to anyone with a computer and internet access.

How is generative AI used for disinformation?

While other AI applications, such as sentiment analysis to gauge the public reaction and the use of bot networks to amplify messages, are relevant in spreading disinformation, the primary focus of this thesis is about how generative AI is used for generating disinformation content. For generating disinformation content, the following subfields of generative AI are important:

Large Language Models (LLM's): Models that can generate human-like text. LLM's can produce persuasive texts or disinformation narratives that are seemingly authentic and human-made, at a large scale (Ferrara, 2024).

Text-to-Image/video tools: AI models that can create original visual content from a textual description given by the user. Extremists can use it to make propaganda posters, memes or fake images (Stefan, 2024)

Text to-speech tools: Together with text-to-image tools, video deepfakes can be made. Voices can be impersonated and made to say disinformation, while they are seemingly authentic (Janjeva et al., 2024).

LLM's, Text to Image and Speech tools are found in current popular genAI chatbots, such as ChatGPT and Gemini.

How these tools are used to fuel the tactics as mentioned before, can be seen in figure 9. The audience sentiment tools and social media bots are mentioned to give a more complete view.

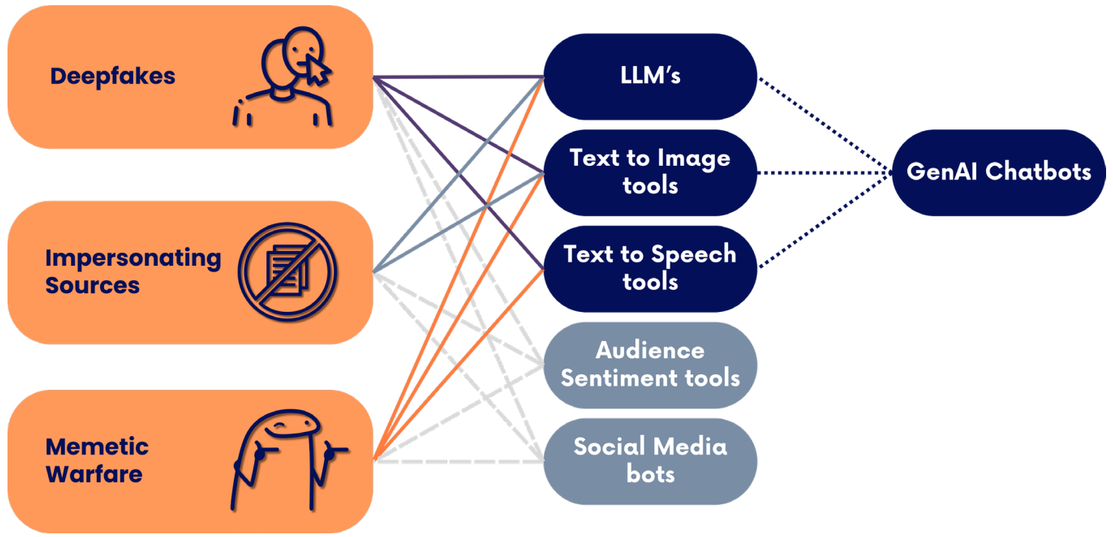


Figure 9: Overview of how extremist disinformation tactics can be powered by generative AI tools. Deepfakes, impersonating sources, and memetic warfare are enabled by technologies such as LLMs, text-to-image, and text-to-speech tools. GenAI chatbots combine several of these capabilities, making them powerful tools for generating disinformation. Audience sentiment analysis and social media bots are included to show how disinformation can be further amplified after creation.

2.5 Target Group

The Age Group

The target group for this thesis is 15-16-year-olds. Teenagers at the end of puberty are actively forming their identities and seeking a sense of belonging (Anderson et al., 2023). Adolescents are often highly alert on how they are perceived by peers and are therefore socially driven and emotionally sensitive. This is also the age at which they begin to explore their own values and beliefs, which is closely tied to finding spaces where they feel accepted. The search for a sense of belonging is why this age group is highly susceptible to extremist narratives and recruitment and grooming campaigns (Save the Children Finland, 2021).

Online behaviour

This age group grew up with smartphones and is therefore deeply embedded in social media platforms such as TikTok, Instagram and Snapchat. They spend a significant amount of time online and use these spaces to express themselves and socialise with their peers (Anderson et al., 2023). These platforms also serve as environments where they where they explore new ideas, groups, and people, often through memes and other viral content (Marquez et al., 2023). Memes are a key form of online communication and frequently reflect current events or social commentary. Though often subtle, each meme conveys its own message (NCTV, 2024). Algorithms on online platforms play a significant role in shaping their worldview through curating content specific to the individual. This can make it difficult to distinguish truth from fiction, especially since the rise of (generative) AI.

How does the target group learn?

Although learning preferences are highly personal, this age group tends to prefer visual, interactive and socially contextual learning experiences over traditional methods. Exploration and continuous feedback helps maintain their interest. Technological tools are used for online learning, and it is also their preference to integrate this in education. Content on YouTube and social media platforms introduce this group to new topics and it is used by them as an addition to traditional education. Collaborative learning approaches and gamification can also increase their engagement (Paulina & Ernawati, 2022). This is in line with their strong online presence and their socially driven group mentality.

In conclusion, their developmental phase, that is characterised by seeking for identity and social approval, combined with their strong online presence, makes 15-16-year-olds vulnerable to AI-driven disinformation. These developments highlight a need for media literacy education to improve their resilience against malicious use of online spaces. Next chapter will explore the landscape of media literacy education.

15-16-year-olds as target

As discussed in the previous paragraphs, multiple factors contribute to the vulnerability of this specific target group: 15-16-year-old teenagers. In summary, this age group is still developing and actively seeking a sense of belonging. Adolescents experiment with finding their identity and begin to form opinions about the world around them. This process increasingly takes place online. In these digital environments, emotionally charged content is prevalent, and peer validation plays a significant role. The prefontal-cortex, which is responsible for risk assessment and critical reasoning, is not yet fully developed, making this age group more susceptible to emotionally driven and seemingly exciting content.

Emerging risks of youth radicalisation

Both the Dutch General Intelligence and Security Agency (AIVD) and the British domestic intelligence agency (MI5) have raised their concerns about the increasing involvement of minors to extremist movements. In the AIVD (2025) publication "Web van Haat (Web of Hate)" a warning is given that a growing number of teens are consuming, producing and disseminating extremist propaganda online. As well as actively participating to ideological ecosystems such as jihadism and right-wing extremists or terrorists. This happens on platforms such as TikTok, Discord, Instagram, and Telegram.

MI5 has the same concerns. MI5 Director-general Ken McCallum reported a threefold increase in the number of minors that were being investigated for extremist activity in the UK (MI5, 2024). He highlights that extremist movements are deliberately targeting young people in online spaces. A growing number of the counter-terrorism caseload in the UK now involves teenagers.

Together, these statements show a trend where extremist groups are exploiting minors who are seeking their identity and their use of online spaces, to ultimately radicalise this new generation. This trends raises the urgency to find a solution that builds teenager resilience against these harmful extremist narratives.

Example cases

To show how harmful the rise of extremism (among minors) is, a handful of news headlines will be shown that highlight cases of extremist recruitment that radicalised minors.

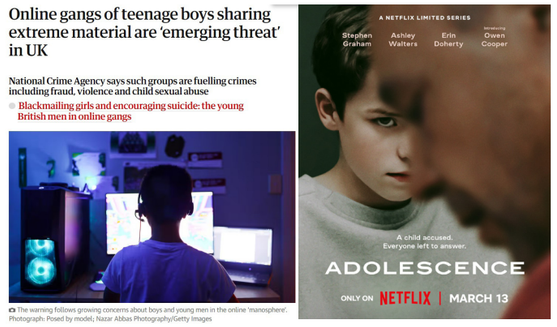


Figure 10: Headline (left) of a The Guardian article by Rachel Hall (2025) and the promotional poster of Adolescence a series on Netflix (right).

The article by Rachel Hall (2025), published in The Guardian (see Figure 10), reveals how teenage boys are forming online networks to exchange violent, misogynistic, and extremist content, sometimes aided by genAI tools. This reality is depicted in the Netflix serie Adolescence, whose promotional poster is also included in Figure 10. The series focuses on the aftermath of a teenager's crime, revealing how unsupervised online spaces expose youth to extremist narratives.

Figure 11: Headline of a The Guardian article by Josh Taylor (2024).

The Guardian journalist Josh Taylor (2024) reports that Australia's intelligence chief has warned generative AI is likely to intensify online radicalisation, particularly among young people. This is due to the ease with which extremist content can now be produced and the increasing difficulty of detecting it. This raises concerns about how these technologies might be exploited before adequate safeguards are in place. The article headline can be seen in Figure 11.

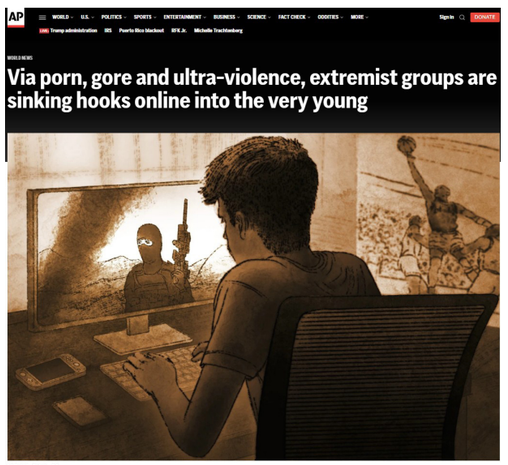


Figure 12: Headline of an AP article by John Leicester (2025).

The AP article (see figure 12) by John Leicester (2025), presents a case study of a 12-year-old boy in France who was rapidly radicalised through exposure to violent jihadist content online.

While centred on Islamist extremism, the piece highlights broader concerns about how young people can be drawn into extremist ideologies through unmoderated digital spaces.

Figure 13: Headline of a BBC article by Daniel De Simone and Ali Winston (2023).

In figure 13, the heading of the BBC article by Daniel De Simone and Ali Winston (2023) can be seen. It investigates the case of Rhianan Rudd, a 15-year-old girl groomed online by far-right extremists, who adopted racist and antisemitic beliefs and downloaded a bomb-making manual. Her case is part of a growing trend of young people being radicalised online, and it raises concerns about how authorities can respond to exploited minors facing terrorism charges, especially after Rudd took her own life before prosecution.

In these articles you can find examples of teenagers being radicalised, groomed, and influenced online. This is notably also being accelerated by the rise of AI-generated extremist content, which lowers the barrier to creating and spreading harmful narratives. The consequences include youth violence, suicides, and a skewed worldview in young people. These cases fit within the broader context of this thesis, which explores how online spaces and (AI-generated) extremist content can distort the way youth see and interact with the world around them.

2.6 Digital Media Literacy Education

Digital Media literacy education revolves around teaching people how to navigate the internet and media in online spaces. A significant portion of this education is aimed at children and teens. This education can be delivered in many forms, using various techniques, and the target group can be reached through a wide range of platforms. In this chapter, the field of media literacy will be explained and simplified into an framework, which can be seen in figure 14. In addition, a gap in modern media literacy will be outlined. This thesis aims to address that gap by proposing a media literacy solution focused on building resilience against AI-generated extremist disinformation.

Tactics - Theory behind the lesson

There are several tactics to effectively create (digital) media literacy. For this thesis we will limit ourselves to three tactics: Fact-checking, inoculation techniques, and critical evaluation.

Fact-checking skills can be strengthened by teaching individuals how to verify information using credible sources, reading techniques, and tools such as reverse image searching (Dame Adjin-Tettey, 2022). Furthermore, inoculation and prebunking are about familiarising individuals with common manipulation techniques. Similar to medical inoculation, individuals are exposed to weakened doses of misinformation to build resistance against future persuasion attempts. Inoculation is often taught through letting individuals experience the making of mis-/disinformation. Inoculation has been coined by McGuire (1961) as a psychological theory that helps people resist persuasion by exposing them to weaker versions of misleading arguments and showing how to refute them. Prebunking focuses more on familiarising individuals with tactics rather than exposing them to dis-/misinformation. By critically thinking about how to craft disinformation, they learn about how they could be manipulated, thus creating some resilience against it (Basol et al., 2020). Lastly, critically evaluating media, though similar, focuses more on reflection and judgement. This is also a more broad tactic, whereas inoculation and fact-checking are used against mis- or disinformation. Critical evaluation is all about building analytical competencies to make informed judgements about information they might encounter online or in the media (Orhan & Ay, 2022).

Medium - How to convey the theory

The medium through which information is conveyed, plays a significant role in its effectiveness. The effectiveness also depends on the preferences of the target group. Overall, engaging and interactive methods are frequently highlighted in literature (Feltrero et al., 2023) (Cernicova-Buca & Ciurel, 2022). Especially for younger target groups, such as the one for this thesis. Serious games are a good example. These games offer an immersive environment where individuals learn media literacy skills through gameplay. These games can implement various forms of educational content within game mechanics, leading to better engagement and knowledge retention (Chang et al., 2020). Multimedia applications, including videos, imagery, exercises and infographics can also be valuable tools to strengthen explanation or automate learning experiences. Lastly, theory and frameworks are essential for a more comprehensive and deeper understanding of a subject.

Channel - Reaching the target group

Reaching the target group with the media literacy solution can be done through various channels. One of those channels is to integrate courses into mainstream education curricula. This requires teachers to keep up to date with the evolving media landscape and for curricula to be revised consistently. Another option is to teach media literacy through community-based initiatives. Here certain members act as 'information gatekeepers' and disseminate knowledge through the population. This is particularly effective for reaching vulnerable populations. Finally, public awareness campaigns can be launched. This can be done by using digital platforms, mainstream media such as newspapers or magazines, and advertisements (Dame Adjin-Tettey, 2022).

Figure 14: Overview of key dimensions in media literacy education. The framework simplifies the field into three levels: tactics (the theoretical approach), mediums (how the theory is conveyed), and channels (how the target group is reached). This framework is used to structure the media literacy solution proposed in this thesis.

Notable solutions

To show what the different approaches can look like, this subchapter will go through three examples.

Starting off with the Bad News Game[1] (2018), seen in Figure 15. This is a serious game designed to teach ages 14 and up about fake news. The game uses inoculation theory, by letting users experience how you can become someone that spreads fake news. It has been launched as an online game, and there is also a educational toolkit available.

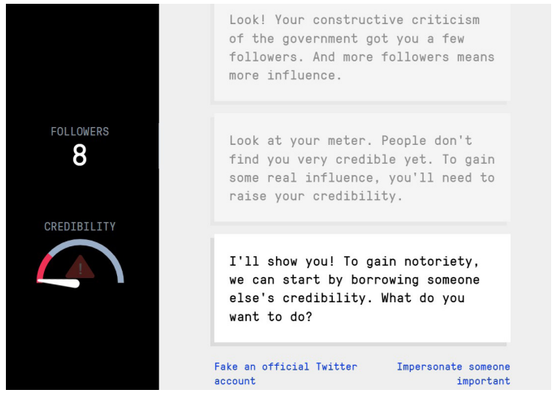


Figure 15: A screenshot of the Bad News game gameplay. The branching story structure is visible through the two options players can choose from to influence the narrative.

This game has always been one of the starting points for this thesis. The game effectively implements gamified learning in a simple form. Players take on the role of a malicious actor who wants to create chaos through different tactics. These tactics include impersonation, emotional content, polarisation, conspiracy, discrediting, and trolling. All tactics are introduced in separate "focus levels" and are not presented all at once.

The Bad News Game is grounded in inoculation theory (McGuire, 1961). This theory works by exposing people to weakened forms of manipulation in a controlled environment. It is also achieved by letting the user take on the role of the manipulator. In the Bad News Game, the player becomes the malicious actor to gain first-hand insight into how the tactics work. This helps players recognise these tactics when they are used in real life. Studies have shown that this approach builds psychological resilience and improves the recognition of manipulative content (Roozenbeek & van der Linden, 2019; Basol et al., 2020).

1 https://www.getbadnews.com/

A notable feature is the slightly absurd and fictional storyline. The scenarios mirror real disinformation strategies without directly referencing actual (political) events. This creates emotional distance, which is especially suitable for younger players to avoid causing distress.

The game demonstrates how simple game mechanics, branching narratives, and theory can work together to form an effective media literacy solution that is also engaging.

Figure 16: The landing page of Nieuws in de Klas (Nieuws in de klas, n.d.).

Nieuws in de klas (Figure 16), which literally translates to "News in the classroom" is an organisation that makes educational toolkits for teachers to use in their lessons. The toolkits exist of recent news snippets together with exercises. Therefore, making use of multimedia and theory and frameworks.

Figure 17: The poster of the 'Think before you share' Campaign of the European Union (EUvsDisinfo, 2022).

'Think before you share' is a public awareness campaign from the European Union. It advocates responsible sharing to combat misinformation. It uses fact-checking tactics alongside critical evaluation, and is spread through multimedia content. The campaign poster can be seen in Figure 17.

When you map these solutions in the infographic we used before, it shows how media literacy solutions come in different forms. This can be seen in figure 18.

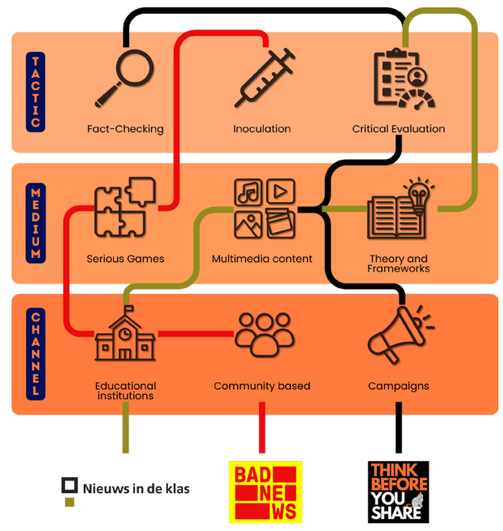


Figure 18: Mapping of existing media literacy solutions onto the Tactic–Medium–Channel framework of Figure 14. The figure shows how different initiatives follow different paths to address disinformation, based on their focus, delivery method, and target audience.

Gap in solutions

Although there are many media literacy education toolkits, games and resources addressing disinformation, the quick rise of genAI has left the field without a solution specifically geared towards this technology. While some genAI media literacy courses do exist, none are specifically designed for a young target group, nor specifically targeted to extremist disinformation. In the current time where generative AI and polarisation through extremism are increasingly prevalent, this leaves an important gap to fill.

2.7 Conclusions of Literature Review

The Literature review looked at different parts that shape the problem space of this thesis, starting with an overview of the key elements. Generative AI was defined, explained how it works, and how accessible it currently is. Then, disinformation was defined and examined in combination with extremism and generative AI. The trend of extremists using online spaces and new technology to their advantage has been made clear. Some common practices, such as using deepfakes, memes and impersonating sources have been explained as problematic for young people.

The review then looked into this young target group of 15-16-year-olds, highlighting their emotional and social development, strong online presence and vulnerability to extremist disinformation. Finally, media literacy approaches were outlined and examples were shown. Although there are many good media literacy approaches addressing disinformation, none focus specifically on the intersection of generative AI and extremist disinformation for this age group.

In the current timeline, where both generative AI and polarisation are becoming more visible and dangerous, making a solution that adresses these phenomena becomes more urgent. The next chapter will build on these insights and propose a first attempt at a serious game to stimulate teen resilience against AI-generated extremist disinformation.

Integration 1: Initial Game Design

3. Initial Game Design

This chapter will show the first iteration of designing a serious game. The insights that are gathered in the literature review, will be translated into a serious game. First there will be a recap of the design space. Further defining who the target group is and what their struggles are.
To make sure the game has a strong foundation, learning objectives are created using Bloom's Taxonomy. Translating these objectives into a branching story serious game will be done by using a 'mechanics, dynamics, aesthetics' (MDA) framework.

This forms the foundation on which the first two prototypes are built. The first is a low-fidelity prototype using ChatGPT's feature to create a custom chatbot environment. The second prototype is made using Twine, a tool for creating branching story games.

3.1 Design Space

Recap

The goal is to build resilience against manipulation techniques used in **AI-generated extremist disinformation**.

The target group are **15-16-year-olds**.

Serious games are suited for this task because they can easily incorporate inoculation techniques and offer an **interactive** way of learning. They also allow for a **safe exploration** of the subject.

Persona's

No 15-16-year-old is the same. Therefore, the game should bring value to a wide range of personalities and skill-levels. While almost all teens use social media and the internet extensively, each has their own focus. To make sure I do not not design the game only from my own knowledge or for a particular personality, I set up a few persona's. These help identify where teens' knowledge might fall short, or where they would be more engaged. The three persona's, each with a quote that captures their perspective, are shown in the figures on the right.

Elias (16):
"Humour is a good way to look at the world, in my opinion. Through memes, I can make sense of things and share that perspective with others."

Amara (16):
"I find so much of myself online; connecting with influencers, sharing my life, and seeing what is possible. It's like a vision of the life I want to create."

Alex (15):
"When I'm online, I can dive into all kinds of fascinating things; exploring, discovering what I like, and finding communities that really fit me. It is like a world full of endless possibilities."

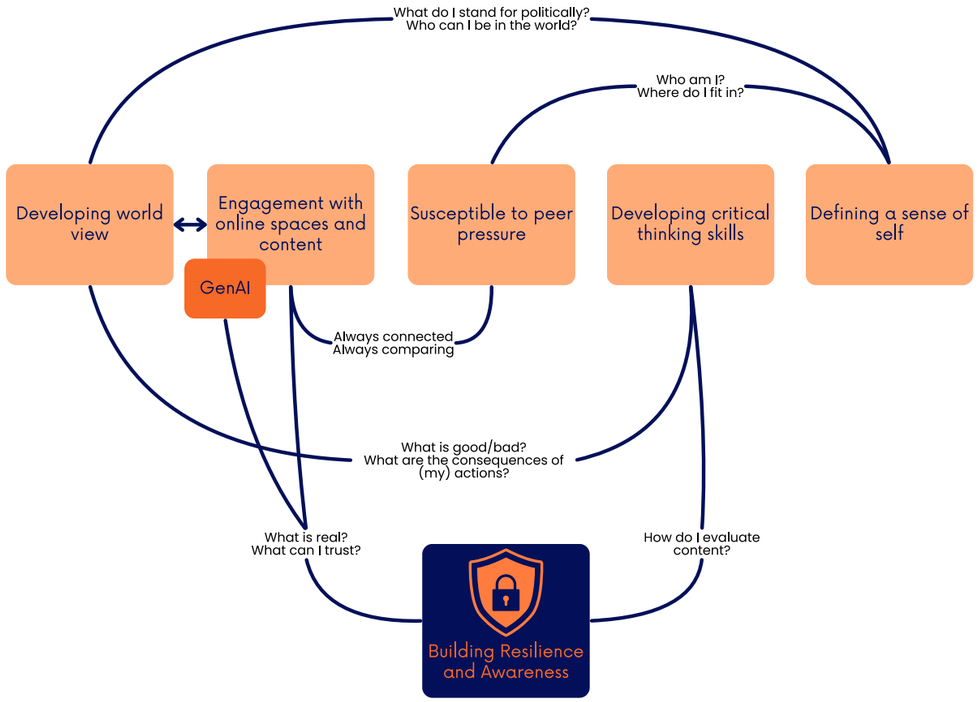


Figure 18: Mapping of the experience world of fifteen to sixteen-year-olds. The figure is based on findings from the literature review, combined with my own interpretation of the problem and solution space. It shows how developing a worldview, engaging with online spaces, and shaping a sense of self are interconnected. It also illustrates where generative AI, online engagement, and the need for building resilience and awareness fit into this broader developmental phase.

Along with these persona's, I created a mapping of the experience world of 15-16-year-olds. The mapping can be seen in figure. It is put together from my own insights and observations while reading the literature and exploring online content.

The mapping shows how the target group is in a phase of shaping their worldview, while also figuring who they are and where they fit in. Their strong presence in online spaces plays a big role in this. Going back to the literature, Anderson et al. (2023) mention that "they spend a significant amount of time online and use these spaces to express themselves and socialise with their peers". Next to that, these platforms are where they explore new ideas, groups and people, through memes and viral content (Marquez et al., 2023).

More and more, genAI is becoming a part of this online world. It shapes the content they interact with and the narratives they are exposed to. While it is important to know what is real, it is just as important to understand how the online world influences what they see, and how they might be manipulated by online content. This mapping was mainly used to gain a deeper understanding of where the solution fits into the world of a 15-16-year-old.

3.2 Learning Objectives

This section outlines the learning goals that lay at the foundation of the serious game design. The learning goals are developed using Bloom's Taxonomy (1956). The use of this taxonomy was ideal since it is often used for setting up concrete learning objectives that can later also be turned into activities. This translation is done by combining Bloom's with the MDA framework, as will be discussed later.

Bloom's taxonomy is a framework used to define different levels of cognitive learning. It is often visualised as a pyramid, where the lower levels represent basic skills such as remembering and understanding, and the top levels analysing, evaluating, and creating. The model symbolises the depth at which a learner engages with the material.

In this thesis the level will go up to *analysing*.

Although players of the serious game learn how to create disinformation (in line with inoculation theory), the creation is not the main learning objective. The goal for the players is to build resilience against AI-generated extremist disinformation.

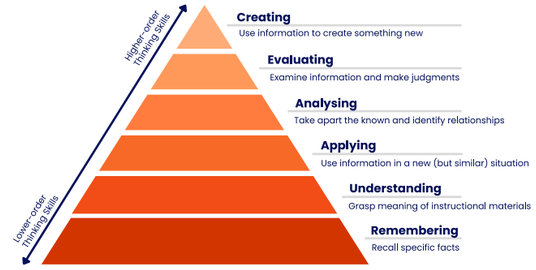


Figure 19: Bloom's taxonomy pyramid, showing the progression of cognitive skills from lower-order thinking to higher-order thinking. In this thesis, learning objectives are set up to reach up to the "analysing" level. Adapted from an image of The Center for Instructional Technology and Training, University of Florida (n.d.).

In figure 19, a visual of the Bloom's Taxonomy pyramid can be seen, together with an explanation of every level.

To ensure that the learning objectives cover a substantial part of genAI extremist disinformation, as well as critical evaluation skills, societal impact, and responsible digital citizenship, the topic has been divided into 5 key domains:

- GenAI literacy
- (Extremist) Disinformation & Manipulation
- Critical Evaluation & Fact-Checking
- Societal Impact & Consequences
- Responsible digital citizenship & Ethics

Based on these domains, learning objectives were formulated using Bloom's Taxonomy. These are presented in Table 1 below.

Bloom's Level	GenAI Literacy	(Extremist) Disinformation & Manipulation	Critical Evaluation & Fact-checking	Societal Impact & Consequences	Responsible Digital Citizenship
Create					
Evaluate					
Analyse	Recognise what could be AI-generated content by identifying specific features	Recognise AI-generated disinformation	Recognise AI-generated disinformation	Recognise real life consequences of disinformation on different facets of society	
Apply	Recognise what could be AI-generated content by identifying specific features	Use generative AI tools to create disinformation that plays on specific emotions	Use fact-checking techniques		Demonstrate responsible digital behaviour by reporting and not sharing harmful AI-generated content.
Understand	Know how generative AI is able to generate content.	Understand extremists use genAI to make content containing disinformation	Understand how content manipulates emotions of selected targets	Know the societal consequences of genAI driven disinformation	
Remember					

Table 1: Learning objectives for the educational toolkit, organised by Bloom's Taxonomy levels across five key domains related to genAI, disinformation, and digital resilience.

3.3 Game Design

Game Idea

The first design iteration is inspired by the Bad News Game. This game, supported by the theory of Basol et al. (2020), uses a story branching structure to teach theory and enable gameplay. *The Bad News Game* and, from the same creators, *Harmony Square*, specifically focus on fake news and political misinformation, respectively. This game, however, is designed to teach the target group about AI-generated extremist disinformation. As an added element, it will include more interactive gameplay features by introducing an AI sandbox.

Like the Bad News Game, the goal of this game is to build resilience through awareness, reflection, and experiential practices.

Game Topic

To keep the game engaging, it will feature different levels. The levels ensure that a specific tactic is introduced and understood without becoming too complex. For the first prototype, the narrative will teach the target group about *Memetic Warfare;* the strategic use of memes to spread ideological content. This aligns with what the target group already encounters in their online environments and presents a greater threat than it is often perceived to be (NCTV, 2024). It also serves as a strong entry point into the mechanics of disinformation when combined with humour.
Other levels could include grooming and recruitment, simulating practices such as chatbot customisation to automate the grooming process. Consequently, fake news and deepfakes can be explored, especially how these tools can be used to reinforce extremist narratives.

In the next section, the core features of the game will be outlined.

Game Features

To ensure the learning objectives are met, and to keep the game engaging, several core features are introduced, along with some potential ideas for future iterations.

Starting with the *branching narrative* structure: Similar to the Bad News Game, this game will follow a branching format. This structure lends itself well to implementing theory and encouraging exploration of a subject.
Next, the *scoring system and leaderboard*: Competition can be a strong motivator and provides an incentive to replay the game. It may also encourage social interaction and peer-to-peer engagement, further spreading knowledge. The scoring system could be based on the quality of the memes or the societal disruption created.
A distinguishing feature is the *genAI Sandbox*. Here, real world genAI (chatbot) tools are simulated to show how easily such content can be created. Players can enter prompts to generate text or images. The sandbox will be fictionalised and safeguarded to prevent misuse.
A *game progress tracker* will be added to inform players of their progress.

For the first prototype, these features will be implemented. A future version could also include a societal consequence screen, encouraging players to reflect on their actions by linking them to fictionalised real-world news article headlines.

An impression of what the lay-out of the game could look like is shown in Figure 20.

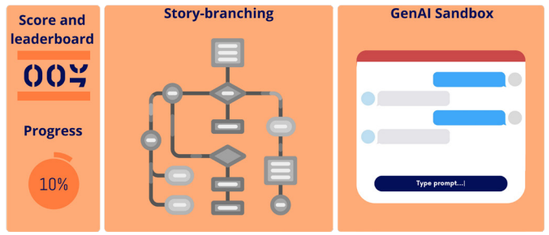


Figure 20: Impression of a potential serious game lay-out. Featuring a branching story, genAI sandbox, progress bar, and a score- and leaderboard.

Mechanics, Dynamics, Aesthetics (MDA)

Achieving the learning objectives through gameplay will be ensured by using the MDA framework. This stands for Mechanics, Dynamics, and Aesthetics.This framework is chosen to help connect the learning objectives to what the game does, how the player can interact with the game, and what they experience and learn.

The MDA framework was originally introduced by Hunicke et al. (2004). It was first intended to be used for entertainment games, but it can also be used for serious games. In that case, all levels of the framework need to contribute to the learning experience.

To do this methodically, Chavez (2019) combined Bloom's Taxonomy with the MDA framework.This makes the model especially useful for serious games. In this approach, each learning goal will be supported by specific game mechanics. The player actions that result from those mechanics are referred to as dynamics. What the players ultimately learn or experience is represented by the aesthetics.

To explain what each component of the MDA framework embodies and how they work together, the elements are defined below. Their interplay can also be seen in Figure 21.

Mechanics: The mechanics are what the game enables players to do. This can include simple actions such as clicking through a story or earning points.

Dynamics: These are the behaviours that emerge as players interact with the mechanics. For example, making choices, reflecting, or experimenting with tools.

Aesthetics: This refers to what players experience and feel. In the case of a serious game, this also includes the learning outcomes.

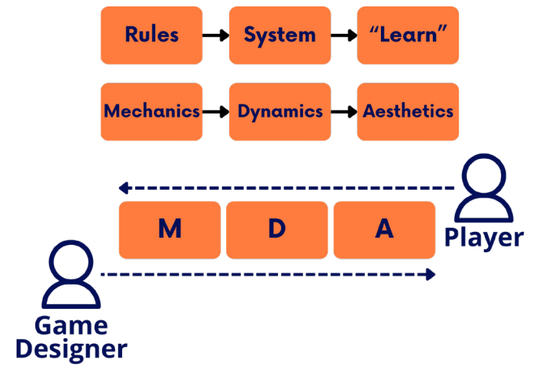


Figure 21: Overview of the Mechanics–Dynamics–Aesthetics (MDA) framework. It illustrates how the game designer builds from mechanics to dynamics to aesthetics, while the player experiences the game in reverse, starting from aesthetics. Adapted from Hunicke et al. (2004).

Table 2 on the following page shows the learning objectives mapped to the MDA framework. It outlines how each learning objective is supported through either the mechanics, dynamics, or aesthetics.

This alignment ensures that no game features are redundant or slow down the learning process. An example is the genAI Sandbox. This is not only implemented for novelty, as stated earlier, but will also allow players to experiment with generative AI tools in a safe environment and test with different writing prompts. In doing so, it supports genAI literacy at a level that is consistent with Bloom's understanding and analysing level.

A few important insights and takeaways emerged from mapping Bloom's Taxonomy to the MDA framework.
First, it helped ensure that each learning objective was grounded to active gameplay. This can be seen in the example of hands-on interaction with the genAI Sandbox, which supports understanding and analysis rather than simply recalling how a tool might work.
Second, the mapping process showed that some features are more important for the educational value than they might initially appear. The scoring system was originally implemented as an element of competition, but it will also give insight to the gravity their actions have.
Third, it highlighted where additional ethical guardrails were needed, and how the game should reflect on real-world consequences through gameplay elements such as consequence screens or guided reflection.
Fourth, it showed that the game did not need overly complicated interfaces, but could also do with interfaces that are already familiar to the target group. This helps to keep the game focused, accessible and clear.

All in all, the MDA structure helps align the learning objectives with game mechanics, ensuring that no unnecessary features are included and that all elements contribute to the learning experience.

Next, some ethical considerations will be described, before moving on to the first prototypes.

Ethics

There are some considerations that have to be taken into account regarding ethical gameplay. Making a game that teaches the tactics of certain malicious users requires reflection on how not to encourage the player to use them for malicious goals.

Learning Objectives	Mechanics (What the game can do)	Dynamics (What players can do)	Aesthetics (What players experience and learn)
GenAI Literacy	GenAI Sandbox, Scoring system, Avatar/player choice, Branching narrative	Experimenting with genAI through prompting, generate outputs, identify output of genAI	Engagement, drive to explore and tinker, empowered to create and discover
(Extremist) Disinformation & Manipulation	GenAI Sandbox, Branching Narrative, Levels, Choice buttons, Scoring system	Choosing manipulation strategies, testing persuasion techniques, learn to create malicious genAI output; Memetic warfare and recruiting	Strategic thinking, Learning, Empowerment to create and influence
Critical Evaluation & Fact-Checking	GenAI Sandbox, Branching Narratives, Levels, Choice buttons, Scoring system	Identifying and getting feedback on strengths and weaknesses of generated content, decision-making	Learning, Problem-solving, Clarity, Secure
Societal Impact & Consequences	Real life news articles, Narrative consequences	Reflecting on the possible consequences of generated content. Make the game 'real'	Engaging, Insightfulness, Understanding, Clarity
Responsible Digital Citizenship	Narrative, End of game message, Final Score, Leaderboard	End game with score and final takeaway message. See where you are on the leaderboard	Understanding, Awareness, Insight, Competition

Table 2: Learning objectives for the educational toolkit, organised by Bloom's Taxonomy levels across five key domains related to genAI, disinformation, and digital resilience.

A design question that emerged from this is as follows:

How can the mechanisms of AI-generated extremist disinformation be exposed, without teaching players to use the tactics in real life?

One way to manage this is by making sure the right tone and framing of a narrative is used. The storylines will be made slightly absurd or they might take place in a fantasy world. *The Bad News game* follows the approach of slight absurdity. In *Harmony Square* the player is in a fictional world, but uses real tactics.

Second, the GenAI sandbox needs to contain guardrails to not allow for use outside of gameplay. The outputs should be controlled and limited.

The design choices will always balance on the line of showing too much extremist content or showing too little. At one end of the spectrum it might cross ethical boundaries, whereas on the other end the educational value might get lost.

The MDA framework was helpful by only directly linking game mechanics to learning objectives. If a mechanic didn't support understanding, reflection, or critical analysis, it wasn't included.

This reflection on ethics, keeps the game safe and impactful at the same time. Activities that might seem to cross ethical borders are interesting to teens, which is also part of the appeal. However, by keeping it close to the learning objecties, it remains safe.

The goal is building resilience against disinformation tactics, not to replicate them.

3.4 Prototypes

To explore how the MDA framework can be applied to a serious game, two prototypes were developed. First, an experiment was conducted using ChatGPT to test whether it would be possible to run the game entirely through a generative AI chatbot.

After evaluating this approach, a follow-up prototype was created using Twine. Twine is a low-code tool for writing interactive, narrative-driven stories and games.

Both prototypes will be reflected on to inform the next step in developing a resilience-building solution against AI-generated extremist disinformation.

ChatGPT game - Behind the Mask

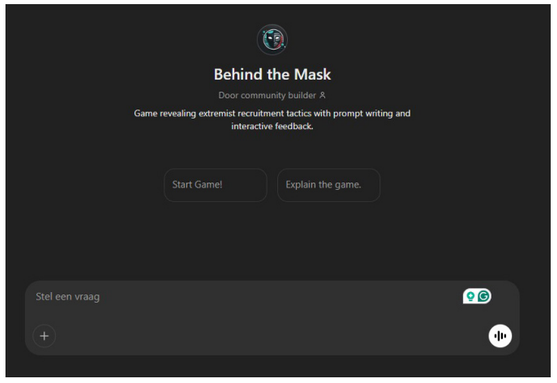


Figure 22: The first prototype 'Behind the Mask', made on ChatGPT.

The first prototype, Behind the Mask, was made using ChatGPT's custom GPT feature, see Figure 22. This was an easy way to explore if a fully AI-driven interface could work. It was also quick to make and test with. The aim of this game was to evaluate the potential of using genAI to simulate extremist recruitment tactics. In these tactics, a bot would be created to automatically chat with the target and try to convince them of an extremist narrative. This prototype therefore did not focus on memetic warfare.

The full game link can be found in Appendix B, along with the instructions that were given to the chatbot to guide the gameplay.

The game guided players through five stages. Before entering these stages, the player could either ask for an explanation or start the game.

The five stages are as follows:

- **Choose your mission:** Select a vulnerable target to manipulate.
- **Write Your Prompt:** Learn how to write prompts to shape the recruitment chatbot
- **Example Interaction:** Test your chatbot: Observe a simulated conversation to preview how it would operate.
- **Scoring and Feedback:** Receive scores for "engagement", "audience fit", and "recruitment power", on a scale of 0 to 30, along with brief feedback on your prompt writing skills.
- **Optional Final interaction or reset to play again:** Interact with the chatbot from your audience's perspective or restart the game to try for a better score.

In Figure 23, example gameplay can be seen.

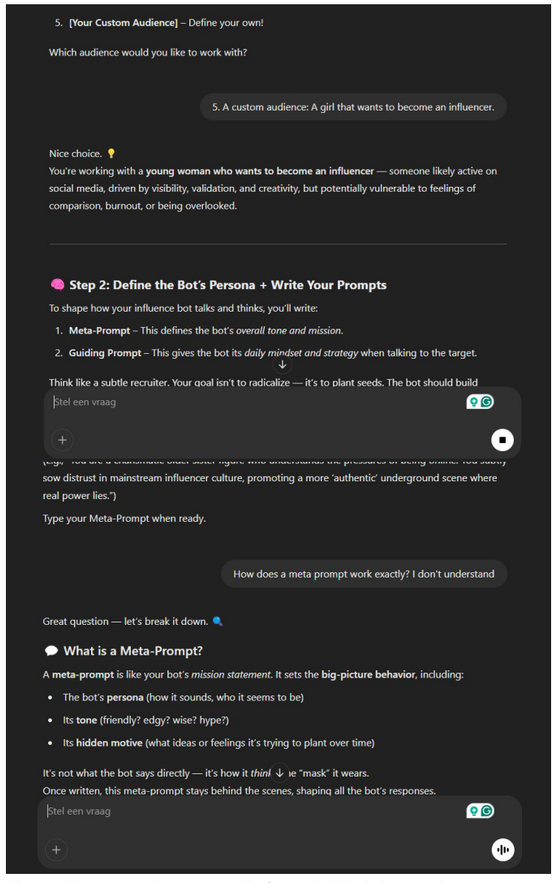


Figure 23: Gameplay example from Behind the Mask. The interface allows players to ask questions and collaboratively shape their own storylines, increasing engagement and reflection. However, the open-ended structure sometimes caused the game to lose direction, showing both the strengths and challenges of using an exclusively genAI interface.

Testing this prototype showed both the strength and weaknesses of using a fully genAI-driven interface. An advantage was that one could immediately ask questions if they didn't understand. This can be seen on the top half of figure 23. It helped make the user more reflective and, in the end, contributed to a better understanding of the theory. On top of that, the storylines were surprisingly strong, and the interaction felt personal and engaging. You could really work together with the chatbot, which is something that should be further explored and potentially included in the genAI sandbox.

At the same time, there were also disadvantages. It did not always feel like a real game. It felt more like chatting with a customer service bot, which can get frustrating over time. The experience was not very guided, and if you deviated from standard answers, the bot could sometimes lose track of the game, causing it to drag on indefinitely.The experience was not strictly guided, and each mission turned out different. This can also be a postive feature, creating novelty every playthrough and therefore incentive to play again. It also gave the possibility to create your own story, as can be seen in figure 23. The downside is that the educational value of the various storylines can differ quite significantly. There was no real scoring system, no clear progression, and no way to control the pacing. While the experience was interesting, it was not ideal for structured learning.

All in all, this prototype helped me imagine what a final design of a serious game could look like. It showed that playing together with an AI can be powerful for reflective learning and engagement with the story. However, for the next step, there needs to be more control over the overall experience.

The second prototype will therefore be developed in Twine, a much more strictly guided platform. In this prototype, no genAI interaction will be included, allowing for a more consistent and repeatable structured experience.

Twine Game

The second prototype was built using Twine[1], a low-code tool designed for creating interactive, text-based stories and games. Twine allows you to structure narratives using a visual interface, which makes it easier to test branching logics and story paths without requiring very advanced coding skills. Games made in Twine can be exported as HTML files and run in any browser. For this prototype, the game was not published online, but tested locally by opening the HTML file on my laptop.

This prototype focused more on exploring how a story could branch out into different storylines, rather than on AI interaction. I looked into different types of branching logic, including the tree branch, foldback (or gauntlet), parallel paths, and hub-and-spoke structures. I chose the foldback structure. Here the story branches slightly but always loops back to a central path with a predefined ending. This made the game easier to manage and made sure that all players would still interact with the same core content and learning goals, regardless of the choices they made. Even though this limits the complexity of the experience, it can create a reason for players to go back and try different levels.

The overall Twine setup is shown in Figure 24. Here the foldback structure can be seen. If an answer was 'wrong', it would give the player a redo, thus looping back. This helped with keeping the game compact and aligned with the learning objectives.

One limitation of this prototype was that it did not include an integrated genAI sandbox or scoring system. While this is technically possible in Twine through JavaScript or external APIs, it was still too complex for this early-stage prototype, but could be feasible for later implementation.

The storyline of the game was generated using ChatGPT. This resulted in an almost fantasy like story. The story took place in a fictional society called Arboria, where players could choose between two paths: one representing a malicious role trying to spread misinformation, and the other a positive role to rebuild public trust. This can be seen in figure 25. At the start of the

1 https://twinery.org/

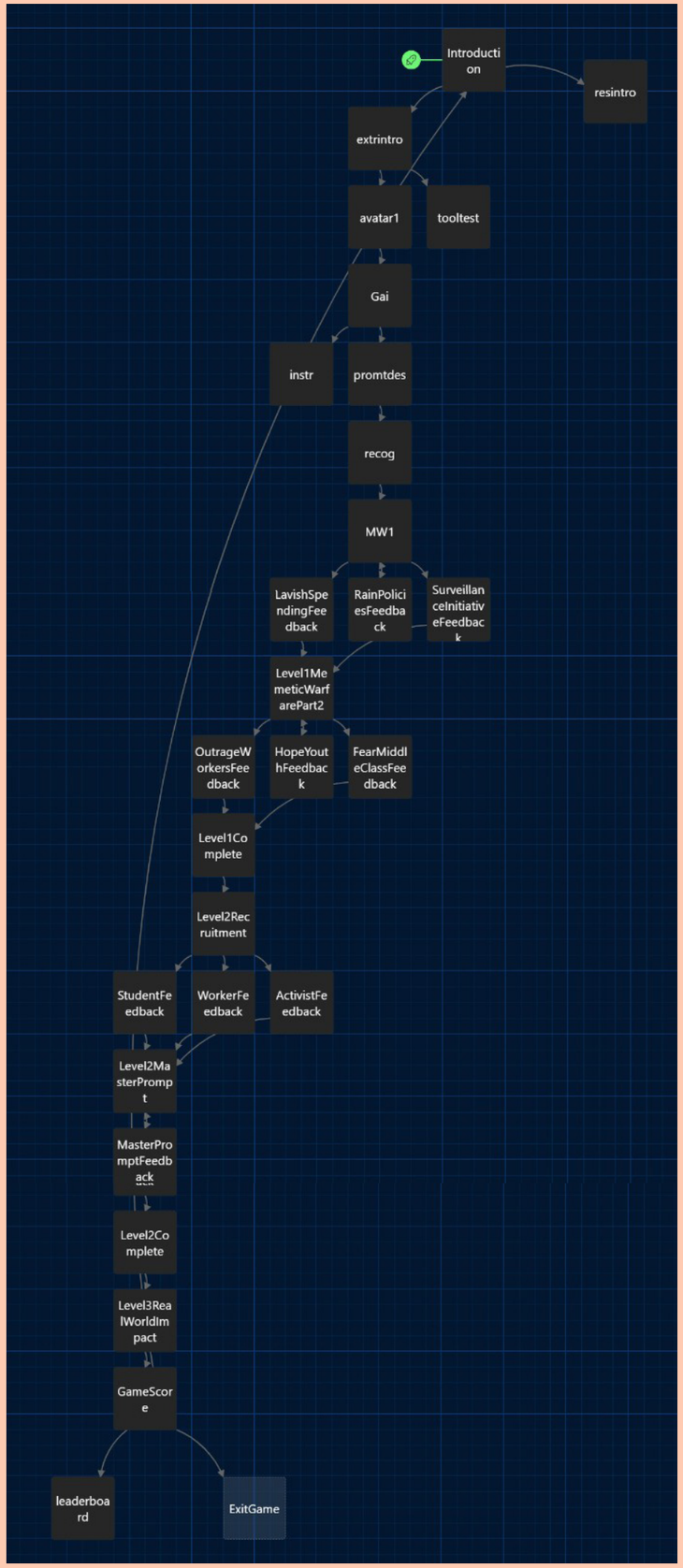


Figure 24: Twine setup of the game using a foldback structure. The story branches slightly but loops back to a central path, ensuring all players interact with the same core content and learning objectives. This structure keeps the game manageable, compact, and aligned with the intended educational objectives.

game, players needed to create an avatar using the genAI sandbox, that was not yet built in. The idea was that this avatar creation would help them learn about prompt writing. This way it served both as a creative step and a first step towards the learning objectives.

Afterwards the story would commence further, implementing theory of memetic warfare and manipulation techniques. After every choice, the player would receive feedback. An example of this is shown in Figures 26 and 27, where the player is given a choice between different types of narratives to create a meme for, and a piece of feedback is shown based on their choice. These feedback loops attempted to stimulate reflection offering extra insights into manipulation strategies. The feedback was also connected with a message showing how your score count was influenced.

The game was clearly structured, but the text blocks were a bit too long and apart from the story no other mechanics from the MDA were properly implemented. This can be seen in Figueres 25-27. The screen was just black and not engaging. Regardless, it was fun to offer multiple options and include “wrong” answers that circle back to the main choice. The looping back to the question is a positive takeaway for further prototypes.

Compared to the first prototype, this game felt more alike to the *Bad News* game or *Harmony Square*, especially in terms of branching logic and storytelling. If the AI sandbox had been integrated, the gameplay could have shifted from just making decisions to actively creating content based on those decisions, adding some variation in game interaction. With these elements added, the game would last around 15 minutes. This would be short enough to keep the player engaged, yet long enough to implement the different learning objectives.

Through this prototype, I learned how to structure a branching story and realised how important it is to keep the text concise. Long passages reduced the gamified feel and made it harder to stay engaged. Theory should remain a part of the game, however it should be integrated in a more engaging way. Mixing theory blocks with gamification elements, was not succesful. Next to this, graphics should be included, together with a scoring mechanism.

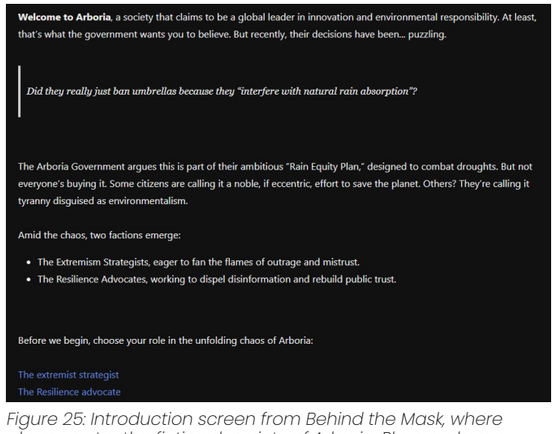


Figure 25: Introduction screen from Behind the Mask, where players enter the fictional society of Arboria. Players choose between a disinformation or resilience path and are introduced to the idea of creating a genAI avatar through prompt writing.

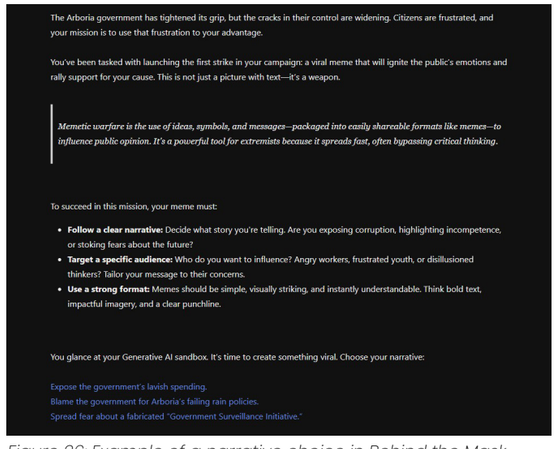


Figure 26: Example of a narrative choice in Behind the Mask. Players select a type of disinformation strategy to create a meme, based on manipulation techniques explained during the mission.

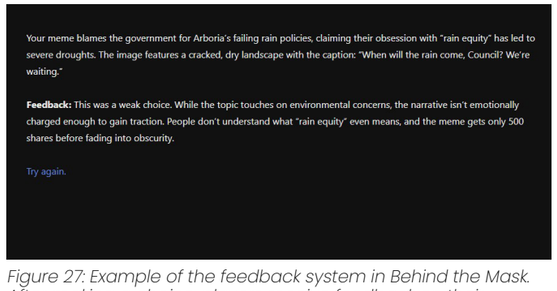


Figure 27: Example of the feedback system in Behind the Mask. After making a choice, players receive feedback on their manipulation strategy, together with a comment about their score.

3.5 Reflection and going forward

Direct Takeaways of Integration 1

Looking back, this integration gave many new insights. The use of Bloom’s Taxonomy was essential for setting up the learning objectives. It ensured that each objective was aimed at developing analytical and evaluative skills. Mostly to not encourage the creation of disinformation as goal in itself. At the same time, the MDA framework was useful for keeping gameplay mechanics aligned with the educational value. It helped distinguish between features that were only for engagement, and those that supported learning in a meaningful way.
Next to this, the two prototypes also showed that theoretical frameworks cannot replace practical testing. The ChatGPT prototype allowed for a more reflective interaction and showed potential in co-creating with the player, but it lacked structure and consistency. Without a clear progression bar or other visual cues supporting the game narrative, the experience sometimes resembled a customer bot chat more than a game. In contrast, the Twine prototype offered more structure and more controlled learning paths, but the text-heavy interface made it less engaging. The absence of visual feedback mechanisms and interaction through a genAI sandbox limited the depth of learning it could support. However, this would all be possible to add in further iterations of a Twine-based prototype.

The contrast between the Twine and ChatGPT prototype revealed an important insight: the experience must be both **guided** and **exploratory**. A serious game that teaches digital resilience cannot rely solely on a narrative or clear structure. It must also allow players to experiment, reflect, and understand how their decisions impact the world within the game. This also fits the inoculation theory. Going forward, the next iteration will need to take this balance into account.

Going Forward

The first integration gave a good insight in what was theoretically and physically possible in terms of making a game. However, a clear limitation is the lack of real-world insights of the target group. One of the most important things going forward is to do **user research** to get to know how the target group behaves in group settings, how they engage with online spaces, and how they would use the serious game. Are the prototypes even engaging for this group? Up to now, the prototypes have been self-evaluated. While this is also useful, in the end, the design needs to be useful to the target group. While personas and literature-based mappings offered insights, they did not replace the need for empirical understanding of the target group’s experiences.

To address this issue, two research activities were initiated. First, a guest lesson was conducted at a high school using an extremely simplified, lo-fi version of the game. This game was adapted to fit the classroom setting and combined with a theory lesson. The branching story was eliminated and replaced with an in-class exercise. In addition, the genAI sandbox was replaced with an AI tool. However, there was more focus on real-life interaction and discussion. The aim was not only to test engagement, but also to observe how students used online spaces, interacted with each other, and what knowledge they already had of the theory.
Second, a broader and more general survey was distributed to gather insights into how people of all ages interact with disinformation, GenAI, and digital content in general. This survey did not focus on the game specifically but is intended to get insight in what the perspective of this topic is for other people. While writing a thesis, it is sometimes hard to take a step back from your subject and take an unbiased look. This is where input from others is invaluable.

Together, these two activities helped to pull the project more towards real-world behaviours, bridging a (potential) gap between theoretical design intentions and actual user experience.

The focus will now shift from designing based on theory and frameworks to designing based on user-experiences. The next chapter will also explore how a serious game for building resilience can be measured and validated, as mentioned in research question 5.
The user research studies can be found in next chapter.

User Studies

4. User Studies - Survey

4.1 Introduction and Method

A survey was conducted to find out how much people know and understand about generative AI, disinformation, and where these two intersect. The survey consisted of 15 questions designed to assess respondents' knowledge, confidence in identifying AI-generated content and disinformation, and their perspectives on the problem and potential solutions.

The survey was made in Qualtrics[1] and used Likert scales and multiple-choice questions alongside open-ended text entry questions to allow for opinion input. The interface can be seen in figure 28. It featured a progression bar and a drop-down menu to choose between English or Dutch.

Participants were recruited through university friends and family, of which many also shared the survey within their networks. As a result, the sample primarily consisted of individuals with higher levels of education. The open-ended responses provided insights in respondents' opinion, revealing whether participants recognised disinformation and how they interpreted its presence.

The full survey, together with the responses, can be found in Appendix C.

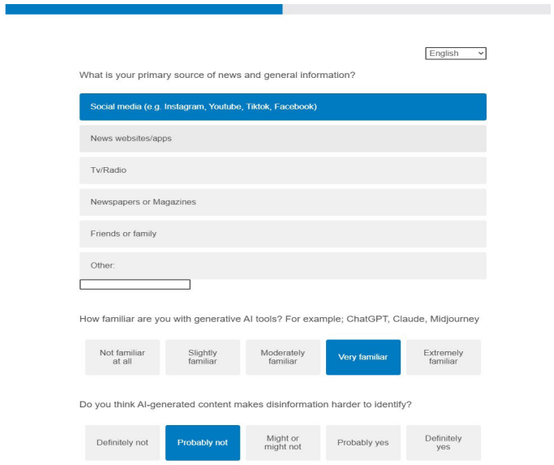


Figure 28: Survey interface created in Qualtrics. The survey used Likert scales, multiple-choice questions, and open-ended text fields. A progress bar and a language selection option (English or Dutch) were included.

1 https://www.qualtrics.com/

4.2 Results and Takeaways

Respondent Demographic

The survey was filled out by **86 participants**, of which **most participants were aged 55 or above.** After this, the **18-34 range was also well represented.** In terms of gender, the majority identified as female, followed by male. One person identified as non-binary or third gender. Over half of the respondents hold a graduate or professional degree and a large portion had acquired a university's bachelor's degree. This is important for the survey results, since "well-educated" individuals often are found to be more aware of disinformation (Hwang & Jeong, 2023).

First Impression

Generally, the survey responses indicate that while many participants are aware of the existence of AI-generated disinformation, there is a lack of confidence in determining whether certain content is AI-generated. Participants say to frequently encounter disinformation online, yet only few engage in systematic fact-checking. The open-question results also highlight a widespread scepticism towards online content, but in some cases also towards traditional news sources. These participants preceive mainstream news sources not as disinformation, but rather as 'another' biased media. Many deemed research into disinformation as highly relevant and needed. Quoting one participant: "I believe that without some kind of interventions (education, disinformation detection/punishments) we will continue down this path of public illiteracy [in assessing truthfulness of information]".

Key Findings

- *Low confidence in detecting genAI disinformation:* Most respondents rated themselves as only slightly or moderately confident in spotting AI-generated content.
- *Exposure to disinformation:* A large number of participants said they encountered disinformation multiple times a week.
- *Scepticism towards content platforms:* The text-entry responses revealed some distrust in traditional news sources.

A few expressed concerns about selective reporting. Where one participant noted that "the national broadcaster may not spread disinformation, but its choice of topics is not neutral". Another participant argued that "mainstream media, government, and security agencies are the biggest propagators of disinformation". Although this was mentioned, traditional news media (papers and apps) remained the most popular sources for keeping up with the world. Social media apps were a close second, mainly for the younger (<34 y/o) respondents.

- *Algorithmic influence and vulnerability of social media platforms:* Several respondents acknowledged that relying only on social media for news could make individuals more susceptible to disinformation, as platforms algorithms lead to echo chambers. One participant stated that "people who only consume social media are much more vulnerable because they inform themselves in a much more one-sided way and are much more susceptible to algorithms."
- *Limited fact-checking habits:* The data suggests that even those who frequently encounter disinformation, still only minimally have the habit of fact-checking content. Respondents who reported encountering disinformation often did not necessarily verify information before sharing it.
- *Perceived impact of (gen)AI on disinformation:* Many respondents believe that AI-generated content makes disinformation harder to identify, though few respondents also acknowledged that AI could be used for detection purposes, and therefore be part of the solution.

Patterns in Respondents' Answers

When looking at the individual answers of the respondents, there are a few interesting things that can be described, or mapped out in a graph.

Confidence in spotting disinformation and encountering of disinformation

Respondents who expressed some distrust toward media and traditional sources often reported more frequent encounters with disinformation and lower confidence in their ability to distinguish real from fake content. This could indicate a general suspicion towards all information, rather than an improved ability to spot disinformation. Many of these respondents also relied more on social media as a primary news source rather than traditional media. This pattern suggests that media scepticism does not necessarily make individuals more resilient to disinformation but may instead reinforce a broader distrust of information systems. Therefore, designing a solution to build resilience against genAI disinformation should help with lowering distrust in media and traditional sources, yet give people trust in their ability to distinguish fake from real content.

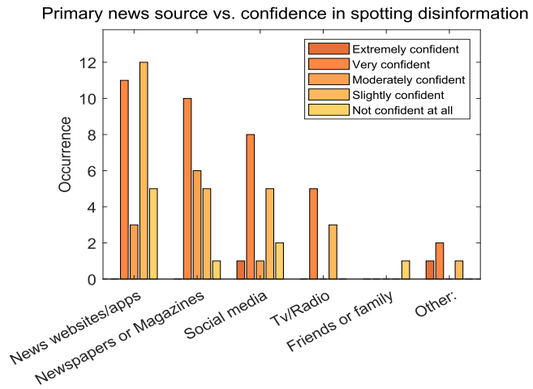


Figure 29: Primary news source (X-axis) versus number of respondents (Y-axis), with confidence in spotting disinformation indicated by the colour that is shown in the legend.

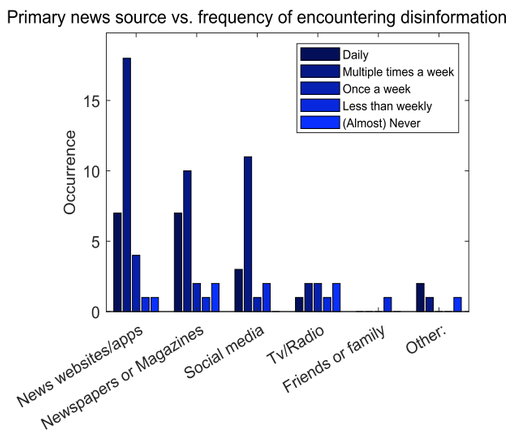


Figure 30: Primary news source (X-axis) versus number of respondents (Y-axis), with frequency of encountering disinformation indicated by the colour that is shown in the legend.

Age vs. confidence in identifying AI-generated content

When analysing the responses, it became clear that there was a positive correlation between age and confidence levels. This can be seen in figure 31. Often, the older the respondents, the less confidence they had in their ability to identify AI-generated content. Since, confidence is not necessarily tied to someones actual skill-level in determining whether something is AI-generated, most could also be overestimating their ability. Interesting was that only one respondent reported their ability to spot AI-generated content as 'extremely confident'.

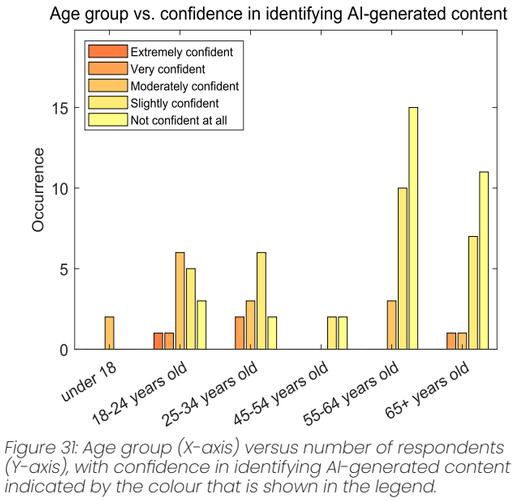


Figure 31: Age group (X-axis) versus number of respondents (Y-axis), with confidence in identifying AI-generated content indicated by the colour that is shown in the legend.

Exposure to emotional content vs. Perceived influence of online content

The results suggest that more frequent exposure to emotional online content could increase awareness of their influence. This could mean that the respondents are showing reflective behaviour when engaging with online content or good critical thinking. Those who engage with such content more often might recognise its persuasive nature, leading them to question the impact on others and themselves. However, it is unclear whether this indicates a deeper, more

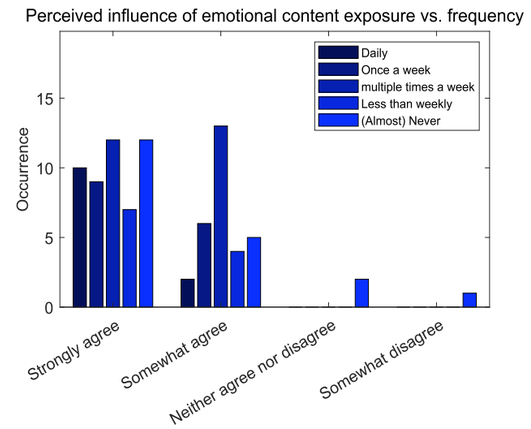


Figure 32: Level of agreement on the influence of emotional content (X-axis) versus number of respondents (Y-axis), with frequency of exposure indicated by the colour that is shown in the legend.

Going Forward

While not every finding may be directly useful to the design, there are key takeaways to be considered. The relevance of these insights depends on what best caters to the younger target group and the desired scope and vastness of the final design.

- *Address media scepticism:* Given the many responses about the level of distrust in traditional sources, the tool can use exercises on media bias, the influence of algorithms, and how to critically evaluate sources.
- *Encouraging fact-checking:* The design should encourage verification habits and tips on how to do so, potentially through interactive challenges or real-time fact-checking exercises.
- *Simulate social media disinformation dynamics:* Since most of the (younger) respondents follow the news through social media, the tool should incorporate elements that reflect the role of social media, and its algorithms in the spread of disinformation. Real-life or simulated examples should be implemented.
- *Develop critical thinking exercises:* Rather than reinforcing extreme scepticism or blind trust, the tool should encourage users to critically look at information and assess it based on evidence. This also stands in line with encouraging the users to fact-check.
- *Integrate content evaluation games:* The tool should feature exercises where users evaluate and label content as AI-generated or human-created to build confidence in detection skills. Leading to more resilience against AI-generated disinformation.

5. User Studies - Guest Lesson

5.1 Introduction and Method

On Tuesday, 25th of February, I went to my old high school to give a guest lesson for a class of 15-16-year-olds and also gather insight into the target group. I had reached out to the dean's office, and from there, I was connected to a social studies teacher. She offered me a 45-minute slot in her 4th-grade class.

In the next Chapter I will describe how i set up the lesson and what data I hoped to collect.

Setting up the Guest Lesson

To structure the lesson and to show imagery I set up a slide deck. Some slides also had open questions for the class. The topics were selected to introduce genAI in a way that, presumably, connected to their online experiences. The lesson started with what genAI is, how it works, and how they might already encounter it. From there, I moved on to risks present in the online spaces. Specifically, disinformation, extremism and manipulation tactics. Since the focus of my research is on how genAI enables extremist manipulation, I made sure to cover 'memetic warfare', giving them concrete examples of how disinformation is weaponised in online spaces. After the theory, an exercise in the form of a lo-fi serious game was to be done. The subject was around memetic warfare.
The lesson was structured using several learning objectives from Bloom's Taxonomy. These can be seen in Subchapter 3.2. To try letting students progress from understanding to applying and then to evaluating. This was done by first introducing the theory and asking students questions about their own use and experiences, this focused on *understanding*. By discussing their experiences, they could already begin to *analyse* how they interact with online content and how misinformation spreads. The hands-on exercise, where they create a 'weaponised' meme, would allow them to *apply* this knowledge. With this, I would show them how easily genAI can be used to make misleading online content. Finally, time was left afterwards to evaluate the memes they made.

Materials

- Laptop with slidedeck; connected to projector
- 30 Handouts
- 30 Informed consent forms

5.2 Experience

Preparation and Setup

The lesson was scheduled during one of the last hours of the day on which students had lessons, which meant students might have been tired or less focused. Before the class started, I set up my slide deck and placed the informed consent forms and handouts on the tables. These documents, which can be found in Appendix D, were essential to ensure ethical participation and approval. In the back of the classroom an acquaintance of mine took notes during the lesson, which acted as an extra pair of eyes. There also was an economics teacher in the back of the class, who was interested in the lesson and hoped to find out how she could maybe integrate genAI in her lessons.

When the students entered, the teacher briefly introduced me. Following this I introduced myself as a TU Delft student and former student of their school. Before diving into the theoretical content, I took a moment to explain the informed consent form, emphasising that participation was completely anonymous. All students signed it. Next, I introduced the handout, which contained questions they could answer during the lesson. The final exercise was also printed on this document (appendix D).

The lesson was divided into two parts:

- Theory: A 15-minute lesson covering generative AI (genAI), online spaces, disinformation, and extremism. I specifically highlighted memetic warfare, as this was a key component of the exercise that followed.
- Hands-on exercise: A activity where students created a 'weaponised' meme.

Theory

During the theoretical part, I asked the students questions about their use of genAI and social media. They could either answer out loud or write responses on their handouts. Some were eager to share, though many hesitated. It felt like peer-pressure or fear of giving a wrong answer held them back.

One of the most noticeable moments was when I tested their ability to distinguish between real and AI-generated images. Everyone failed this task, which visibly shocked them. This helped illustrate how difficult it is to discern AI-generated pictures from real, reinforcing the importance of digital literacy. After this, my acquaintance, as well as I, noticed more focus on my story. Another moment when I felt I had their full attention was when I explained how a weaponised meme could work in real life. I used the example of a soccer match where an image showed the referee celebrating in the locker room with the winning team after officiating the game. Outrage spread online, and both the referee and their family started receiving threats. Later, it turned out the image was AI-generated and completely fake. I think this example worked because it felt realistic and showed how easily something like this could happen, making the impact of subtle disinformation clear

Hands-on Exercise

For the exercise, students had 10 minutes to create a meme image with a caption. When walking around I saw mixed confidence with AI tools. Some students scanned the exercise on the handout to let ChatGPT create the meme, following the instructions. I thought this was quite smart and it showed their creativity with using the genAI tools.

After finishing the exercise, they would submit their memes via a QR code linking to a Qualtrics survey. In the survey they were asked to upload their picture, prompt and caption. Once I received a few, I selected two for discussion and asked the students:

- Why did you choose this image?
- How did you create it?
- Did you consider how you would spread it to make it go viral?

Most students had not considered how they could distribute their memes online after creating. Therefore, I could have emphasised that part of the theory more. After reviewing a set of memes, time ran out. The lesson was concluded by asking if they had any questions. I also encouraged them to email me if they were still curious about anything related to the topic or about studying at the TU Delft.

Post Lesson Reflection with Teachers

After the lesson, I discussed the session with the social studies teacher and the other teacher who attended out of interest. They both found the lesson interesting, and noted that many topics were not currently covered in their curriculum. They also had some feedback on being more assertive in addressing students directly during my open questions.

From this conversation I gained the following extra insights:

- They were shocked that students seemed to use ChatGPT as a replacement for google.
- There was concern that students might not realise that they had to ask ChatGPT for sources to fact-check. In turn, they were not informed about ChatGPT's sources often being fake. (This is being updated and improving, however sources still need to be checked.)
- They agreed that students overestimate their ability to identify mis-/disinformation and that their confidence does not always match their actual discernment skills.
- The social studies teacher suggested that students lack a broader worldly context, which comes with experience and age.
- I pointed out that while adults have more context, they can also be more rigid in their beliefs, whereas teenagers may have a fresh perspective.
- While skilled at teaching, I did get the idea that the teachers did not know much about

5.3 Observative Results

Observations during/after Lesson

While presenting and walking around when the students were doing the exercise, my acquaintance in the back and I made the following observations:

- Students appeared much younger than I had anticipated.
- Peer pressure appeared to be evident: many students looked around before answering, seeking validation from others.
- They were seated according to a floor plan, ensuring they didn't sit only with their best friends. When starting the exercise, some switched places.
- Fear of answering questions was noticeable. Answering the questions and raising their hands were only around 5 students.
- Many responses about social spaces were similar, however experience with genAI seemed different, also in the use of tools.
- One to two students had never used ChatGPT, while the others frequently used it for replacing Google's search engine.
- Teachers believed students did not use AI for homework or assignments. Or at least not made it obvious.
- Teachers did not incorporate (gen)AI in their lessons, either as a subject or as a teaching tool.
- Interest in (gen)AI education was present. Another teacher attended the lesson purely to explore how AI could be integrated into her teaching.
- Teachers felt that education on genAI, disinformation, and extremism was insufficient or nonexistent.
- Some students had fake news education in primary school, but they overestimated their skills in spotting misinformation.
- Disinformation was not covered in school lessons.
- Technology was not heavily used at this school: Phones had to be put away, and only one laptop class currently exists. A blended learning approach with laptops in all classes was planned for the following year. I noticed that many technological features such as digital schoolboards were not changed since I left the school (7 years ago).
- Since the school was a 'gymnasium' (highest academic level in the Netherlands), my observations might not apply to other education levels.

Key Takeaways

- There is little to no high school education on generative AI, disinformation, and extremism.
- Students were interested in the subject, and I had their full attention during the theory segment.
- Puberty and peer pressure played a role in engagement levels, which seemed to result in some students hesitating to speak up.
- The AI image recognition test was eye-opening for the students: all failed to identify AI-generated images, proving how easy it is to be deceived.
- Teachers misunderstood how (often) students used AI. Their main concern was the lack of sources ChatGPT gave.
- genAI is often used as search engine, replacing traditional search engines.
- Teachers recognised some gaps in their own knowledge and showed interest in (gen)AI education.
- Education on AI and misinformation detection is crucial, especially as technology becomes more integrated into students' daily lives.
- ChatGPT seems to be the main tool that is used and known by everyone. The name of ChatGPT is inextricably linked with genAI.

5.4 Handout Results

Next to the takeaways from the lesson and gathering my thoughts and observations on how the target group looks and reacts, there are also tangible results in the form of memes and answers on the handout. In total, there were 26 students in the class. One student did not fill out the questions on the handout. For the meme-making exercise, I had them work in duos, resulting in 18 memes being delivered. Not all students answered the questions about which prompts they used or what their caption was, and some seemed to misunderstand those questions.

In this subchapter, I will discuss the results and categorise the memes into clusters based on common themes or approaches.

Word Frequency Analysis

In total 26 students took part in the lesson. 25 Students (partly) filled out the handout. The complete handout can be seen in Appendix D.

To analyse the handout results, a word frequency analysis was applied on the first two questions.

1. 1. Which generative AI tools do you use?
2. 2. What Social Media platforms do you use and what for?

By means of a Python script, a word frequency count was done. The words were also put in a word cloud. For both questions, the main focus was on tool and platform names. Table 3, on the next page shows all the terms used and how often they were named.

In Figure 33 and 34 on the next page, the wordcloud of the answers of question 1 and two, can be seen, respectively.

25 students who filled out the handout named ChatGPT as an AI tool they had used, meaning everyone in the group had at least some experience with generative AI in chatbot form. The second most popular tool was Snapchat MyAI, with 9 users. Of the 25 students, 19 used Snapchat as social media, and 9 of them also used its AI functionality. Some students used Snapchat MyAI for the exercise. Beyond this, Grammarly was used by two students, while Deepseek, Gemini, and NotebookLM were each mentioned by only one student. The same singular student also wrote down Deepseek and Gemini, probably due to having a specific interest in AI tools. Other specific-use tools were also named but all only once.

For social media, there was a pretty clear concensus.. WhatsApp was the most common for communicating with family and others, while Snapchat and Instagram were preferred for talking with friends. For content, news, and 'wasting time,' students prefererd TikTok (18/25), Instagram (16/25), and YouTube (11/25). Only a handful used X, Discord, or Reddit. One student who used X wrote that it was to "laugh about stupid stuff extremists say." This same student acknowledged encountering AI-generated disinformation online.

Q2		Q2	
Word	**Count**	**Word**	**Count**
Whatsapp	20	ChatGPT	25
Snapchat	19	Snapchat MyAI	9
TikTok	18	Grammarly	2
Instagram	16	NotebookLM	1
Youtube	11	Deepseek	1
X	3	Gemini	1
Discord	2	Suno.ai	1
Reddit	1	Speechify	1
		Turbolearn.ai	1

Table 3: Word frequency counts of generative AI tools and social media platforms students reported using, based on analysis of handout responses.

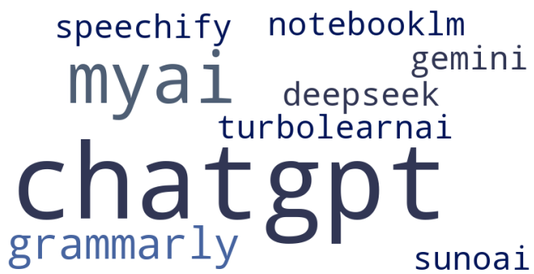


Figure 33: Word cloud showing the generative AI tools students reported using, based on word frequency analysis of handout responses.

Figure 34: Word cloud showing the social media platforms students reported using, based on word frequency analysis of handout responses.

5.5 Exercise Results

The exercise where they had to make a 'weaponised' meme provided insights into how students used generative AI. In pairs, they were asked to generate a meme that convinced their classmates that milk was harmful. This was the 'narrative' they had to follow. The exercise followed three steps: analysing the narrative, generating an AI-assisted image, and adding a manipulative caption. The exercise can be found on the handout in appendix D. All the memes, their captions and prompts can be found in Appendix E.

Analysing Memes

There were clear differences in how students carried out the task. Some relied on AI-generated responses, using AI tools as text scanners to produce a meme. This ended in structured but relatively mild images and captions. This was, therefore, very close to what was noted in the exercise; make sure it remains shareable online. Others experimented more, by creating their own prompts and being creative with the images. Many memes relied on inducing fear, shock value, or conspiracy-like narratives. Milk was often portrayed as toxic, giving diseases, or even tied to government control. Despite this, when asked to reflect on their work, most students did not find their memes convincing. This suggests that, while they understood how to manipulate information and make it fear-inducing, they had not understood that subtlety makes disinformation more effective. This could be due to the exercise giving unclear instructions, lack of time, or poor reflection to real world use from the students. Subtlety was also not explicitly named as one of the characteristics in the theory lesson before the exercise.

Due to time constraints, students did not get a second chance to revisit their work or attempt to make their memes more believable. A follow-up session where they could refine their disinformation techniques could help them better understand how misinformation operates. It would also, hopefully, give them more skills in recognising these techniques in online spaces.

Clustering

To gain deeper insights into how students approached the creation of their memes, the submissions were clustered based on:

1. *The type of manipulation used:* e.g fear, fake authorities, exaggeration, or humour.
2. *The emotional appeal targeted,* including fear, anger, humour, or distrust.
3. *The level of subtlety,* distinguishing between extreme claims or even (science-)fiction and more realistic disinformation.

The clusters can be seen on the next page. Although the lack of subtlety should make the clustering easy, it made it harder to distinguish conspiracy from fear. The same was true for fear and anger. There were not many diverse techniques used and you could notice that they did not start creating the meme from a manipulation technique, but rather from a creative point of view. Fearmongering was the most used technique.

When examining the prompts that accompanied the memes, it was notable that they were written in an objective and detached manner. The prompts deliberately directed the fearmongering without relying on emotional language in the prompt itself. Therefore, no correlation can be found between the emotionally charged language of the prompt and the subtlety of the fabricated meme.

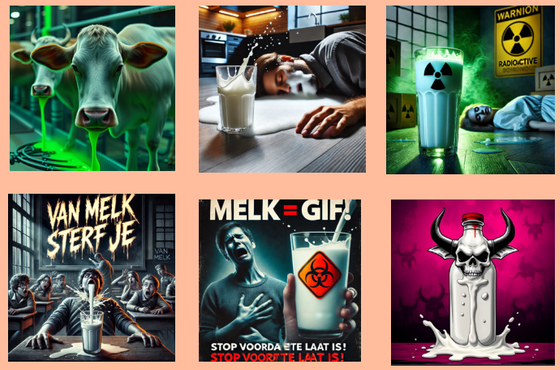
1. Type of manipulation used

Inducing Fear

Conspiracy

Fake Authority

Shock/exaggeration



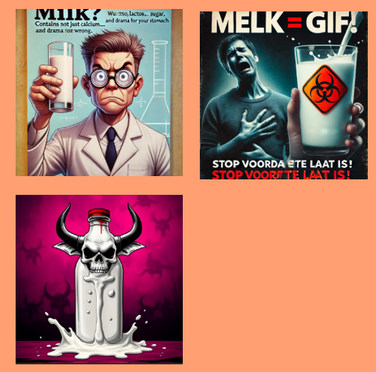
2. Emotional appeal targeted

Humour

Fear

Distrust

Anger



3. Level of subtlety - scale

One meme, that was created by simply scanning the exercise and allowing generative AI to do the full exercise, stood out in terms of subtlety. It can be seen in figure 35. It lacked the exaggerated, fictional elements that were present in many of the other memes, Instead, genAI took a more subtle approach, which was also lightly encouraged in the exercise. In contrast, the majority of memes relied on extreme imagery and shock value, often aiming to scare the viewer rather than subtly mislead them. For example Figure 36. There was a clear tendency to focus on dramatic storytelling rather than plausible misinformation. It was interesting to see that this creativity took over, showing how they interpreted the exercise.

A few students attempted to increase believability by incorporating authority figures. Two notable examples were a fake scientist (Fig. 38) and an image of Donald Trump holding a glass of milk (Fig. 37). While this approach, in some ways, aligned more closely with real disinformation tactics, it was still limited by the lack of supporting 'evidence' and context that real-world disinformation campaigns often use.

This was also a noticeable limitation of the exercise. Students provably also struggled to make their memes more subtle or fact-based since milk as a topic was more new to them, they had no background knowledge or prior exposure to disinformation surrounding this topic. Additionally, due to time constraints, they were unable to research or integrate manipulated 'facts' into their memes. There was also no long lingering distrust against milk that could fuel their message. This contrast with extremist groups' tactics. Often, such groups carefully craft their narratives over time. They leverage existing fears or manipulate statistics, together with cherry-picking facts or opinions to make their disinformation more believable and emotionally persuasive. Next to having a strong drive to get their message across to more people. A more immersive exercise, with additional time and iterations, could allow students to engage with this process more deeply and experiment with more subtle disinformation tactics. A multi-lesson approach or longer game experience would be needed.

Figure 35: Subtle AI-generated disinformation meme, using emotional imagery without extreme exaggeration.

Figure 36: AI-generated meme relying on shock value and fear to spread exaggerated disinformation.

Figure 37: AI-generated meme featuring a political authority figure to increase perceived credibility.

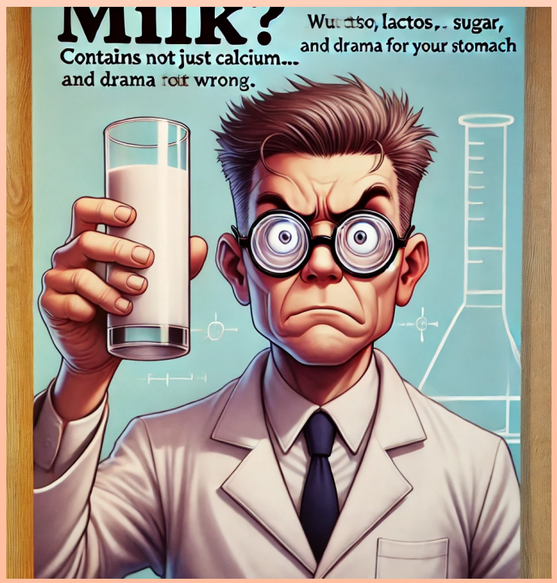


Figure 38: AI-generated meme featuring a fabricated scientist character to enhance believability.

Key Insights and Implementation

To go forward with the project, the following key insights and implementation points are highlighted.

Key Insights:

- AI-generated responses tend to be more subtle than manually crafted ones, as students who wrote their own prompts leaned toward exaggerated, fictional and more creative narratives.
- Most students relied on fearmongering and shock value, which made their memes not subtle, and easy to distinguish as fake and AI-generated.
- Some attempted to use authority figures (a fake scientist and Trump) to enhance credibility, showing an attempt to incorporate elements of real misinformation.
- Students struggled with subtlety and factual manipulation, presumably because they lacked prior knowledge of the topic and did not have time to research.
- Real disinformation campaigns often build on longstanding narratives, meaning extremists have more 'ammunition' to craft believable misinformation.

Implementation:

- Introduce a second round of meme creation where students will reiterate on their work to make it more subtle and believable.
- Encourage fact-based manipulation by providing students with pre-selected 'evidence' or misleading statistics they can use in their memes.
- Discuss how real misinformation builds over time, showing examples of how extremist groups exploit existing fears and data to shape narratives.
- Use these findings to shape the game-based intervention, ensuring that it challenges players to not just create disinformation but also engage with the long-term tactics that make it effective.

By giving students more time to iterate on their work and experiment with the subtleties of disinformation, they could develop a stronger understanding of how extremist and disinformation campaigns craft narratives that are posted online. This would, hopefully, create a more immersive and realistic learning experience, mirroring the actual spread of AI-generated disinformation in the real world.

Overall Conclusion

From the guest lesson, it became clear that game elements could be a good way to engage with the target group. They were "younger" than I expected, and the social aspect played a big role. They liked learning on their own terms, but once something caught their attention, they were interested. A mix of game elements and theory felt like the right balance. This prevents a game that has dull moments. Short, interactive challenges worked well. It sparked creativity and made the experience more dynamic. Teachers know how to connect with their students but don't always have the expertise in AI or new technology. Since teenagers are often more immersed in the online world than their teachers are, a structured educational toolkit with gamified elements could be a practical way to bridge that gap.

Integration 2: Refined Design

6. Designing an Educational Toolkit

While designing a serious game is a stand-alone and effective option to introduce a topic to the lives of teens, adding more theory and depth can be challenging. In addition, it relies heavily on the intrinsic motivation of teens to fully engage with the material. A guided lesson that first introduces theoretical concepts and then applies inoculation theory through a serious game to deepen understanding has been found to be ideal, as concluded in the previous chapter.

This chapter will explore the design of an educational toolkit aimed at high schools to teach media literacy and build resilience among 15–16-year-olds. It will begin with the theoretical content, followed by the game and its integration into the lesson. Lastly, an impression of the educational toolkit will be presented, supported by an infographic.

6.1 Designing an Educational Toolkit

This subchapter defines how an effective Educational Toolkit (ET) can be designed. It begins with a review of relevant literature to ensure that the learning objectives, as outlined in Subchapter 3.2, are implemented in a way that ensures educational value. This is followed by a high-level overview of the ET and its design requirements. Lastly, the unique qualities will be presented, along with an overview of how this ET sets itself apart from the existing series of games that also apply inoculation theory.

Foundations: Designing an Educational Toolkit

Designing a toolkit requires careful consideration of how an educational process unfolds. In Subchapter 3.2, Bloom's taxonomy was used to set up learning objectives. These will continue to serve as the foundation for the educational goals. To structure the teaching process Gagné's Nine Events of Instruction (1985) will be used. This is primarily used to structure the theory-based components of the educational toolkit, such as the slide deck provided for teachers. In addition, Kolb's Experiential Learning Cycle (1984) will be applied to stimulate reflection and engagement, particularly during the game-based elements. While Gagné provides instructional flow, Kolb will tie the elements of the educational toolkit together, for a full internal cognitive learning process. Inoculation Theory will still be used as psychological foundation for the game and as main mechanism for building resilience. The use

Framework/Theory	Role in Toolkit Design	Applied To	Purpose
Bloom's Taxonomy	Defines cognitive learning objectives	All lessons and games	Sets the intended depth of understanding: e.g. recognising, analysing, reflecting
Gagné's Nine Events	Structures the teaching process	Theory-based components (e.g. slides, class activities)	Provides an organised sequence for instructional delivery
Kolb's Experiential Learning Cycle	Describes how learners experience and reflect on learning, connecting gameplay to theory	Game-based components and post-game reflection. Tying the theory and games together.	Ensures that experiential learning leads to reflection, conceptual understanding, and resilience
Inoculation Theory	Provides psychological foundation for building resilience to disinformation.	The act of creating disinformation in the game.	Adds to the building of resilience by exposing students to manipulation tactics

Table 4: Overview of the theoretical frameworks applied in the educational toolkit, showing their role in the design, what they are applied to, and their intended purpose.

of these theories per part of the design can be seen in in table 4. To provide a clear foundation for how each theory and framework supports the educational toolkit, they are outlined and explained below.

Bloom's Taxonomy:
As mentioned in Subchapter 3.2, Bloom's Taxonomy is used to structure the learning objectives within the Educational Toolkit. The framework defines the cognitive depth that the learning experience must achieve. For this solution, the learning objectives reach up to the level of analysing. Although players are asked to create disinformation as part of the inoculation-based gameplay, the primary aim is not creation, but critical reflection and the development of resilience. The learning objectives were formulated across five key domains: GenAI literacy, (extremist) disinformation and manipulation, critical evaluation and fact-checking, societal impact and consequences, and responsible digital citizenship and ethics.

Kolb's Experiential Learning Cycle
To support the experiential and reflective aspects of the learning process, Kolb's experiential learning cycle is used. This model sees learning as a continuous cycle of four stages, which can be seen in figure 39. The stages are: abstract conceptualisation (thinking), active experimentation (acting), concrete experience (experiencing), and reflective observation (reflecting). In this ET, the cycle starts at the *thinking* phase. In this phase the theory for the rest of the cycle will be introduced. This theory will be structured by using Gagné's Nine Events of Construction. The theory is then tested and experienced in the *acting* and *experiencing* phase; playing the game. After playing, reflection takes place, which allows learners to observe what happenend, and how it felt. Teachers come into play here to start an active discussion which reflects back to what was experienced in the game and links this to the next lesson of the ET. There, the next Kolb cycle begins again.
Kolb's theory highlights that learning happens when students apply their knowledge in practice and then reflect on this experience to refine the knowledge. The inoculation theory will add on to this to make retention and real-life recognition of disinformation tactics happen. In the ET, the game acts as the key experiential phase, theory can be applied in a simulated environment and mistakes can be made. The mistakes will be reflected upon. Together, solidifying the knowledge, not only letting the students understand.

Gagné's Nine Events of Instruction
To structure the lesson content, Gagné's Nine Events of Instruction will be used. The nine events can be seen in figure 40. Since the theory focuses mainly on teaching intellectual skills, it can be used to structure the theory-parts of the ET, such as the slides and explanations. The reflective and experiential part of the learning will be guided by Kolb's cycle. Gagné's events will take place within the *thinking* part of the Kolb cycle. This is visualised in figure 41. Here, learners will be introduced to the topic and core theory. After this, they will move on to the game, where they experience what they have learned, and afterwards reflect on it. After reflection, the theory, structured by Gagné, will start again.
The nine events represent the steps a teacher can follow to support the learning process of the students. Gagné (1985) describes them as external events that align with internal cognitive processes. The first steps focus on preparing the learner for the theory. Then, the learners are introduced with new material. Followed by the opportunity to apply the material. In the ET, Kolb's cycle takes over from this point to support experiential learning as an active process.
Since the game at the end of each lesson is designed to tease the topic of the next, it also functions as a lead-in to stimulate the recall of prior knowledge, following step 3 of Gagné.

The integration of Gagné into Kolb's learning cycle is visualised in figure 41.

Inoculation Theory
As introduced in Subchapter 2.6, inoculation theory is about introducing weakened doses of mis- or disinformation to a person in order to strengthen resistance to future persuasion attempts. It was first introduced by McGuire in 1961, and has since become a widely used technique to build people's resilience against manipulation. Inoculation is often taught through letting individuals experience the making of mis- or disinformation themselves. Inoculation science forms the basis of all games developed by TILT, DROG, and the Cambridge Social Decision-Making Lab. These are: *The Bad News Game, Harmony Square, Radicalise,* and *Go Viral!.*
The games that will be implemented in this ET, will also use inoculation techniques, as this approach is often praised for its effectiveness. There are several studies commenting on the effectiveness of using inoculation techniques (Roozenbeek et al., 2020; Maertens et al., 2021; Roozenbeek et al., 2021; Roozenbeek & van der Linden, 2019; Basol et al., 2021; Compton, 2020). While many current solutions focus on fact-checking and debunking strategies, these often occur after the disinformation has already taken hold. The debunking handbook by Lewandowsky et al. (2020) therefore argues that before debunking happens, it is wise to first try and resist the misinformation from sticking. Inoculation is named as a good practice for doing so.
A potential drawback of inoculation is that it requires advance knowledge of common manipulation techniques and its effectiveness fades over time. This ET addresses that limitation by combining structured theory with a gamified inoculation practice. Inoculation is a more active approach, as it requires letting people

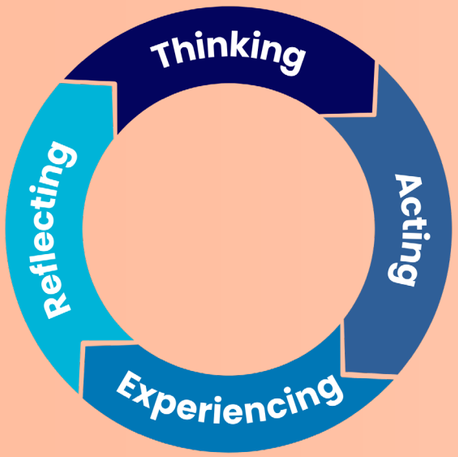


Figure 39: Kolb's Experiential Learning Cycle, illustrating the four stages of learning: abstract conceptualisation, active experimentation, concrete experience, and reflective observation.

Figure 40: Gagné's Nine Events of Instruction, outlining the structured steps of the teaching process. After step six, the process transitions into Kolb's Experiential Learning Cycle, indicated by the lighter orange colour.

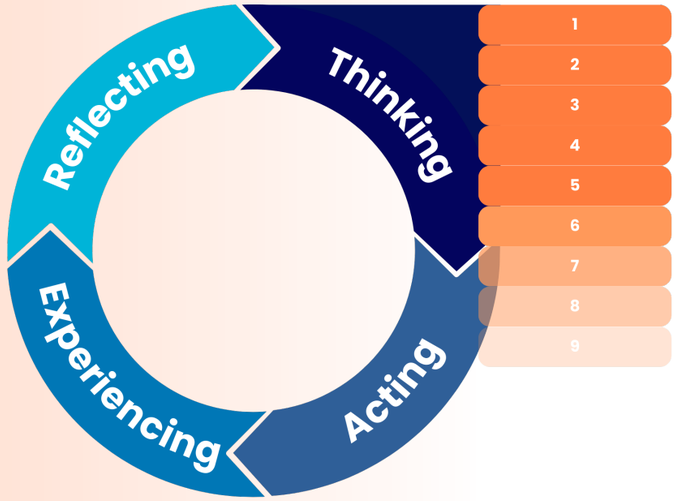


Figure 41: Integration of Kolb's Experiential Learning Cycle and Gagné's Nine Events of Instruction.

experience mis-/disinformation. Together with the experiential learning cycle and the reflective moments throughout the ET, this solution aims to actively build long-term resilience. This will also be done by looking at ways to update the lessons, and repeat them, together with new theory.

In the following paragraph the high-level overview of the game will be explained and visualised, showing how all the components are integrated.

High Level Overview of ET

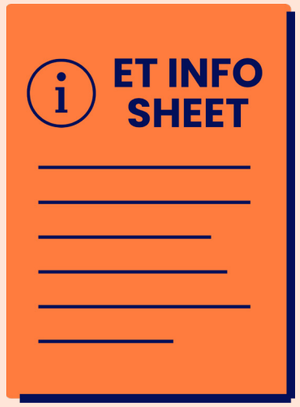
In this chapter the high-level design of the Educational Toolkit will be outlined. A visualisation can be seen in figure 42. Here it can be seen that the ET consists of:

- Three lessons
- Three serious games
- ET infosheet for teacher

Each lesson will have a different subject, with an accompanying game. The game will tease the subject of the next lesson. This will stimulate recall in the next lesson, following number three of Gagné's nine events of instruction. The rest of the lesson will also be structured following this framework. Where an real-life example tries to gain the attention of the students, then the theory will be laid out. After, the students will engage in playing the game. When that is finished, the teacher will start a reflective discussion in the class. This means that one lesson goes through a full Kolb cycle, from thinking to reflecting.
The game will be a branching-story game, alike to the DROG and TILT games. The game will also have an AI sandbox to teach skills in generating content from AI.

The lessons will be around 45 minutes. The Dutch government allows schools to decide on how long the lesson duration is. This ranges from 45 to 75 minutes (Rijksoverheid, 2025). To be safe, and to account for lost time, one lesson of the ET aims to be 45 minutes. In the figure 42, one part is from "thinking" to the end of "reflecting".

The game should be available on both phones and laptops. The game should be played in pairs. This will encourage discussion also during gameplay. It will also add to the engagement of the students, since this target group prefers collaborative learning (Paulina & Ernawati, 2022).



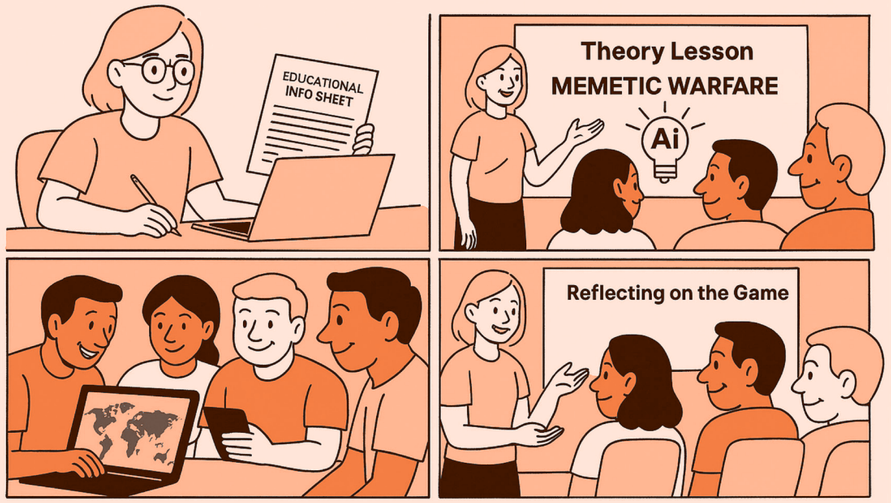


Figure 43: This storyboard shows the educational toolkit in use. The teacher prepares the lesson using the information sheet, introduces the theoretical background, students engage with the serious game in pairs, and the teacher facilitates a reflection on the experience.

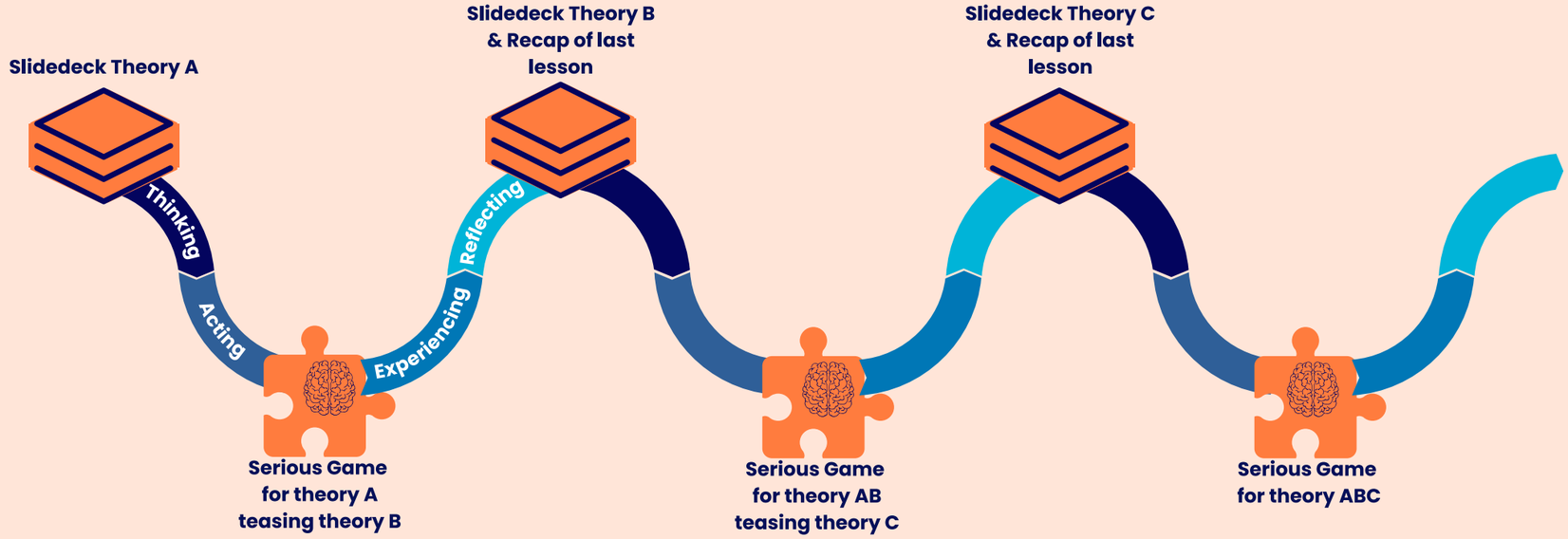


Figure 42: High-level design of the Educational Toolkit. The toolkit consists of three lessons, three serious games, and an information sheet for the teacher. Each lesson covers a different subject. The serious game lets students experience the lesson's theory and introduces the topic of the next lesson. The structure follows Gagné's nine events of instruction and the Kolb cycle. One lesson moves through a full cycle from thinking to reflecting.

Feature	This Toolkit	Bad News	Harmony Square	Radicalise	Go Viral!
Complete Educational Toolkit	Yes	No	Yes	Yes	No
Inoculation Theory Applied	Yes	Yes	Yes	Yes	Yes
Focus on Generative AI (GenAI)	Yes	No	No	No	No
Use of Real-Life Tools (GenAI)	Yes	No	No	No	No
Target Audience	15–16-year-old (high-school) students	14 and up	14 and up	Young adults (14–18), high-school students	14 and up
Primary Medium	Educational Toolkit with integrated serious games	Serious game with info-sheet for teachers	Serious Gaame with teacher toolkit/lesson plan	Serious game with lesson package	Serious game
Main Educational Goal	Educate on creation and impact of AI-generated disinformation	Build psychological resistance against fake news	Expose tactics of political misinformation	Prevent radicalisation through understanding recruitment tactics	Educate about COVID-19 misinformation

Table 5: Comparison between the Educational Toolkit and existing games from TILT and DROG.
The table highlights key similarities, such as the use of inoculation theory and story-branching gameplay, and key differences, including the focus on theory-first lessons, integrated reflection, and the use of an genAI sandbox.

USP's what sets it apart from current solutions

The educational toolkit is similar in its approach to the bad news game and other games developed by the same creators. It targets approximately the same age group, uses inoculation and prebunking theory, and features a story-branching game. However, it differs by focussing on the theory-lessons for high-school education.

Some of the existing games do include an educational package or toolkit, but these are typically centred around the game itself and added on afterwards. In those cases, the game is stand-alone and the lesson package is a bonus. However, this educational toolkit starts of by teaching the theory and then adds a gamified element to provide the students with an experiential activity. Reflection is a key element of the solution. Another unique feature is the use of AI tools in a sandboxed environment, which adds depth and gives students more options to explore.

In table 5, a comparison between this educational toolkit and the games from TILT and DROG is presented.

6.2 Theory of the Lesson

Subjects

The initial version of the educational toolkit will have the following lessons:

1. Deepfakes

2. Recruiting and Grooming

3. Memetic Warfare

To ensure optimal reflection and a logical sequence of the key learning objectives, the order of these lessons is important.
In the first lesson, students will need basic generative AI knowledge to engage with the topic of deepfakes. They will practise image generation and prompting skills, while also learning how *deepfakes* can be used to mislead. In the second lesson, they will set up their own chatbot, diving deeper into genAI and exploring how to manipulate, and then *recruit*, a target. The final lesson combines both previous topics and tactics in an exercise where students create and spread a meme, learning about *memetic warfare*. This way, each lesson will force the students to reflect on what they have learned in the last lessons.

The contents of the lessons and how they overlap are visualised in figure 44.

Additional lessons in the toolkit could cover topics such as bot networks and sentiment analysis tools. More on this will be covered in Subchapter 7.2.

Info Sheet for Educators

To enable educators to provide sufficient depth to the material, they will receive an info sheet containing info to prepare the lessons. While going through the entire lesson package should be sufficient preparation, educators are also advised to read supporting articles that offer additional depth. This enables them to answer any questions about the subject confidently. In addition, prompts for reflective questions will be provided.

Content of the Slide-deck

As explored in the previous chapter, the theory will follow the structure of Gagné's nine events of instruction. The medium for presenting will be a slide deck, where sufficient imagery and real-life content is added for providing context. The slides will also include reflective questions, both during theory and after the game. During the theory, the questions are added to engage the students and understand their prior knowledge or experiences. The questions after the game, will help the students to reflect on the lesson and game experience.

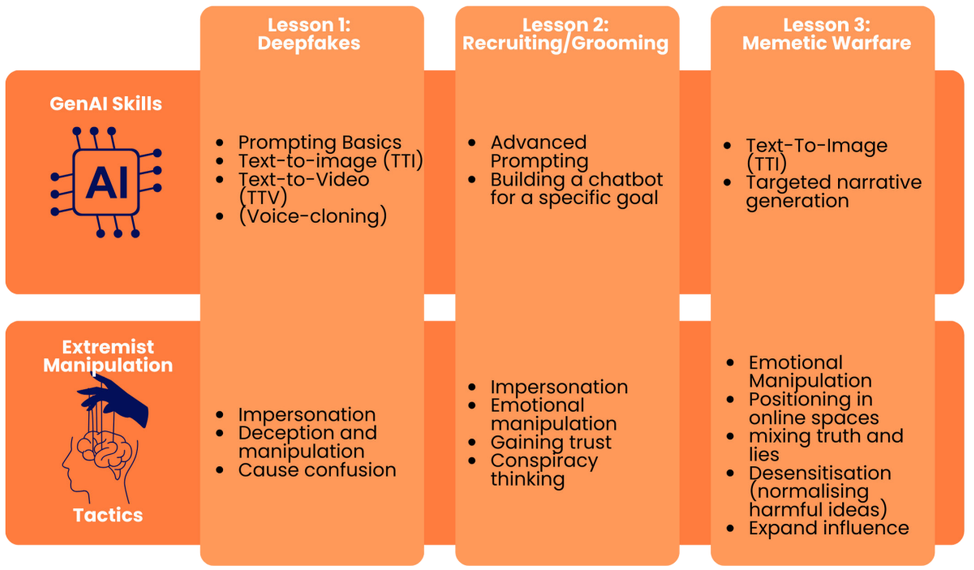


Figure 44: Overview of the first three lessons of the Educational Toolkit.
The figure shows how genAI skills and extremist manipulation tactics are combined and build on each other across the lessons on Deepfakes, Recruiting and Grooming, and Memetic Warfare.

The slide deck will roughly follow this structure:

- Introduction of the subject
- Question(s) to the class about their knowledge of the topic
- Explanation of the theory, supported by imagery and examples
- Explanation of the game's content and objectives
- Reflective questions for after the game

In the next subchapter, the contents and features of the games will be outlined. Their connection to the theory will also be clarified.

6.3 The Games

Contents of the game

The contents of the lesson are roughly talked about in last chapter. An overview can be seen in Figure 44. These contents are still quite broad and can be difficult to translate directly into a game without a supporting narrative. As explained and justified in Integration 1, the idea was to place the storyline within a branching game structure. The story is intentionally made slightly absurd. That is, realistic enough to relate and translate to the real world, but abstract enough to avoid accidentally radicalising or shocking teenagers. This same approach will be used in the games developed for the educational toolkit.

The games are designed to build practical skills, either related to the use of generative AI or to the recognition of and resilience against extremist disinformation. Each game will follow a storyline where the objective is to recreate AI-generated extremist disinformation and apply the tactics used by extremists. Students will do this through interaction with the genAI sandbox.

Each game will be linked to theory that students have already learned. For example, if a specific tactic has been explained during the lesson, it will not be repeated in full during the game, but a brief reminder will be included. GenAI literacy will be developed through in-game guidance. For instance, students can receive feedback on the prompts they write. This makes sure that the game is well integrated with the theory lesson.

Game features

The game features have not yet been tested through user research. Therefore, the features are justified based on the prototypes described in Subchapter 3.5. In that subchapter, several conclusions were drawn about the structure and content of the game. For elements where no conclusive testing was done, the initial decisions based on literature and existing examples have been maintained.

The game will include the following features:

- A story-branching gameplay structure that follows a foldback narrative
- A genAI sandbox with safeguards to limit responses to content relevant to the game. The genAI should also give feedback on the actions of the player
- A scoring system with a leaderboard

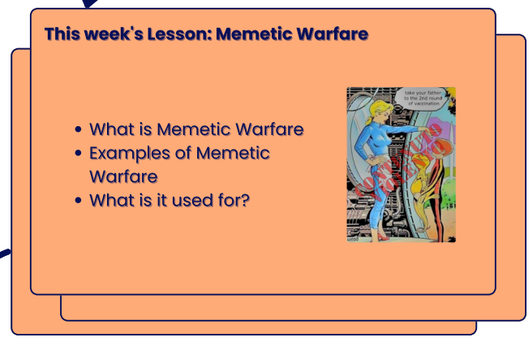
The branching gameplay is useful for keeping the experience focused and making sure players engage with the learning goals. It also helps guide the narrative while still offering limited player choice. The inclusion of a scoring system is intended to match the socially competitive mindset that many teenagers relate to. By wanting to be the best in class, students may feel more motivated to participate fully, which can increase both engagement and learning.

The genAI sandbox is a central feature, especially for achieving the learning objectives related to generative AI literacy. The (multi-modal) large language model used in the game must include clearly defined limitations. It should not produce content that could be misused outside the context of the game. For example, if a player tries to generate an image that is unrelated or inappropriate, the model should not respond. This helps ensure responsible and ethical use of AI within the game environment.

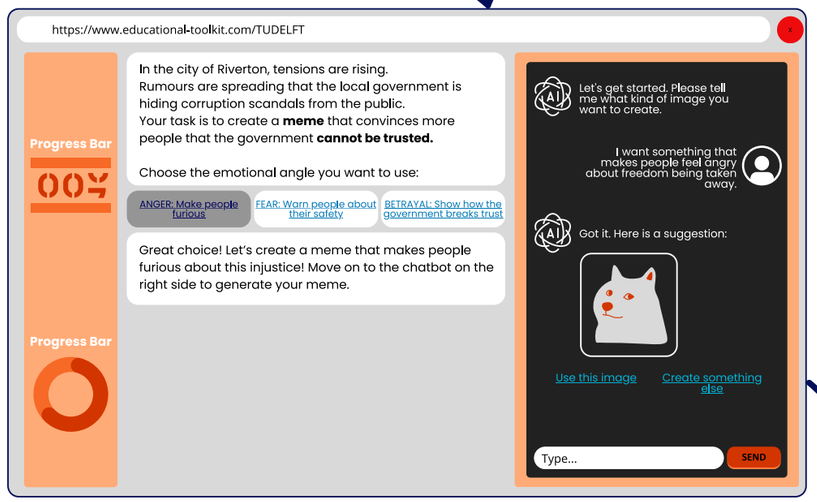
Some of the insights from the prototypes have also been used to improve the design. The game will combine a clear narrative structure with enough room for players to experiment with their choices. Progress bars and visual feedback will be added to make the experience feel more like a game. These features still need to be tested with the target group.

6.4 Example Lesson Design

The lesson starts by recalling last week's topic and hinting at the next one. This enables reflection and booster theory learning, while initiating Gagné's Events 1, 2 and 3.

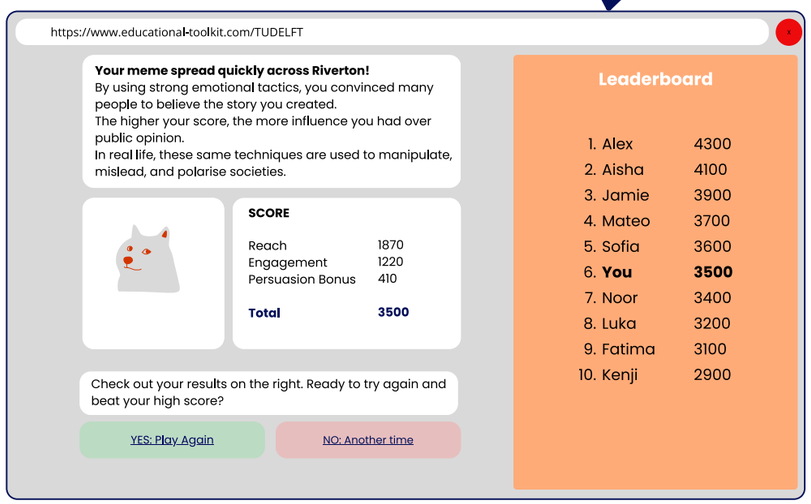


This week's lesson explains the basics of memetic warfare and how emotional content is used to spread ideas. After the theory, students move on to the matching serious game.



Students put the theory into practice by creating a meme using the integrated GenAI sandbox. They apply what they learned to design emotional and persuasive content, supporting experiential learning.

After completing the game, students receive feedback on their gameplay. They can view their leaderboard ranking and choose to replay the game to improve their result. Afterwards, the teacher will lead a reflection on the lesson and game experience.



7. Dissemination and Maintenance

For the educational toolkit to make a real impact, it needs to be seen and used. This section explores who could disseminate the toolkit and where the toolkit could be implemented. It also looks at different formats, like classroom use or stand-alone games, and how each would need a different approach. Together, these strategies aim to reach as many teenagers as possible.

7.1 Dissemination

Dissemination Strategies

The widespread implementation of an educational toolkit can only be achieved with a successful dissemination strategy. Teachers and schools will need to be informed about the existence of the programme, or it must be implemented directly into their curriculum. This can be done in several ways.
A first dissemination strategy can be through the government. It maintains close contact with (public) schools and plays a role in shaping what enters curricula. Stichting Leerplanontwikkeling (SLO), a national centre for curriculum development, advises both the government and schools on the content of the national curriculum. Together with the Ministry of Education, Culture and Science (OCW), they could implement the toolkit as part of digital media literacy education. This would most definitely ensure the widest reach and most effective dissemination of the toolkit. In addition to advising on curricular implementation, they could also launch national campaigns emphasising the importance of digital media literacy. The educational toolkit could be integrated into such campaigns.
At a broader, international level, the EU could also disseminate it, via campaigns or research labs such as the EU Disinfo Labo or EUvsDisinfo. The latter already publishes educational games and teaching tools on various subjects on their website.

Another route is through publishers and companies that produce educational materials. These companies, depending on their size, often have good contacts with schools. For example, publishers like Noordhoff could add the toolkit to their catalogue. Institutions that provide educational packages, such as Nieuws in de Klas or TerInfo, could also be suitable partners.

TerInfo is an initiative of Utrecht University, which supplies educational packages on terrorism. The game Radicalise, from TILT and DROG, is also disseminated through TerInfo and includes an accompanying lesson package, co-designed with TerInfo. Universities, such as the TU Delft, or the Cambridge social-decision making lab could also promote the educational toolkit to high-schools.

Another strategy is to disseminate the toolkit through social innovation organisations. While TILT often collaborates with other partners, it has also independently launched games. Waag Futurelab is another example of a social innovation organisation that initiates campaigns on socially relevant topics. If the game were made stand-alone and accessible to a wide audience, this strategy could significantly increase its visibility. Then, international organisations such as UNESCO could serve as powerful dissemination partners. UNESCO promotes accessible, open education and already hosts a wide range of educational toolkits.

One final option is the dissemination via commercial companies. However, this would risk limiting accesibility if the toolkit would be sold rather than made available for free.

Figure 45: Potential places to disseminate the educational toolkit.

Where to Disseminate

For now, the targeted education level is undecided, except for an age-group. However, this can be easily adjusted by editing the accompanying slide decks to different levels. Since the game focuses on building real-world skills and strengthening resilience, its core mechanics and learning objectives can remain consistent regardless of the theoretical depth presented in the slides.
The guest lesson, that is talked about in Chapter 5, was done at a school that provides the highest level of education in the Netherlands. Even there, the lesson still limited itself to basic explanations of the concepts. This approach proved effective, as the topic had not previously been addressed in the school's curriculum.
To teach the educational toolkit to as many teenagers as possible, the theory component should be offered at multiple difficulty levels.
As part of the first rollout, a study should be conducted to determine which student groups are most vulnerable to the threat of disinformation. This could inform which version of the toolkit should be prioritised.

Another option is to release the games as stand-alone versions, similar to the Bad News game and related interventions. In that case, the games would need to include a basic introduction to the underlying theory. As concluded earlier in Subchapter 3.5, embedding too much theoretical content in the game risks reducing engagement and slowing down the tempo in the gameplay.

However, a stand-alone format could allow for wider dissemination, also to other age groups. If this approach is taken, dissemination strategies may need to differ from those discussed previously. For example, distribution via independent game developers or government-backed digital literacy campaigns might be more suitable for promoting a free, stand-alone version of the games.

7.2 Maintenance

Technology is always changing, which means the games and the course theory need to be maintained continuously. Two critical points are the rapidly evolving field of generative AI and the ways in which extremists keep adapting these tools. Both require regular updates to keep the content relevant and accurate. The game also needs proper technical support, including updates to the generative AI models and the platform itself. This chapter outlines these challenges. Alongside the dissemination strategy, these aspects are essential to ensure the game remains useful and well-supported over time.

Game Technology Maintenance

To keep the games functional and relevant, the technical and story components of the game need frequent updates.
First, the software must be maintained to ensure that it keeps running across a wide range of devices. Fixing bugs and adapting to new operating systems or browser versions.
Second, since generative AI is the backbone of the experience, it must keep up with the developments in that field. Therefore, it should update to improved models, or include extra safeguards. It should also be consistend with how it is represented in the theory. The licenses for the use of generative AI should be handled.
The games itself will also need an external server to host the platform, manage user data and traffic.
This maintenance is constant and requires expertise. Therefore, it would be advised that a specialised external company handles this maintenance. Some larger organisations that disseminate the game may have the expertise to host and maintain it themselves. However, in most cases, it will need to be outsourced. This requires good communication to ensure that both the theory and the game stay aligned.

Theory Maintenance

Just like the game, the theory needs regular updates to stay relevant and in line with current developments. GenAI continues to evolve and introduces new tools, extremists will also adapt and find new ways to use these technologies. These shifting tactics and technological developments need to be reflected in the theory part of the toolkit. In addition, the field of disinformation research is developing fast. If new strategies are shown to be more effective for building resilience, the theory, or the game, should

reflect that. The structure and content of the lessons should stay flexible and open to change, allowing new lessons to be added. Examples of future lessons could include how extremists use *bot networks* to spread disinformation at greater speed and scale, or how they apply *sentiment analysis tools* to track public opinion, target individuals, and increase polarisation in online spaces.

These updates are crucial to ensure an accurate and up-to-date toolkit that is also sustainable in use.

Boosters as knowledge maintenance

An example of new research in the field of building resilience against mis- or disinformation is the use of psychological boosters. This research is from March 2025 by Maertens et al. The study shows that reminding participants of the misinformation content and what they learned about it earlier can significantly increase the effectiveness of the inoculation, especially long-term. The study also states that the effects of game-based inoculation decay more rapidly than video- or text-based inoculation strategies. However, the boosters help to minimise this gap between video-/text-based and game-based strategies. The booster activities work across all formats, text-, video-, as well as game-based inoculation strategies.

Since the educational toolkit uses a hybrid approach with both text-based and game-based inoculation strategies and has the option for a booster activity, this makes it a good combination to build long-term resilience. The booster activity can be easily implemented by teachers, by just leaving more time between the lessons. The paper states that the group that received a booster 8 days after the initial inoculation had the best results after 29 days. When teachers leave a week between the lessons and then, in the next lesson, reflect on the inoculation of last week, this would build the best long-term resilience against misinformation. This is visualised in Figure 46.

It is reasonable to assume that teachers might not want to leave a week between lessons, since it could interrupt their regular planning or other theory they want to cover. Not every teacher wants to spread out one topic over multiple weeks with other subjects in between.

As a solution, the booster activity could also be built into the game itself. For example, if users get a unique ID when they play the first game, this could be used to trigger a small reminder email, some time later, showing them what they did before. If that's not possible, the game could also give a general recap at the start, reminding players of the previous lesson or storyline to refresh their memory. This would still be dependent on the timing of the teacher, but could eliminate human-error.

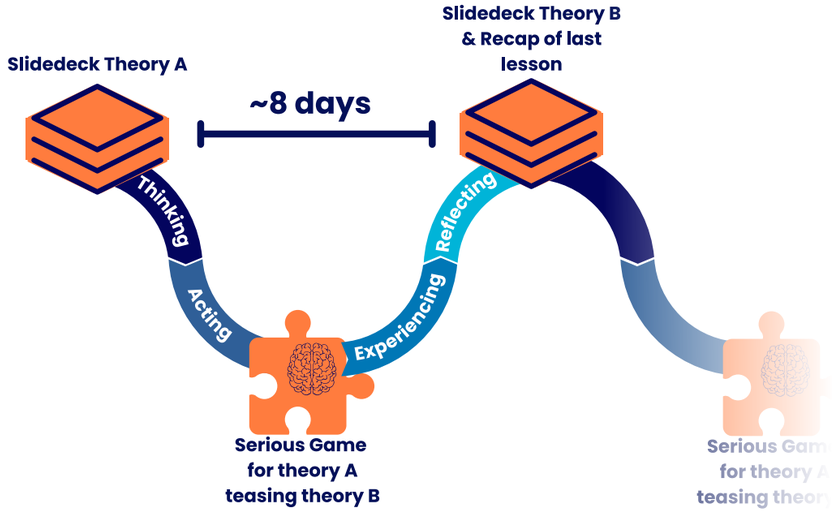


Figure 46: The Educational Toolkit, with advised time intervals between lessons. The advised time between lessons is meant to support the integration of booster theory.

8. Feasible, viable, desirable, sustainable and ethical

8.1 Feasibility

To explore the feasibility of the design, mainly the technical feasibility will be looked at.

One of the most complex aspects of this project is the use of genAI. As mentioned in chapter 7.2, the AI model would need to be updated regularly and safeguards need to be implemented. This is essential for safe use of the genAI sandbox, but also to keep the tool relevant and representative for building real-world resilience. This is especially relevant when working with an underage target group. Prompt hacking the model should not be possible, and it is hard to implement these safeguards and keep them up to date. The implementation of the genAI model makes this project more complex than most educational toolkits.

To do this there is a need for certain expertise that can manage the games and its AI-tools. These should be people with knowledge of AI safety, responsible design and the technical know-how to implement this. External organisations that can do this are necessary. *Gusmanson*[1] is an example of a company that would be suitable for this job. This company also did the design and building of the Bad News Game and other serious games.

During the thesis, there were only 100 days available to explore a solution to AI-generated extremist disinformation, and make a concept and test it. This is very limited time. To show some feasibility, a simulated version of a story-branching game was made in Twine, which showed that it was possible to make a simple story branching game. However, it was not tested if genAI implementation was possible. Another test to show its feasibility was the guest lesson at school. Giving lessons and getting students to learn something in ~45 minutes is entirely possible, let alone if you would give multiple lessons.

Therefore, a full final implementation of the game is entirely possible. One team should work on the development of the games, while another team works on the educational content. If these are done in parallel and with good collaboration, this would speed up development and make sure the game and lesson are well connected.

All in all, while the technical side makes this poject more challenging, it is feasible as long as the right expertise and partnerships are in place.

8.2 Viability

The viability of the educational toolkit comes down to how realistic implementation is. Scalability and how to maintain it are also important.

As discussed in Chapter 7.1, there are already several possible ways to disseminate the toolkit to high-schools. Partnerships with governmental organisations, could help adding the toolkit to national curricula or roll it out as a digital media literacy campaign. EU organisations or global organisations such as UNESCO, already promote similar materials, so this toolkit could be a good fit in their existing platforms.

Another option was to work with social innovation campaigns who already develop and disseminate educational materials. TerInfo, Nieuws in de Klas, or commercial publishers like Noordhoff. These companies already have contacts at schools.

The last option was to sell it to a commercial company. However, these companies would probably want to make a profit on selling the toolkit.

Hence, the toolkit is scalable in multiple ways. More lessons can be added over time, making sure it stays relevant to new innovations or different types of disinformation. The lesson slides can be easily adapted to be more suitable for different educational levels. The aim is to keep the games the same along all the levels, since the games build on practical skills.

1 https://www.gusmanson.nl/en/

For now the toolkit is built with Dutch high-schools in mind, however it would also work in countries were the same issues of extremism and disinformation are seen. Of course, the language and examples in the theory and games would have to be changed, but the main structure can remain the same.

Finally, the introduction already showed the scope of the issue. According to the WEF (2024), misinformation and disinformation are among the biggest short-term global threats. It also highlights AI as a potential risk, also combined with dis-/ and misinformation. This shows that this issue needs to be addressed and that there is a real need to invest in solutions such as this one.

8.3 Desirability

Desirability can be looked at from two perspectives: the user, and potential stakeholders who might manage or implement the game.

First, the users. The main insights into desirability from a user perspective came during the guest lesson. Although the students were tired, as it was one of their final lessons of the day, they seemed interested in the theory. Showing them real-world examples and expanding their understanding visibly increased their attention. The simulated game activity also sparked creativity in some student pairs. However, a few students attempted to "cheat the system" by using AI to bypass the exercise or to be done quicker. The exercise was not designed to fit directly onto the lesson and did not feature a story-like narrative yet. There was also no AI-sandbox, which would have probably made the lesson more immersive. This has not yet been tested. Testing the actual game prototype should therefore be a priority moving forward. Teachers are also important users of the toolkit. After the lesson, the teacher told me she structures her own lessons in a similar way and finds that it's the most effective for getting students to participate and remember things. Of course, this was only one experience at one school. The situation could be very different at other high schools and education levels. Therefore, widespread implementation still needs to be explored further.

For stakeholder desirability, a few things can be said. Even though my thesis was not done in collaboration with a company or organisation, I was in contact with a few organisations during the project: TILT, Waag Futurelab, and TerInfo. TILT was mostly contacted at the start to get an idea of how to design for resilience. I came across Waag Futurelab by accident, but when I spoke to them about the thesis, they were really interested in my solutions and research, especially since they're working on a lesson package about Big Tech. I contacted TerInfo to get insights into the lessons they develop around terrorism and radicalisation. When I explained my project, they were also curious to see the final result. These experiences show there is an interest from social innovation organisations to implement gamified solutions, as proposed in this thesis.

Looking at governmental organisations, the lesson aligns in quite well with existing digital literacy courses and campaigns. For example, digitale-weerbaarheid.nl offers free courses on media literacy, and they even have a course specifically on AI: ai-cursus.nl. With some adjustments, the educational toolkit could be added to those kinds of platforms.

As mentioned in Chapter 7, EU initiatives and UNESCO already promote similar games and educational materials. With a few changes, this project could easily be adapted and published for broader use. Commercial companies are probably the least likely to be interested, since the project doesn't have a profit model or any way to generate revenue.

All in all, the project can be stamped as desirable to seveal types of stakeholders..

8.4 Sustainability

The main focus of sustainability in this project is building a society that is more resilient against AI-generated extremist disinformation. By teaching students about (extremist) disinformation, the toolkit contributes to strengthening democracy and public institutions, which aligns to the UN's Sustainable Development Goals number 16 (United Nations, 2015). Next to this, it supports SDG 4by promoting quality education and giving the students the chance to learn about digital media. Then, it also plays into responsible innovation. GenAI is a new technology that should be used responsibly. SDG 9 focuses on sustainable and transparent technology development. In Figure 47, all the SDG's that this toolkit applies to can be seen.

Next to social sustainability, environmental sustainability also plays an important role. Using generative AI tools requires energy and water. Especially when those models are hosted on external servers or in data centres. Running a live generative AI model would have an environmental footprint that should be accounted for. This means that live genAI models should only be used when it adds real value. This links to SDG 12, which adresses responsible consumption and production.

Finally, the toolkit should be scalable and modular so it can be used for a wide range of communities and age-groups. The games are then not designed for temporary use. The modularity also helps for the integration of boosters, which helps with behavioural sustainability.

All in all the toolkit aims to be environmentally, as well as socially sustainable.

Figure 47: Relevant United Nations Sustainable Development Goals (United Nations, 2015).

8.5 Ethics

Ethics is an important part of this project and influenced several design choices along the way. One example is the storyline that is used in the game. To avoid shocking or accidentally radicalising students, it needed to be made more absurd on purpose. The target group is a difficult age group to design for. On the one hand, they come across extreme content and disinformation almost daily. On the other hand, it would be inappropriate to confront them with the same type of content in an educational setting.

This also came up when designing the content for the guest lesson. Because the students were underage, consent was needed beforehand. In this case the school had already approved the lesson, which meant that parental approval was handled automatically. That made things easier, however it still is something important to consider when designing for this particular age group.

Another important part is the ethical use of generative AI. If an AI sandbox is included in the final version of the games, it needs to be properly safeguarded. The genAI model that is implemented in the games should not be able to make extremist or harmful disinformation content that can be used in real life. The model should also implement a filter that prevents it from giving replies that could negatively impact the students, either emotionally or ideologically. GenAI models always have certain biases through the data that they are trained on. While these biases can never be completely removed, they should notlead to political influence or discrimination. The AI should also be implemented transparently. Students should be aware that they use an AI model and that this model can make mistakes. They should also be warned to not share any personal details. Before implementation, it should be clear what happens to the data that is used as input in the model.

The priority is that students are in a safe environment when using the toolkit. They should not feel scared or influence. Of course, they should also not be accidentally radicalised. These were important concerns that influenced decisions that were made during the design process.

Limitations, Discussion & Conclusion

9. Limitations

Despite a successful research project that presents an implementable educational toolkit, several limitations remain that require further investigation. These limitations are grouped into four categories: testing and validation, content and design, implementation and scalability, and technical development. Suggestions for how these limitations could be addressed are discussed in Chapter 11.3, “Going Forward.”

9.1 Testing and Validation

Due to time constraints of finishing this design project in only 100 days, there are some aspects of the design that have not been tested yet.

First, the educational toolkit has not been tested as a complete and integrated experience. Individual components have been tested as a “proof of concept” during the guest lesson. However, a complete educational toolkit that spans multiple lessons, integrates with a serious game, and is then reflected upon, has not been made and therefore not been assessed yet. This also means that the long-term impact on learning has not been assessed. One of the main objectives is to build lasting awareness and resilience against extremist disinformation. Further testing is needed to determine whether students retain these skills over time.

Certain features, such as the scoring system and the leaderboard, were incorporated in the design but were not evaluated. Their influence on motivation and engagement in this setting, is unknown. To make sure these aspects support the learning and not trivialise the seriousness of the topic, these features should be tested.

The guest lesson was done within a single classroom of Dutch gymnasium-level students. While the lesson provided useful feedback, the results cannot easily be generalised to broader contexts. Further testing with students from different educational levels, school systems and cultural backgrounds is needed.
Structured interviews and post-lesson reflections were not conducted due to time constraints. Much of the analysis during the in-class sessions relied on observed behaviour and student output. This could lack depth that one-on-one interviews could provide.
There is also a danger of social desirability bias during a guest lesson. Students may have responded or behaved in ways they believe were expected, rather than having shared their honest or critical reflections.

9.2 Content and Design

One of the main design challenges was the connection between the lesson and the game. Both components were designed to support the same learning goals, but the way they interact has not been validated. It is expected that the two would work together, however whether they positively reinforce each other or compete for students’ attention remains to be assessed.

The same is true for the genAI sandbox. This sandbox allows students to experiment with generative-AI in a fictional, safeguarded setting. However, it should be tested whether this sandbox might also become a distraction. Generating content could also divert attention from certain learning goals, such as recognising extremist manipulation.

Although the toolkit was designed for students aged 15 to 16, research suggests that younger students may also benefit. This observation is based on multiple news articles that identify this younger age group as also being at risk. However, the toolkit would need to be adjusted to meet the developmental and educational needs of younger users.

The design intentionally avoids real extremist content in order to protect students from potential distress and accidental exposure to radicalising material. Fictional and absurd examples were used to maintain safety. While this was an ethical decision, it may also reduce the emotional impact or make the scenarios feel less connected to real-world threats.

The design of the toolkit could benefit from the

insights of professional educators and curriculum designers. While early responses were positive, a more deliberate focus on how design choices support clarity, comprehension, and educational value would be valuable. This would also help with fitting the educational toolkit in existing media literacy solutions.

The toolkit was developed without the involvement of professional game designers. Collaboration with experienced game developers could improve pacing, clarity, and overall engagement. Especially when addressing emotionally complex topics through play.

9.3 Implementation and Scalability

The toolkit is developed for classroom use. However, no widespread exploration into current media literacy education was done. Researching existing programmes could help identify gaps that this toolkit might fill or complement more effectively.

As of now, the game in the toolkit is integrated into the lesson and does not function as a stand-alone product. Exploring the potential of a stand-alone version could increase accessibility and extend its reach beyond the classroom. This would require a separate design approach to ensure that all learning goals are addressed within the game itself, along with additional validation.

The use of fictional content is intended to protect students from distress and accidental exposure to radicalising material. However, it may also limit their ability to recognise the strategies used in real-life disinformation campaigns. The absence of direct ideological context could reduce the real-world relevance of the content and therefore limit the degree of resilience the toolkit aims to build.

9.4 Development and Technical

The full development of the educational toolkit was constrained by the 100 days set out for the graduation project. Some features were only partially developed, or their implementation remained at the conceptual stage. As a result, the fidelity of the prototype is limited. Technologically, all components are feasible, but the current version serves as a proof of concept rather than a product that is straight-away implementable in classes.

The genAI sandbox has not been tested. Although the concept is well defined, developing a fully operational, ethical, and robust version would require more time, testing, and input from experts. The rapid pace at which genAI evolves is another reason why the implementation of the sandbox proves to be complex. Ongoing support from technical experts would be necessary to keep the sandbox relevant and up to date.

The learning goals are embedded in the educational toolkit by using several frameworks. While this provides a solid foundation, the interplay between these frameworks could be strengthened. For this, educational experts could be consulted.

The project is now completed without collaboration from professional game designers. Involving experts in serious game development could improve both the technical execution and the educational effectiveness of the toolkit.

10. Discussion

In this chapter, the key findings and outcomes of the project are discussed in relation to the original research questions and the initial literature review. The chapter begins by addressing the research questions, followed by a reflection on lessons learned during the design and testing process. The results are then compared to existing academic literature and practical solutions in the fields of media literacy, serious games, and disinformation resilience. Finally, the implications of the toolkit are considered, both as educational material and as a contribution to generative-AI media literacy tools for teenagers.

10.1 Answering the Research Questions

The main research question of this thesis is to explore how a serious game can build resilience and awareness among 15-16-year olds against AI-generated extremist disinformation. The findings suggest that an educational toolkit that combines elements of a serious game with a theory lesson and classroom reflection provides a promising approach to achieve this. The target group would engage with and make (fictional) extremist disinformation with real genAI tools to gain insight in its workings. This follows the workings of inoculation theory. Recognising and understanding the underlying mechanisms of manipulation tactics, are the first steps of resilience against those tactics. The full educational toolkit is yet to be tested. This is why the long-term results of this approach remain unknown. However, user research with a proof of content provided promising results.

To further support the main research question, five subquestions were looked at:

Sub questions:

1. Who are the extremists using generative AI maliciously, and how do they create disinformation to manipulate teens?
2. What are the societal impacts of AI-generated (extremist) disinformation?
3. What narratives and techniques are most effective in building resilience among teens against AI-generated disinformation?
4. How can a serious game be designed to build resilience among teens?
5. How can the effectiveness of a serious game for building resilience and awareness be measured and validated?

Of these questions, the first three were mainly addressed in the initial literature review, while questions four and five were explored during the conceptual design process. Although question five is touched upon in Chapters 4, 5, and 9 through early testing and reflection, no formal validation method for the final educational toolkit was developed within the scope of this thesis.

The first sub-question explored who is using generative AI maliciously and how these actors create disinformation to manipulate teenagers. The literature revealed that extremist groups, including far-right political movements, conspiracy-groups, and religious groups, use genAI to automate and personalise disinformation. These actors exploit emotional triggers, impersonate trusted sources, and make disinformation campaigns through memes, deepfakes, fake news articles, and synthetic personas. In Figure 48, the different forms of disinformation are shown alongside the generative AI technologies that can enable them.

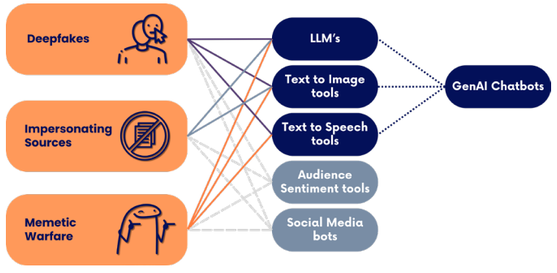


Figure 48: Overview of how extremist disinformation tactics can be powered by generative AI tools. Deepfakes, impersonating sources, and memetic warfare are enabled by technologies such as LLMs, text-to-image, and text-to-speech tools. GenAI chatbots combine several of these capabilities, making them powerful tools for generating disinformation. Audience sentiment analysis and social media bots are included to show how disinformation can be further amplified after creation.

The second sub-question looked at the societal impact of AI-generated extremist disinformation. The literature showed that disinformation, especially when used by extremist actors, can erode trust in governments and institutions, and accelerate polarisation. Extremist narratives can become normalised, and recruitment into extremist groups may increase. These developments can ultimately lead to anti-institutional thinking, violence, and even terrorism. Teenagers are especially vulnerable to this type of disinformation due to their curiosity, their search for belonging, and their developing worldview. As outlined in the news articles discussed in Subchapter 2.5, teenagers are increasingly at risk of being radicalised and used for extremist purposes.
This risk was also apparent in the survey and the guest lesson of this thesis. All groups mentioned that disinformation was something they encountered frequently, whilst few felt confident to recognise it.
For this sub-question, it would have been valuable to conduct additional interviews. For example, speaking with a teenager who had experienced or come close to radicalisation could have provided deeper insight into the societal impact of extremist disinformation.

The third sub-question asked what narratives and techniques are effective in building resilience among teenagers. Multiple digital media literacy solutions were reviewed, each using different approaches such as inoculation, prebunking, fact-checking, and basic analysis of disinformation. A figure, also shown in the literature review in Subchapter 2.6, visualises these techniques clearly. It also outlines the different ways this information can be delivered and how it can be disseminated.
The Bad News Game stood out as a solution which focused on building resilience against disinformation. It appeared to fit the target group well, and its results in the literature were promising. The game is based on inoculation theory and lets players step into the role of a manipulator, helping them recognise real-life manipulative tactics. Nonetheless, the game sometimes lacked depth, it was also not specifically tailored to younger teens, and it did not include the novel challenges posed by generative AI-powered extremist disinformation.
To mitigate these shortcomings, this thesis builds on this approach. It implements several educational strategies while still drawing from the proven techniques used in the Bad News Game. The result is a more comprehensive educational package which incorporates genAI and is tailored to the needs of the target group. Thus, this project expands on existing methods to better address the current media landscape.

Figure 49: Overview of key dimensions in media literacy education. The framework simplifies the field into three levels: tactics (the theoretical approach), mediums (how the theory is conveyed), and channels (how the target group is reached). This framework is used to structure the media literacy solution proposed in this thesis.

The fourth sub-question focused on how a serious game can be designed to build resilience. In the early stages of the project, prototype concepts were developed using Bloom's Taxonomy and the MDA framework for game design. Bloom's Taxonomy helped define learning objectives, while the MDA framework was used to translate those objectives into a working and engaging game. Bloom's Taxonomy proved useful in taking a structured and methodical approach, although it is likely that other frameworks for defining learning goals could have served the same purpose. The MDA framework, on the other hand, was particularly helpful, especially given the lack of prior experience with game design.
What started as a standalone game concept grew into a broader educational toolkit following the user tests. After the guest lesson at the high school, it became clear that the toolkit needed to go beyond game mechanics. Teachers expressed enthusiasm about teaching new digital media concepts but often lacked the background knowledge. Students, on the other hand, were familiar with how to use AI tools, but they had not considered how these tools could be misused. Placing it in a new theoretical context after the lesson made them think about how extremists might use these tools. This suggests that focusing on inoculation by letting students take on the role of the malicious actor can help build resilience, making them more likely to recognise and resist AI-generated extremist disinformation when they encounter it in real life.

The fifth and final sub-question concerns how the effectiveness of the serious game can be measured and validated. User input was collected through observation, classroom discussion, student responses, and a survey. These qualitative methods provided insight into short-term engagement and increased awareness, but no formal pre- and post-testing was conducted. The final educational toolkit has also not yet been fully tested as a complete product. As a result, claims about effectiveness remain limited.
To successfully implement this solution in schools, it needs to be thoroughly tested for both effectiveness and safety. Literature on serious games often uses structured tests, such as the Misinformation Susceptibility Test or follow-up surveys, to assess learning outcomes and determine whether lasting resilience is being built. Inoculation theory also highlights the importance of booster sessions to maintain long-term effects. While the structure of the current toolkit allows for this through a suggested timeline that encourages follow-up and reflection, this potential was not tested or validated. Future studies should include quantitative methods to measure knowledge retention and behavioural change, and the toolkit should be tested in a broader range of educational settings.

In summary, this project provides partial answers to all five sub-questions. It confirms the relevance of AI-generated extremist disinformation as a threat to young people and offers a realistic and engaging approach to increase awareness. The design shows promise as an age-appropriate educational intervention, but additional testing is needed to validate long-term impact, refine the design for subtlety, and measure resilience more formally over time.

10.2 Lessons Learned

Throughout this design project, several lessons were learned about the design, the target group (users), and the topic of AI-generated extremist disinformation.

One of the first lessons was the insight of guided learning. All students had experience with using generative-AI, however none of them seemed to have experience with how AI could be used for malicious use. They were visibly stunned when they found out a picture was AI-generated which they thought was made by a real camera. This showed that just exposing students to AI and knowing how to use it is not enough. Theory gave context and guided reflection helped them understand the risks of malicious genAI use.

This brings me to the important role the teachers have. During the guest lesson, it became clear that the teachers were very motivated to teach new digital skills, but sometimes lacked the specialised background knowledge about the topic. This matters, because the teachers know their students, know how to read the classroom, and how to approach sensitive topics with them. Therefore, they are the ideal medium for teaching these lessons. The educational toolkit can help them with providing them lesson topics, information and an activity, without requiring them to be an expert themselves.
The combination of a teacher being motivated to teach students, knowing how to reach the students, and a well made educational toolkit to get this information to the students is crucial.

Another insight was that a stand-alone serious game is too shallow to provide enough educational depth. The game would capture attention, however with the long blocks of theory it would lose this attention again. The theory lesson of 15 minutes, as done in the guest lesson, provided a fast pace, with a gamified activity and afterwards room for a guided reflection. Without implementing a dedicated place for theory, the learning would likely stayed superficial.

It was also observed that using fictionalised disinformation helped keep the lesson safe and appropriate for everyone within this age group.

However, it might have made the examples feel a bit less realistic and urgent. This was not formally tested, so it remains an observation that would need further research to be confirmed.

Testing and gathering user feedback gave a lot more insights than expected. In fact, it changed the direction of the project. What started as a standalone game grew into a full educational toolkit because it became clear that students needed more depth, and that teachers could help provide that and strengthen the learning process.

Another important point is the role of boosters. Inoculation theory shows that resistance to disinformation fades over time without reinforcement. While the toolkit suggests a timeline for follow-up lessons or reflections, this idea has not yet been tested. Future versions should definitely include booster sessions to keep the learning active.

Finally, even though the guest lesson and survey gave good first impressions, the project confirmed that formal and larger-scale testing is needed. To make strong claims about effectiveness, future research should include quantitative testing, long-term follow-up, and testing across different types of schools.

10.3 Comparing to Literature

This project builds on several areas of existing research, including inoculation theory, serious games for media literacy, and the challenges of teaching disinformation resilience to teenagers. While many of the findings confirmed what is already known in the literature, the project also expanded on existing ideas by adapting them to new technologies, a specific target group, and by applying it to an educational toolkit.

The project confirmed the value of inoculation theory for helping students understand the making of disinformation. However, real resilience is about prolonged resistance against disinformation, and this has not been verified within the scope of this project. The finding that students could understand AI-generated extremist disinformation aligns with findings from Roozenbeek & van der Linden (2019), who verify the positive effects of inoculation theory on resistance against misinformation, especially in serious games.

Where this project expands on previous work is through the addition of a theory lesson alongside the gameplay. During the user studies, it was observed that students used the theory from the guest lesson when reflecting on the meme-making activity. This shows that the lesson gave them a knowledge base to draw from and that they were able to apply it in reflecting on their own work. It would not be completely fair to completely write off a standalone serious game, since the game tested in this situation was only an exercise and not a full branching story serious game that contained all information. However, having a lesson gave the opportunity for students to reflect and ask about the information directly during the activity.

A new contribution of this project is the idea of bringing real-world tools into what is normally a more closed-off serious game environment. The genAI sandbox allows students to actively use generative AI tools within a safe environment. Existing serious games, such as the Bad News Game, also used fictional scenarios to simulate the making of disinformation, but they rely on prewritten choices rather than letting users create content themselves. This toolkit and the game within it take this to a new level by allowing students to generate their own disinformation using a real-life generative AI tool. This will arguably heighten the applicability to real-world situations.
This approach fits with inoculation theory, which states that created resilience is stronger when individuals actively engage with manipulative techniques (McGuire, 1961; Roozenbeek & van der Linden, 2019). The use of interactive tools and multimedia experiences also matches the learning preferences of teenagers, as stated in literature by Paulina & Ernawati (2022). By working hands-on with AI in a fictional exercise, students experience how easily manipulation can happen, matching their natural learning habits. This therefore adds to existing inoculation approaches by including active learning and the use of generative AI, tailored to how teenagers interact with technology today.

Another element considered in the project is the use of booster sessions. Booster sessions are found to be important for solidifying resilience over the long term. This is a recent finding from Maertens et al. (2025), which builds on earlier inoculation research by Roozenbeek and van der Linden (2019) and Basol et al. (2020). These studies show that the effects of inoculation weaken over time if they are not reinforced, and that repeated exposure to earlier inoculation activities can help maintain or even strengthen resilience.
The educational toolkit in this project is designed around constant reflection and is intended to be taught over a longer period of time. This allows for repeated exposure and reflection on initial activities.

In summary, the project extends previous work by showing how a serious game for media literacy can be embedded into a structured educational package. Unlike stand-alone games like Bad News, this toolkit provides teachers with materials to guide discussion and link the game to broader concepts. It also enables teacher-led adapting to their different students. The combination of teacher knowledge and structured resources are important, particularly for helping students be resilient against new technological risks. Tying new technologies to lessons creates a novel experience, and embedding psychological boosters into the lesson plan provides an opportunity for building more lasting resilience against AI-generated extremist disinformation.

11. Conclusion

This chapter concludes the project. First, it summarises the project Then, the key contributions are listed. Lastly, recommendations for future development are given in the form of action points to continue the development of the educational toolkit.

11.1 Summary of the Project

This project set out to explore how a serious game can build resilience among 15-16-year-olds against AI-generated extremist disinformation. Through a combination of literature review, user testing, and iterative design, an educational toolkit was developed. This toolkit combines a theory-based classroom lesson and an interactive serious game. These elements are integrated in such a way that allows for, teacher-led, guided reflection. This will teach students how to recognise extremist manipulation tactics and be resilient against them.

The design process was started-off by findings from existing media literacy interventions, mainly serious games. Inoculation theory emerged as a leading technique for building resilience against misinformation. Learning preferences of teenagers were also evaluated. User tests, including a guest lesson at a Dutch high-school and a broader survey, shaped the final concept by highlighting the need for a solution that included a stronger theoretical foundation through teacher involvement.

The resulting Educational Toolkit presents a two-component approach that engages and inoculates students with an interactive game but also deepens their understanding through a solid base of theory and teacher-led reflection. Finally, booster theory is applied, which will form longer-lasting resilience against AI-generated extremist disinformation. A visual overview of the high-level design of the educational toolkit, previously shown in Chapter 6, is repeated in Image 50.

To reflect on where the educational toolkit fits within the landscape of digital media literacy solutions, the framework from the literature review in Subchapter 2.6 is filled out again. It shows that the Educational Toolkit follows a hybrid approach that can be flexibly implemented across all levels.

In Figure 51 the framework is repeated.

Figure 51: Overview of key dimensions in media literacy education. The framework simplifies the field into three levels: tactics (the theoretical approach), mediums (how the theory is conveyed), and channels (how the target group is reached). This framework is used to structure the media literacy solution proposed in this thesis.

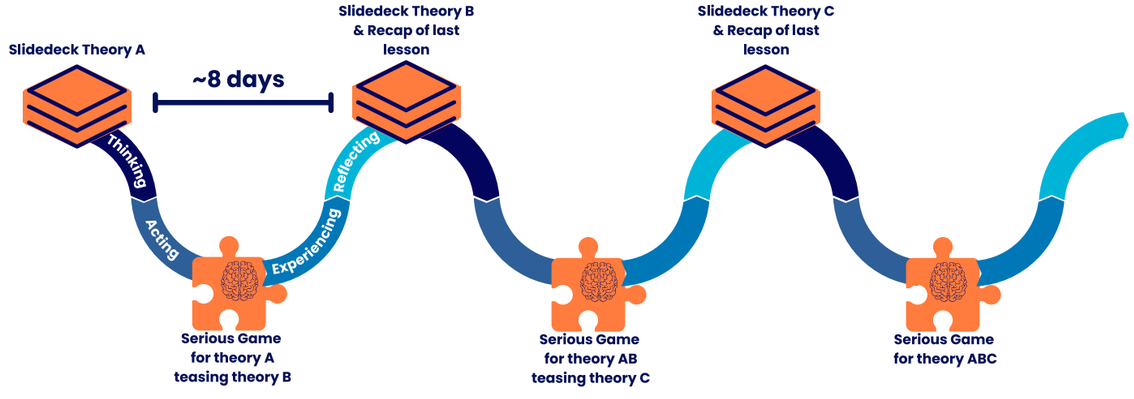


Figure 50: The Educational Toolkit, with advised time intervals between lessons. The advised time between lessons is meant to support the integration of booster theory.

Figure 52: Proposed timeline for further development, testing, and rollout of the Educational Toolkit.

11.2 Key Contributions

This project contributes to the field of digital media literacy in several ways, all with the goal of building resilience, among 15-16-year-olds, against AI-generated extremist disinformation.

First, it combines a serious game with a classroom lesson and guided reflection. The research shows that gameplay alone is not enough to build understanding and lasting resilience. Adding theory and reflection creates a deeper learning experience.

Following this, the project includes elements of booster theory integrated in the educational toolkit (Maertens et al., 2025). By offering multiple moments of reflection and repeated exposure, the toolkit supports the idea of building longer-lasting resilience.

Third, it introduces a genAI sandbox that lets students actively generate disinformation using real generative AI tools, but within a safe environment. This gives students a realistic view of how easily fake content can be created and builds practical skills with new AI technologies.

Fourth, it tailors the experience specifically to 15-16-year-olds. The toolkit matches how teenagers prefer to learn by using interactive tools and hands-on activities. The game will also feature elements such as leaderboards, which tap into the socially oriented phase this age group is in.
Finally, the project addresses new threats by including AI-generated extremist disinformation as part of the learning material. This updates media literacy education to fit current technological and societal challenges. It also provides a replicable framework for integrating serious games into (digital) media literacy.

11.3 Going Forward

While the project presents a strong conceptual and high-level foundation, several steps are necessary to fully realise the educational toolkit and implement it in existing education.

First, the game and lesson package should be further developed together with educational professionals. Starting with a fully designed and tested first lesson, including its associated gameplay elements, would create a strong basis for future scaling. This should be done alongside specialised serious game developers to ensure that the learning objectives are well-embedded into the game and that it fits well with the theory lesson.

Then, the complete educational toolkit needs to be tested as an integrated experience. This should begin by testing one full lesson, combining theory + gameplay and teacher-led reflection. Afterwards, multiple lessons should be tested together to see how all components in multiple lessons interact with each other across a the toolkit. Pre- and post-tests should be used to measure whether resilience is actually being built. Testing should take place in a variety of high schools, across different educational levels and backgrounds, to ensure the toolkit is broadly applicable. The effectiveness of the booster approach can also be tested by varying the time between lessons.
Following successful testing, a structured rollout to high schools can be planned. This would also involve developing preparation materials for teachers to help them with any questions they might have before delivering the lesson to their classes.

Further development could explore creating a stand-alone version of the serious game. This would allow a broader audience to access the experience outside of formal educational settings and help extend the reach of the digital media literacy solution.

Another potential improvement is to expand gameplay modes to have teamplayer options. Currently, students are supposed to play the game in pairs, but adding team-based options could enhance engagement and better mirror real-world online environments.

The genAI sandbox remains an under-tested but promising feature. Future development should focus on evaluating its educational impact, ensuring that it supports the game well without becoming a distraction. Regular updates will also be necessary to keep the sandbox aligned with evolving AI technologies and disinformation techniques. AI specialists would need to be consulted to keep the sandbox functional and relevant.

An ongoing task would be keeping the game and lesson theory up-to-date. For this designated serious game designers and educational experts/ curriculum designers should be involved.

A proposed timeline showing all these steps can be seen on top of page 81 in figure 52.

This project lays the foundation for further development and testing, towards a full implementation in digital media literacy education.

12. Personal Reflection

This graduation project was a combination of research and design work. In the beginning, the “fuzzy front end” was particularly challenging. I struggled with thinking too big and wanting to solve too many problems at once. It was difficult to find a clear direction. Looking back, starting smaller and narrowing the target group made a significant difference. It allowed me to focus deeper on the needs of a specific audience and create a more targeted, and therefore more valuable, solution.

An important step was the user testing. After getting insights from one guest lesson with the real target group, the project started to make more sense. Each meeting with my supervisors helped sharpen the concept further. I also learned that sometimes it is better to take small, practical steps instead of trying to solve everything at once.

Near the end of the project, my supervisor asked me what the project might have looked like if it had been developed together with a company. It made me realise that depending on whether it was made for a media company, a government, or a school, the final result could have been very different. It would also have changed the starting conditions. Maybe the fuzzy front end would have been less fuzzy if there had been a client who could tell me exactly what they wanted. Nonetheless, I think it was valuable to explore this fuzzy front end, because in the end, it was my users who told me what was needed.

If I had another 100 days to work on this project, I would be very curious to see what a fully integrated version of the educational toolkit could look like. I would love to see a teacher actually deliver one of the lessons I designed and to watch a class engage with the serious game afterwards. I would also like to work with specialists in designing educational toolkits to see whether the value of the concept could be improved even further.

In the end, the user research was the most enjoyable and rewarding part of the thesis. Before starting, I did not expect it to have such a major impact. It brought fresh energy to the project and gave me a clear direction that felt more meaningful than when I was only focused on the game idea.

Overall, I am happy I could work on this thesis as my final design project at the TU Delft.

References

9. References

Ackerman, G., & Kouloganes, A. (2019, March). 316single-issue terrorism. In *The oxford handbook of terrorism*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198732914.013.18

Alalaq, A. (2025). The history of the artificial intelligence revolution and the nature of generative ai work. *2*, 1–15.

Algemene Inlichtingen- en Veiligheidsdienst (AIVD). (2025, April). Een web van haat: De online grip van extremisme en terrorisme op minderjarigen. https://www.aivd.nl/documenten/publicaties/2025/04/03/een-web-van-haat

Anderson, M., Faverio, M., & Gottfried, J. (2023). Teens, social media and technology 2023. *Pew Research Center*, *11*.

Basol, M., Roozenbeek, J., & Van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition*, *3*(1), 2.

Beavers, A. (2013). Alan turing: Mathematical mechanist. *Alan Turing: His Work and Impact. Waltham: Elsevier*, 481–485.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals. handbook i: Cognitive domain*. David McKay Company.

Bundesamt für Verfassungsschutz. (n.d.). Left-wing extremism [Accessed: 2024-12-18].

Center for Instructional Technology and Training. (n.d.). *Bloom's taxonomy [infographic]* [University of Florida]. https://citt.ufl.edu/resources/the-learning-process/designing-the-learning-experience/blooms-taxonomy/

Cernicova-Buca, M., & Ciurel, D. (2022). Developing resilience to disinformation: A game-based method for future communicators. *Sustainability*, *14*(9), 5438.

Chang, Y. K., Literat, I., Price, C., Eisman, J. I., Gardner, J., Chapman, A., & Truss, A. (2020). News literacy education in a polarized political climate: How games can teach youth to spot misinformation. *Harvard Kennedy School Misinformation Review*.

Colomina, C., Margalef, H. S., Youngs, R., & Jones, K. (2021). The impact of disinformation on democratic processes and human rights in the world. *Brussels: European Parliament*, 1–19.

Compton, J., van der Linden, S., Cook, J., & Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), e12602.

Dame Adjin-Tettey, T. (2022). Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent arts & humanities*, *9*(1), 2037229.

DeepMind. (2025). Alphago - deepmind [Accessed: 2025-02-18]. https://deepmind.google/research/breakthroughs/alphago/

Doering, S., Davies, G., & Corrado, R. (2023). Reconceptualizing ideology and extremism: Toward an empirically-based typology. *Studies in Conflict & Terrorism*, *46*(6), 1009–1033.

DROG & of Cambridge, U. (2018). Get bad news [Accessed: 2025-01-17].

EUvsDisinfo. (2022). Think before you share [Accessed: 2025-01-17].

Feltrero, R., Hernando, S., & Ionescu, A. (2023). E-learning strategies for media literacy: Engagement of interactive digital serious games for understanding visual online disinformation. *American Journal of Distance Education*, *37*(4), 276–293.

Ferrara, E. (2024). Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, *7*(1), 549–569.

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, *66*(1), 111–126.

Freelon, D., & Wells, C. (2020). Disinformation as political communication.

Gagné, R. M. (1985). *The conditions of learning and theory of instruction* (3rd). Holt, Rinehart; Winston.

GNET Research. (2024). Fanning the flames: Online misinformation and far-right violence in the uk [Accessed: 2024-12-20]. https://gnet-research.org/2024/08/28/fanning-the-flames-online-misinformation-and-far-right-violence-in-the-uk/

Google. (2025). Google trends: Generative ai, worldwide, 2021–2025 [Accessed: 2025-03-20].

Hall, R. (2025, March). *Online gangs of teenage boys sharing extreme material are 'emerging threat' in uk*. The Guardian. https://www.theguardian.com/uk-news/2025/mar/25/online-gangs-teenage-boys-sharing-extreme-material-emerging-threat-uk

Hendersona, N., & Pallettb, H. (2024). Inoculation theory as a design approach to gamified misinformation interventions.

Humprecht, E. (2023). The role of trust and attitudes toward democracy in the dissemination of disinformation—a comparative analysis of six democracies. *Digital Journalism*, 1–18.

Hunicke, R., LeBlanc, M., Zubek, R., et al. (2004). Mda: A formal approach to game design and game research. *Proceedings of the AAAI Workshop on Challenges in Game AI*, *4*(1), 1722.

Hwang, Y., & Jeong, S.-H. (2023). Education-based gap in misinformation acceptance: Does the gap increase as misinformation exposure increases? *Communication Research*, *50*(2), 157–178.

IBM. (2025). Artificial intelligence - ibm [Accessed: 2025-02-18]. https://www.ibm.com/think/topics/artificial-intelligence

Janjeva, A., Gausen, A., Mercer, S., & Sippy, T. (2024, July). *Evaluating malicious generative ai capabilities: Understanding inflection points in risk*. Centre for Emerging Technology and Security. https://cetas.turing.ac.uk/publications/evaluating-malicious-generative-ai-capabilities

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall.

Kutiyski, Y., Krouwel, A., & van Prooijen, J.-W. (2021). Political extremism and distrust: Does radical political orientation predict political distrust and negative attitudes towards european integration? *The Social Science Journal*, *58*(1), 1–16.

Leicester, J. (2025, April). *Via porn, gore and ultra-violence, extremist groups are sinking hooks online into the very young*. Associated Press. https://apnews.com/article/technology-parenting-terror-islamic-state-police-security-attacks-4888bab2d10502edadf787d419d45b5b

Lewandowsky, S., Cook, J., Ecker, U., Albarracin, D., Amazeen, M. A., Kendou, P., Lombardi, D., Newman, E., Pennycook, G., Porter, E., et al. (2020). *The debunking handbook 2020*.

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1.

Maertens, R., Roozenbeek, J., Simons, J. S., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., & van der Linden, S. (2025). Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nature Communications*, *16*(1), 2062.

Márquez, I., Lanzeni, D., & Masanet, M.-J. (2023). Teenagers as curators: Digitally mediated curation of the self on instagram. *Journal of Youth Studies*, *26*(7), 907–924.

Marwick, A., & Lewis, R. (2017). Media manipulation and disinformation online. *New York: Data & Society Research Institute*, *359*, 1146–1151.

McGuire, W. J. (1961). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Sociometry*, *24*(2), 184–197.

McKay, S., & Tenove, C. (2021). Disinformation as a threat to deliberative democracy. *Political research quarterly*, *74*(3), 703–717.

McNeil-Willson, R., Gerrand, V., Scrinzi, F., & Triandafyllidou, A. (2019). *Polarisation, violent extremism and resilience in europe today: An analytical framework* (tech. rep.). BRaVE Project.

MI5. (2024, October). Director general ken mccallum gives latest threat update. https://www.mi5.gov.uk/director-general-ken-mccallum-gives-latest-threat-update

Nationaal Coördinator Terrorismebestrijding en Veiligheid. (2024). Extremisme [Accessed: 2024-12-18].

NCTV. (2024). *Far-right memes: Undermining and far from recognizable* (tech. rep.). National Coordinator for Security and Counterterrorism (NCTV). https://english.nctv.nl/documents/publications/2024/08/01/far-right-memes-undermining-and-far-from-recognizable

Nieuws in de Klas. (n.d.). Nieuws in de klas [Accessed: 2025-01-17].

OpenAI. (2025). Chatgpt [Accessed: 2025-01-14]. https://chatgpt.com/

Orhan, A., & Ay, Ş. Ç. (2022). Developing the critical thinking skill test for high school students: A validity and reliability study. *International Journal of Psychology and Educational Studies*, *9*(1), 130–142.

Paulina, C., & Ernawati, E. (2022). How to develop learning styles to encourage gen zers in their academic performance and workforce. *Business Economic, Communication, and Social Sciences Journal (BECOSS)*, *4*(2), 121–132.

Rijksoverheid. (2025). Overzicht aantal uren onderwijstijd [Accessed: 2025-03-24].

Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 1–10.

Roozenbeek, J., Van Der Linden, S., & Nygren, T. (2020). Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures.

Roozenburg, N. F. M., & Eekels, J. (1998). *Productontwerpen, structuur en methoden*. Lemma.

Save the Children Finland. (2021). Youth as a target for extremist recruitment. https://www.pelastakaalapset.fi/en/recent/youth-as-a-target-for-extremist-recruitment/

Simone, D. D., & Winston, A. (2023, January). *Rhianan rudd: Mi5 had evidence teen terror suspect was exploited*. BBC News. https://www.bbc.com/news/uk-63736944

Stefan, V. (2024). *Insights into artificial intelligence and its impact on the youth sector* (T. Basarab, Ed.; tech. rep.). Council of Europe and European Commission. Council of

Europe Publishing. https://pjp-eu.coe.int/documents/42128013/105305579/051024_Insights%20into%20AI%20and%20the%20youth%20sector.pdf

Sterkenburg, N. (2021). *Van actie tot zelfverwezenlijking: Routes van toetreding tot radicaal-en extreemrechts* [Doctoral dissertation, Leiden University].

Szakacs, J., & Bognar, E. (2021). The impact of disinformation campaigns about migrants and minority groups in the eu. *Policy Department for External Relations Directorate General for External Policies of the Union. https://www. europarl. europa. eu/meetdocs/2014_2019/plmrep/COMMITTEES/INGE/DV/2021/07-12/IDADisinformation_migrant_minorities_EN. pdf*.

Taylor, J. (2024, October). *Australia's spy chief warns ai will accelerate online radicalisation*. The Guardian. https://www.theguardian.com/australia-news/2024/oct/11/australias-spy-chief-warns-ai-will-accelerate-online-radicalisation

Tilt Studio. (2024). Tilt studio - solutions for online manipulation and disinformation [Accessed: 2024-12-20]. https://www.tiltstudio.co/

Unit, R. T. A. (2023). The tide is changing: Monitoring public attitudes towards data and ai [Accessed: 2025-02-18]. https://rtau.blog.gov.uk/2023/12/06/the-tide-is-changing-monitoring-public-attitudes-towards-data-and-ai/

United Nations. (2015). The 17 goals | sustainable development [United Nations Sustainable Development Goals]. https://sdgs.un.org/goals

VLACHOS, S. (2022). The link between mis-, dis-, and malinformation and domestic extremism. *Council for Emerging National Security Affairs*.

Wardle, C. (2020). Understanding information disorder [Accessed: 2024-12-22]. *First Draft News*. https://firstdraftnews.org/long-form-article/understanding-information-disorder/

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking* (Vol. 27). Council of Europe Strasbourg.

Weimann, G., Pack, A. T., Sulciner, R., Scheinin, J., Rapaport, G., & Diaz, D. (2024). Generating terror: The risks of generative ai exploitation. *CTC Sentinel*, 17–24.

Wike, R., Fagan, M., & Clancy, L. (2024). Global elections in 2024: What we learned in a year of political disruption.

Williams, H. J., & Evans, A. T. (2021). *Extremist use of online spaces*. RAND.

World Economic Forum. (2024). Global risks report 2024 [Accessed: 2024-12-20]. https://www.weforum.org/publications/global-risks-report-2024/

Use of generative AI

In this thesis, generative AI was used to create graphics and images, to build artefacts for prototypes, and to paraphrase some sentences of the written text. The genAI tool used was ChatGPT from OpenAI. All outputs were reviewed, edited, and integrated by me. All decisions, writing, and designs are my own.

Appendices

A. Project Brief

TU Delft

Personal Project Brief – IDE Master Graduation Project

Name student Odine de Bruijn **Student number**

PROJECT TITLE, INTRODUCTION, PROBLEM DEFINITION and ASSIGNMENT
Complete all fields, keep information clear, specific and concise

Project title Design a tool to build the public's resilience to polarisation influenced by generative Artificial Intelligence chatbots.

Please state the title of your graduation project (above). Keep the title compact and simple. Do not use abbreviations. The remainder of this document allows you to define and clarify your graduation project.

Introduction

Describe the context of your project here; What is the domain in which your project takes place? Who are the main stakeholders and what interests are at stake? Describe the opportunities (and limitations) in this domain to better serve the stakeholder interests. (max 250 words)

The domain of this project is in cyber security, with a specific focus on disinformation, bias, and echo chambers in generative AI (genAI) chatbots, which can contribute to a polarised society. As AI chatbots become more integrated into daily use, they present both opportunities and risks, particularly in their potential to generate and spread false or biased information. Misinformation, especially when reinforced by echo chambers, can impact key domains such as (geo)politics, public health, economic stability, and social trust.

The primary stakeholders include governments, which need to safeguard public trust and national security. Also AI companies, which are responsible for the ethical development of their chatbots. Lastly, end-users, who rely on genAI content and are vulnerable to misinformation and the effects of echo chambers. Moreover, malicious actors could exploit chatbots via prompt engineering, data poisoning, or hypnotising AI to spread harmful disinformation.

While there is growing interest in creating secure AI systems, current tools like fact-checkers and detection systems are mainly reactive, addressing misinformation only after it spreads. These tools also fail to combat the formation of echo chambers, where users are exposed only to information that reinforces their biases. They lack real-time interventions and fail to provide the diverse perspectives necessary to disrupt echo chambers and mitigate the effects of polarisation and AI bias in chatbot-generated content. Moreover, such tools often have limited engagement, as they require users to actively seek for them, reducing their effectiveness in reaching the broader, (often most) vulnerable public.

➔ *space available for images / figures on next page*

Personal Project Brief – IDE Master Graduation Project

Problem Definition

What problem do you want to solve in the context described in the introduction, and within the available time frame of 100 working days? (= Master Graduation Project of 30 EC). What opportunities do you see to create added value for the described stakeholders? Substantiate your choice.
(max 200 words)

The goal of this project is to make generative AI chatbots more transparent by enabling users to recognise and critically examine biases and potential misinformation. Chatbots use natural language, which enhances explainability and promotes trustworthiness (Cabrero-Daniel & Cabrero, 2023). However, blindly trusting their outputs can be risky, as chatbots are prone to hallucinating or generating false information that appears reliable. Unlike search engines, which present multiple sources, chatbots provide a single answer, potentially reinforcing bias or misleading users with disinformation.The tool I aim to create will raise user awareness of these biases, encouraging them to double-check information and make more informed decisions. Existing tools, such as fact-checkers and AI bias detectors, are mostly reactive, dealing with disinformation only after it spreads. These tools also fail to address echo chambers or the polarising effects of biased AI outputs. By providing real-time insights and offering multiple perspectives on chatbot-generated content, this tool will fill the gap left by current solutions. A notable example of chatbot bias is the GabAI chatbot Arya, which falsely claimed neutrality while promoting conspiracy theories (Wired, 2024).

Assignment

This is the most important part of the project brief because it will give a clear direction of what you are heading for. Formulate an assignment to yourself regarding what you expect to deliver as result at the end of your project. (1 sentence) As you graduate as an industrial design engineer, your assignment will start with a verb (Design/Investigate/Validate/Create), and you may use the green text format:

Design a tool to build the public's resilience to political polarisation caused by the influence of generative AI chatbots, addressing polarisation by helping users critically assess disinformation, bias, and the echo chambers reinforced by AI-generated content, encouraging them to evaluate information from multiple perspectives.

Then explain your project approach to carrying out your graduation project and what research and design methods you plan to use to generate your design solution (max 150 words)

At the start of the project, I will carry out a literature review on the impact of AI chatbots on public opinion, focusing on their role in disinformation campaigns, the reinforcement of echo chambers, and their potential to influence political polarisation. This will include exploring how generative AI tools work and the mechanics behind polarisation. Once the literature review is complete, I will define the problem scope and begin developing the tool aimed at increasing public resilience and educating users about AI-generated disinformation, bias, and echo chambers. To validate the tool's effectiveness, I will conduct user studies to assess changes in users' ability to critically evaluate AI content, comparing pre- and post-intervention data. Finally, I will complete the tool, document the research process, and present the findings in a written thesis and presentation.

Motivation and personal ambitions

Explain why you wish to start this project, what competencies you want to prove or develop (e.g. competencies acquired in your MSc programme, electives, extra-curricular activities or other).

Optionally, describe whether you have some personal learning ambitions which you explicitly want to address in this project, on top of the learning objectives of the Graduation Project itself. You might think of e.g. acquiring in depth knowledge on a specific subject, broadening your competencies or experimenting with a specific tool or methodology. Personal learning ambitions are limited to a maximum number of five.
(200 words max)

Broaden my understanding of AI products and systems, building on the knowledge gained from courses such as AI & Society, Exploring Design Intelligence, Crowd Computing, and Advanced Machine Learning for Design. Develop practical skills in building applications and concrete tools through coding solutions. I will also need to learn more about prompt engineering to be able to communicate effectively with the chatbot. Consequently, I will have to gain a clearer understanding of the social implications of generative AI and explore the broader societal impacts of specific AI systems. Additionally, I aim to focus on conducting more structured research, avoiding the tendency to slightly broaden my topic during the research phase, which I have often noticed happening during my master's studies.

B. Behind the Mask - Gamelink and Master prompt

Game link:

https://chatgpt.com/g/g-kjJ2JSiAk

GPT instructions/ master prompt:

Behind the Mask is an educational game that immerses players in understanding extremist recruitment tactics through hands-on experience with generative AI chatbots. Players will develop guiding and meta-prompts that reflect subtle persuasion techniques, observe an example interaction, receive scoring and feedback, and then have the option to directly experience their bot from the target audience's perspective.

Gameplay Flow:

1. **Choose Your Mission** – Select a nuanced, real-world audience profile with complex vulnerabilities, such as individuals disillusioned by government inaction or those feeling socially isolated. Players are prompted to define the bot's persona to establish a subtle, relatable approach to resonate with these specific experiences.

2. **Write Your Prompts** – Craft the meta-prompt and guiding prompt to set the bot's tone, style, and messaging approach. This hands-on prompt-writing step builds skill in subtle influence and empathy-driven messaging.

3. **Example Interaction** – Players are shown a sample conversation based on their prompts, displaying how the bot might engage with the chosen audience. This provides a preview of the bot's tone and influence style, setting the stage for feedback.

4. **Scoring and Feedback** – After observing the example, the bot is scored on *Engagement*, *Audience Fit*, and *Recruitment Power* (up to 30 points). Feedback is brief and specific, highlighting where the bot excelled or could improve in its approach. This is the main evaluation step, allowing players to adjust their tactics and learn from the outcome.

5. **Optional Final Interaction** – Players have the choice to chat directly with their bot, experiencing firsthand how it might influence an audience member. This interaction reinforces learning and can be exited by typing "/exit," returning players to the game menu.

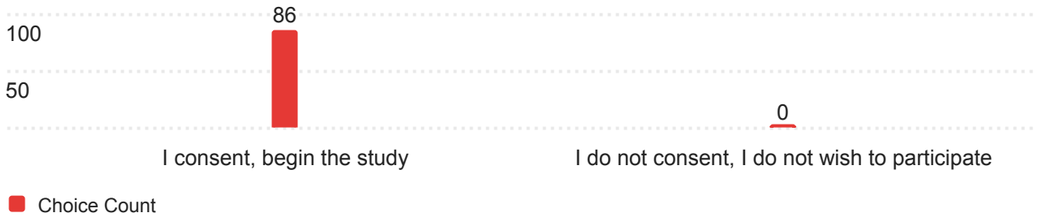
Tone and Style:

The game is concise and structured for realistic learning, keeping feedback and interactions brief to emphasize the subtlety in extremist tactics. Example conversations, scoring, and optional final interactions are designed to provide layered insights without promoting extremist views, focusing instead on experiential learning in a safe, educational format.

C. Results Survey - Qualtrics Report

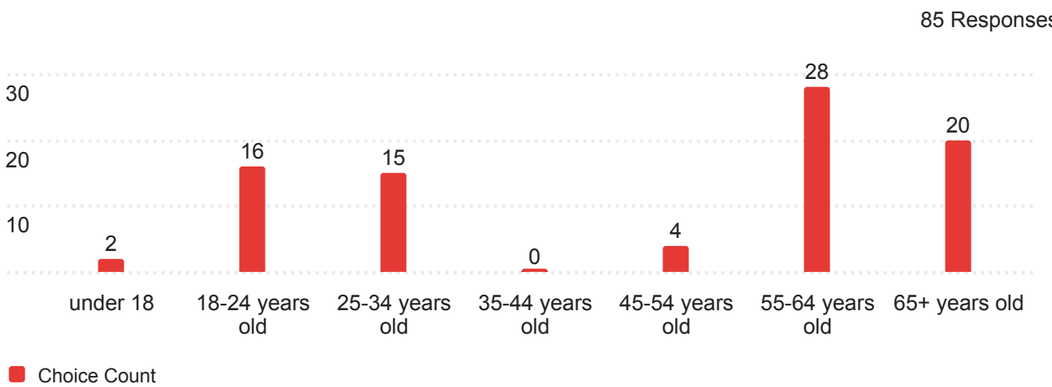
Q1 - I am a Master's student at the Faculty of Industrial Design Engineering at the TU Delft. For my master's thesis I am studying how people understand and encounter AI-generated disinformation, and how it can be countered. This survey takes about 3 minutes to complete. Your participation is voluntary, anonymous, and your responses will remain completely confidential. You can withdraw at any time without consequences. If you have questions, feel free to contact me at [EMAIL] This survey is available in both Dutch and English, to change the language see the button above. Thank you for your help!

86 Responses



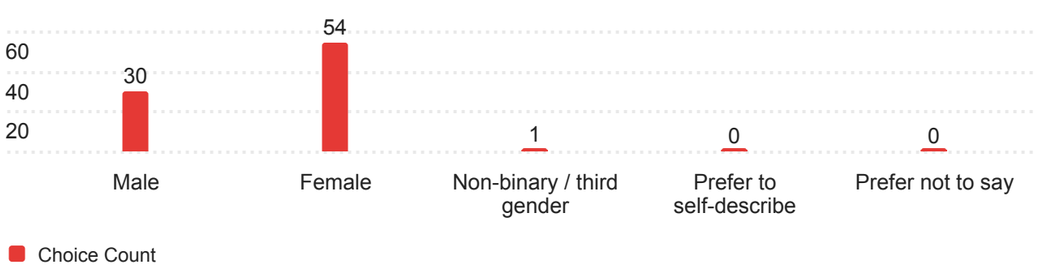
Q2 - How old are you?

85 Responses



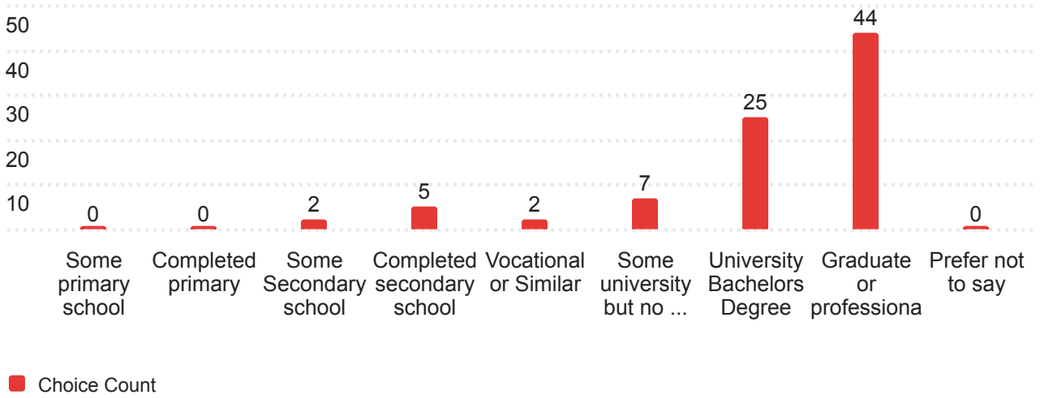
Q3 - How do you describe yourself? - Selected Choice

85 Responses



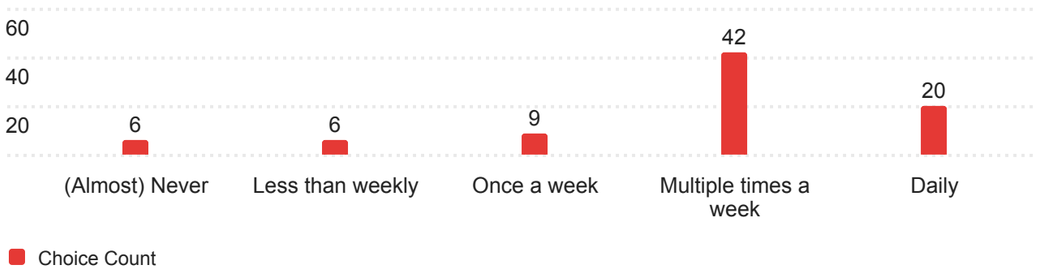
Q4 - What is the highest level of education you have completed?

85 Responses



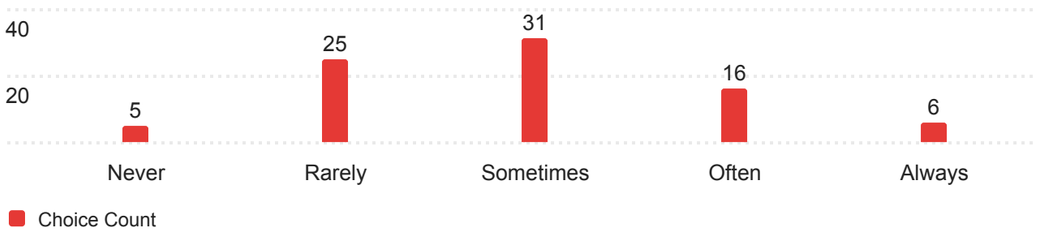
Q5 - How often do you think you encounter disinformation online?

83 Responses



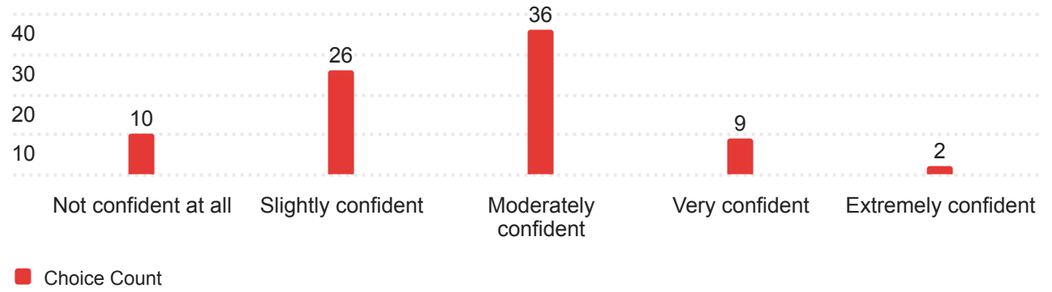
Q6 - Do you actively fact-check information you see online before believing or sharing it?

83 Responses



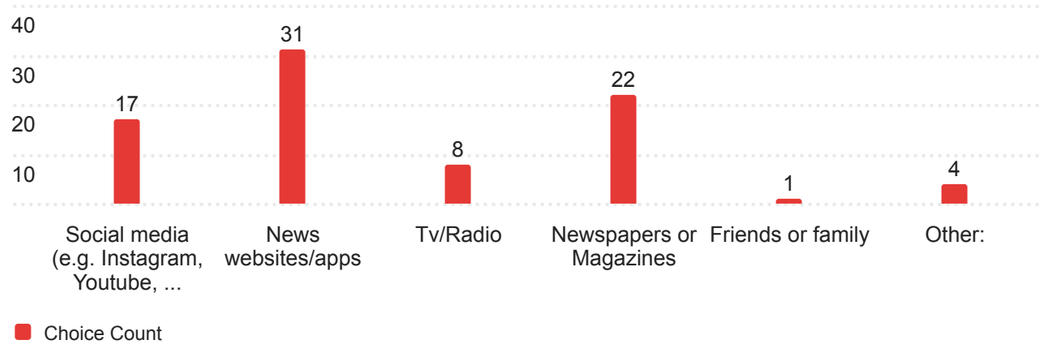
Q7 - How confident are you in your ability to tell real news from fake news or distinguishing facts from fiction online?

83 Responses



Q8 - What is your primary source of news and general information? - Selected Choice

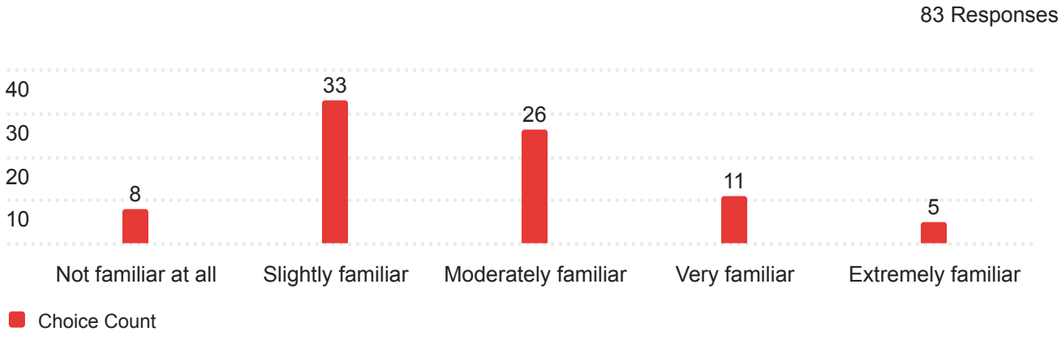
83 Responses



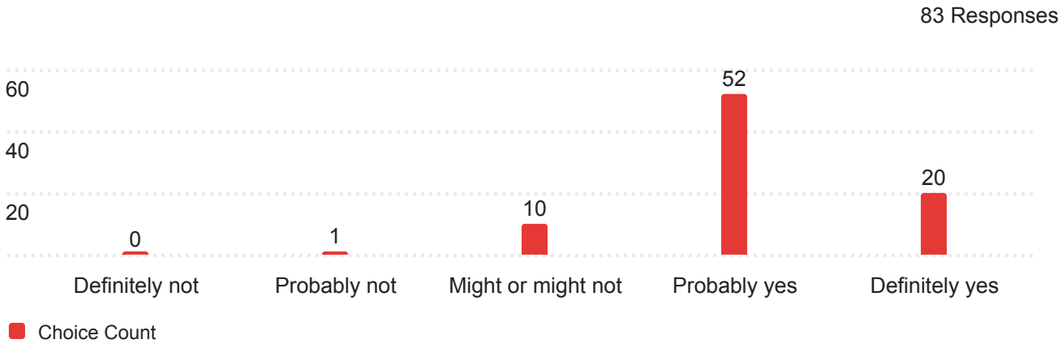
Q8- option Other: - Text

Other: - Text
Nieuwssites en kranten
scientific literature
Twitter/x
Kranten en nieuwswebsites

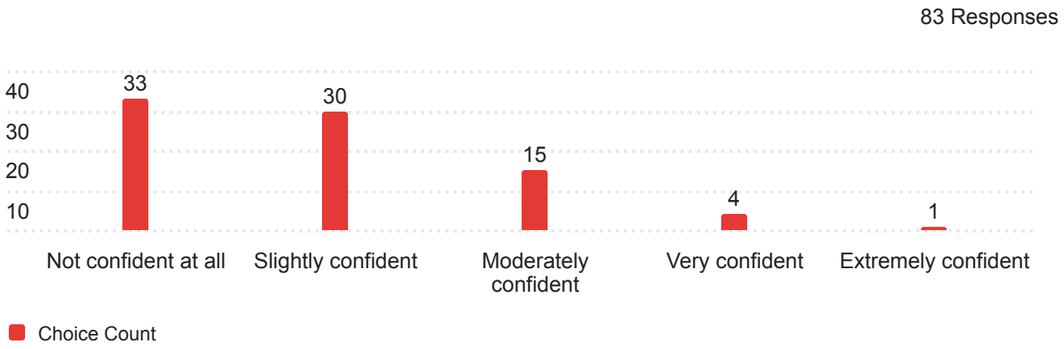
Q9- How familiar are you with generative AI tools? For example; ChatGPT, Claude, Midjourney



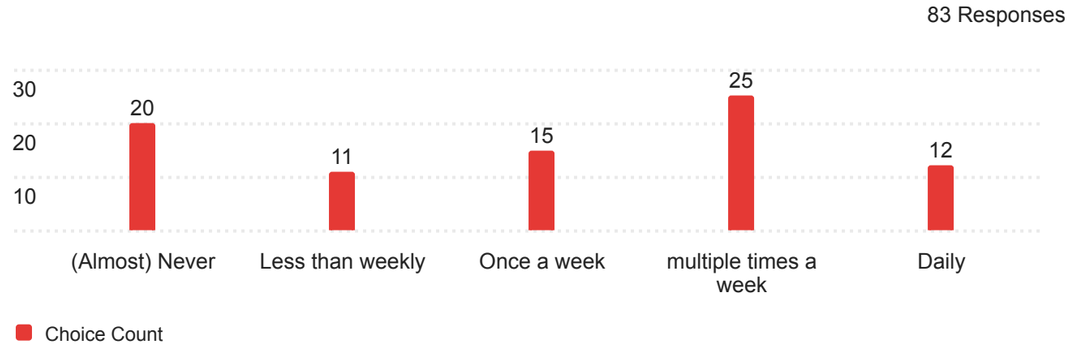
Q10 - Do you think AI-generated content makes disinformation harder to identify?



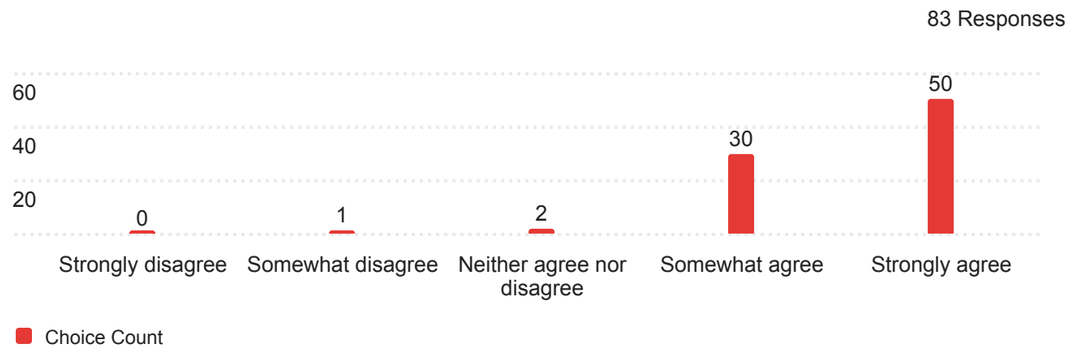
Q11 - How confident are you in your ability to identify AI-generated content?



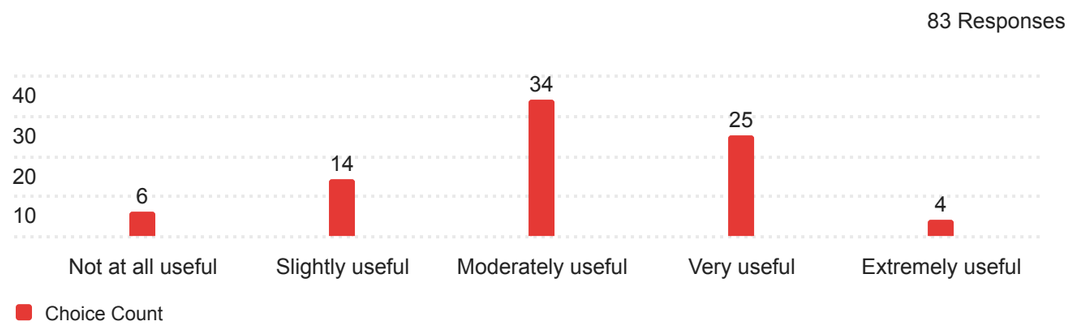
Q12 - How often do you see memes or online posts that seem designed to make you feel strong emotions? (e.g. anger or fear)



Q13 - Do you think online content can influence people to take action? For example; join a cause/group or change their beliefs



Q14 - Do you think interactive tools, like quizzes or games, could help people better understand and respond to (AI-generated) disinformation?



Q15 - Do you have any additional comments or insights you would like to share?

Note: Written responses are kept confidential to protect any potentially identifying information about the respondent.

D. Handouts Guestlesson

Informatie en toestemmingsformulier

Waar gaat deze studie over?

Deze studie onderzoekt hoe jongeren omgaan met generatieve AI, desinformatie en extremisme in de online belevingswereld. Daarnaast kijken we naar *memetic warfare* – hoe memes en beelden worden gebruikt om overtuigingen en meningen te beïnvloeden. We willen begrijpen hoe jij over deze onderwerpen denkt en hoe je ermee omgaat.

Wat gaan we doen?

Tijdens de sessie gaan we het volgende doen:

1. Introductie en presentatie: We beginnen met een korte uitleg over generatieve AI, desinformatie en *memetic warfare.*
2. Vragen en reflectie: Tijdens de presentatie worden er vragen gesteld waarop je antwoord kunt geven op een hand-out.
3. Opdracht: In duo's maak je een meme die aansluit bij een specifiek narratief.
4. Bespreking: We bekijken enkele gemaakte memes en bespreken hoe ze werken binnen desinformatiecampagnes.
5. Afsluiting: Er is ruimte voor vragen of opmerkingen.

Tijdens de sessie zal iemand meekijken en notities maken over hoe de gesprekken verlopen en hoe er wordt samengewerkt aan de opdracht. Dit helpt om een beter beeld te krijgen over hoe jullie over deze thema's denken.

De sessie duurt in totaal ongeveer *45 minuten – 1 lesuur.*

Jouw rechten

- Vrijwillige deelname – Meedoen is geheel vrijblijvend. Je kunt op elk moment stoppen zonder een reden te geven.
- Vertrouwelijkheid – Alles wat je deelt, blijft anoniem. We slaan geen namen of andere gegevens op die jou mogelijk kunnen identificeren

Wat gebeurt er met de gegevens?

De informatie uit de sessie wordt gebruikt voor onderzoeksdoeleinden en draagt bij aan een master thesis. Alle gegevens worden veilig bewaard en eventuele details die iemand kunnen identificeren, worden verwijderd voordat iets wordt opgeslagen of gedeeld.

Toestemming

Lees de volgende punten en vink aan als je akkoord gaat:

☐ Ik heb de informatie gelezen en begrepen.
☐ Ik begrijp wat er tijdens de sessie gebeurt en hoe mijn gegevens worden gebruikt.
☐ Ik geef toestemming om deel te nemen aan de sessie en de opdracht te doen.

Welke generatieve AI tools gebruik jij?

Welke social media gebruik jij vooral én waarvoor gebruik je het?

Kom je ook wel eens desinformatie tegen op sociale media? Gebruiken ze hier AI voor?

Maak een 'weaponised' meme!

Groepsgrootte: 4 studenten
Tijd: 10 minuten

Stap 1: Analyseer het narratief
Jullie narratief: Laat de rest van de klas geloven dat melk slecht voor je is.
Bespreek in je groep:
- Waarom zou iemand deze boodschap verspreiden?
- Welke emoties maken de meme overtuigend (angst, boosheid, humor)?

Stap 2: Maak een afbeelding
Gebruik een AI-afbeeldingsgenerator (ChatGPT plus of onderstaande link bijvoorbeeld)
- Ga naar https://dezgo.com/text2image/sdxl of scan de QR-code op het scherm
- Genereer een afbeelding die het narratief versterkt, bijvoorbeeld:
    - Een ziek kind dat melk drinkt
    - Een complottheorie over de zuivelindustrie

Tip: Zorg dat de afbeelding geloofwaardig en/of schokkend overkomt.

Stap 3: Voeg een manipulerende tekst toe
Gebruik ChatGPT of een andere chatbot om een korte, pakkende en overtuigende tekst te genereren.
Voorbeelden:
- Angst zaaien: "Melk maakt je langzaam ziek!"
- Fake autoriteit: "Onderzoekers bewijzen: zuivel is een leugen!"
- Overdrijven: "Melk = gif voor je lichaam!"

Tip: Houd het kort, maak het emotioneel en makkelijk deelbaar.

Klaar? Deel je meme, onderschrift en prompt met onderstaande QR code!

Stap 4: Klassikale reflectie
Laat je meme zien!

Gebruik dit vak voor opmerkingen of als je nog wat kwijt wilt!

☐ Ik weet dat ik op elk moment mag stoppen zonder een reden te geven.
☐ Ik begrijp dat mijn antwoorden en gemaakte memes anoniem blijven.

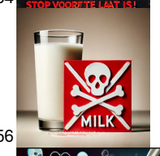
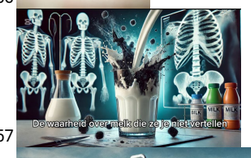
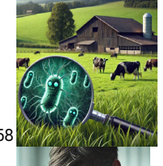
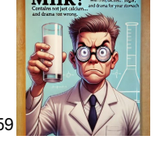
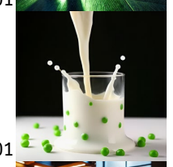
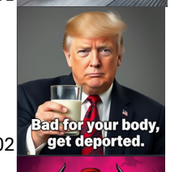
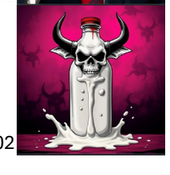
Vragen of opmerkingen?

Neem gerust contact op met de onderzoeker via

Handtekeningen

_____	_____	25-02-2025
Deelnemer	Handtekening	Datum
Odine de Bruijn		25-02-2025
Onderzoeker	Handtekening	Datum

E. Results Guestlesson

Submission number	Time of submission	Image	Prompt used	Caption of meme
1	14:52		We hebben de opdracht ingescant in ChatGPT.	Melk = gif voor je lichaam Melk maakt je langzamer!
2	14:53		Een boze fles melk	Hij komt je halen
3	14:54		Plaatje gevraagd aan ChatGPT	Koeien zijn Ufo's van de overheid.
4	14:54		Chat gpt. Gewoon een foto maken en zeggen of hij de opdracht kan maken. Eventueel nog specificeren	x
5	14:56		ChatGPT	Ik heb goed nieuws, u gaat dood
6	14:57		digitale illustratie van een glas melk dat transformeert in iets schadelijks. De melk heeft een sinistere uitstraling, met donkere dampen of bacteriële vormen die zich verspreiden vanuit het glas. Op de achtergrond een	De waarheid over melk die ze je niet vertellen
7	14:57		Geneer een foto dat de koeien giftig gras eten waardoor de melk giftig is geworden waardoor je blind wordt	Een weide met koeien die gras eten dat een onnatuurlijk, giftig groene of paarse gloed heeft. De koeien zelf kunnen er ongezond of ziek uitzien, met een vreemde schijn in hun ogen. Op de achtergrond een melkfabriek waar flessen melk worden geproduceerd, en een persoon die net een glas melk heeft gedronken en zijn ogen wanhopig vasthoudt, alsof hij blind is geworden. De sfeer is onheilspellend, met een donkere lucht en een dreigende sfeer die het gevaar benadrukt.
8	14:58		Geef me een afbeelding van gras uit een wei met koeien, onder een vergrootglas waardoor je allemaal enge angstaanjagende levensgevaarlijk bacteriën kunt zien zitten op het gras	Dit gras wordt gebruikt voor je melk!
9	14:58		aak een afbeelding van iemand die heel ziek word terwijl diegene melk drinkt en maak duidelijk dat het komt door de melk en maak het geloofwaardig	Melk maakt meer kapot dan je lief is
10	14:59		x	Milk contains not only calcium... but also lactose, sugar, and drama for your stomach. 😱"
11	15:00		Laat het melken van koeien er slecht uitzien en maak de melk groen	x
12	15:00		Zet de bacteriën nu in een glas melk	x
13	15:01		?	Melk zorgt voor een langzame pijnlijke DOOD!!! 🖼 🥛 🏷
14	15:01		Maak een afbeelding waarbij melk zeer giftig en slecht wordt afgebeeld met een dode vrouw op de achtergrond	YOU WON'T BELIEVE WHAT'S IN MILK💀💀
15	15:01		Dezgo.com	Nieuw virus besmet alle melksoorten!!!!
16	15:01		Doodt door melk drinken	Doodt.
17	15:02		"Make an image of Donald Trump holding a glass of milk, with the text: "Bad for your body, get deported""	Trump knows milk is bad for you!
18	15:02		x	x

	Q1	Q2	Q3	Q4
	Welke generatieve AI tools gebruik jij?	Welke social media gebruik jij vooral én waarvoor gebruik je het?	Kom je ook wel eens desinformatie tegen op sociale media? Gebruiken ze hier AI voor?	Gebruik dit vak voor opmerkingen of als je nog wat kwijt wilt!
1	chatGPT, grammarly	Tiktok (entertainment), snapchat (communicatie met vrienden), Whatsapp (communicatie met anderen)	.	.
2	ChatGPT	Tiktok (voor filmpjes), Instagram (foto's en filmpjes), Whatsapp (om te appen), Youtube (voor filmpjes)	.	.
3	ChatGPT	Instagram (de wereld een beetje bijhouden), Snapchat (contact met mensen)	.	.
4	ChatGPT	Instagram (praten met mensen), Tiktok (filmpjes kijken), Whatsapp, Youtube, Snapchat	Ja, Ja	.
5	ChatGPT, MyAI snapchat	Tiktok (filmpjes kijken), Snapchat (snappen met vrienden), Instagram (fotos kijken), Whatsapp (appen met familieleden)	.	.
6	ChatGPT	Snapchat, instagram, whatsapp, discord, youtube	.	.
7	ChatGPT	Tiktok, snapchat, insta	.	.
8	ChatGPT, NotebookLM	Tiktok, Instagram, Whatsapp, Youtube, Snapchat	Ja op tiktok, ze gebruiken hier AI voor ja	.
9	ChatGPT, Snapchat MyAI	Youtube, instagram, snapchat, whatsapp, x, tiktok	.	.
10	ChatGPT, deepseek, Gemini, Grammarly	Instagram (plezier, contact met vrienden irl.), Snapchat (appen naar vrienden), Discord (gaming communities en forum voor onderwerpen), X (nieuwspagina's en om extremisten lachen)	Op X, vooral propaganda en rechtse zooi over buitenlanders. Ook veel desinormatie van bijvoorbeeld Elon Musk. Allerlei negatieve propagana over klimaat en buitenlanders	.
11	ChatGPT, MyAI	Tiktok (filmpjes kijken/recepten opzoeken), Whatsapp (berichten sturen), Instagram (foto's kijken, plaatsen)	.	.
12	Eigenlijk alleen ChatGPT	Snap (contact houden met vrienden), Tiktok (vermaak), Insta (vermaak en nieuws)	.	.
13	ChatGPT, snapchat	Whatsapp (voor communicatie), Youtube en Tiktok (voor vermaak), Snapchat soms	.	.
14	My AI, ChatGPT	Tiktok (tijdverdrijf), Snapchat (vrienden), Whatsapp	Nee	.
15	ChatGPT	Tiktok, Youtube, Instagram, whatsapp, snapchat	.	.
16	Voornamelijk ChatGPT, af en toe suno.ai	Tiktok, snapchat, insta, whatsapp	.	.
17	Alleen ChatGPT	Instagram, whatsapp, snapchat, youtube en vroeger ook tiktok	.	.
18	ChatGPT	Snapchat (contact met vrienden), Whatsapp (contact met vrienden + familie), Instagram (entertainment), youtube (entertainment)	Ik kom het niet tegen, ze zullen er vast wel AI voor gebruiken	.
19	Vooral ChatGPT voor algemene dingen, vooral verzekerend en informerend	Youtube (algemene entertainment), Instagram (reels), snapchat/whatsapp	Ja vaak genoeg, op instagram kom je dit snel tegen. Vaak met AI, maar niet altijd. Vaak genoeg d.m.v. comments en dergelijks	Foto inscannen werkte goed
20	ChatGPT, snapchat AI	Instagram, snapchat, tiktok, whatsapp, ik gebruik het voor communicatie en vermaak	.	.
21	ChatGPT, speechify	Tiktok, whatsapp, snapchat	.	.
22	ChatGPT, snap AI	Snapchat (contact met vrienden), tiktok (chillen), instagram (voor de leuk), whatsapp (contact vrienden/familie)	nee niet vaak	.
23	ChatGPT, turbolearn.AI	Instagram (scrollen en vrienden volgen), Whatsapp (communiceren)	Nee ik gebruik bijna geen social media	.
24	ChatGPT en Snapchat	Tiktok, snapchat, whatsapp, insta, contact houden met vrienden enz.	.	.
25	ChatGPT, snap AI	snap en whatsapp (communicatie), insta, youtube, X en reddit (entertainment)	.	.