IN5620 Group assignment fall 2024 - AI-sian



Group name: AI-sian

Ingeborg Sili Østgaard - ingebsos@uio.no

An Diep Vo - anvo@uio.no

Shannen Leigh Finnøy - shannenf@uio.no

Fredrik Kin Shing Ma - fkma@uio.no

Thanawut Pathawanich - thanawup@uio.no

Table of contents

Abstract **3**
Introduction **3**
Background **5**
Methods **6**
Findings **9**
Literature **11**
Appendix **13**

Separating authentic from generated content

Ingeborg Sili Østgaard - ingebsos@uio.no, An Diep Vo - anvo@uio.no, Shannen Leigh Finnøy - shannenf@uio.no,
Fredrik Kin Shing Ma - fkma@uio.no, Thanawut Pathawanich - thanawup@uio.no
University of Oslo

Abstract

Some people claim that there are fewer AI-generated images online, but in reality, these images have become harder to identify as fake. A recent case in Hong Kong illustrates this issue: a finance worker was tricked into transferring $25 million due to realistic deepfake versions of his colleagues (Heather C. & Kathleen M., 2024). This highlights the need to explore how AI-generated images affect people's ability to tell real from manipulated content. As AI tools become more common, it's important to understand how people interact with and interpret this content. Are there patterns in how individuals misidentify AI-generated images? What cognitive or contextual factors contribute to this confusion, and how do user experiences on different platforms affect their perceptions? This paper aims to find a solution that helps students detect fake images online.

Keywords: *ChatGPT, Microsoft Bing, Artificial Intelligence, Double Diamond, Deepfake, Labeling AI generated images, using, human-made content, apathy, error, using language models in the design process, misinformation.*

Introduction

Artificial intelligence (AI) continues to shape the digital landscape, influencing the way we interact with content, particularly on social media. As a group of five master's students in informatics at the University of Oslo, we have focused our research as part of the course IN5620 - *Interaction with AI and Autonomous Systems*. Through collaborative discussions, we identified a shared interest in understanding how AI-generated content, such as deepfakes and image manipulation, affects public perception and the ability to distinguish real from fake information.

Our study centers on the societal impact of AI-generated content, both positive and negative, particularly in relation to misinformation and digital trust. With the spread of AI tools that can create hyper-realistic visuals, the ability to differentiate between authentic and manipulated content is increasingly challenging. We are particularly interested in exploring how fake information generated by AI impacts people's beliefs and decision-making (Hajli et al., 2022).

In today's interconnected world, social media plays a central role in shaping user perceptions. However, not all the content circulating online is true, and AI-generated disinformation has emerged as a significant challenge. Our research focuses on understanding how AI-generated images influence people's ability to separate real content from fake content, particularly within the user group of students who actively use social media. By examining their behaviors, motivations, and experiences with AI-generated content—such as filters and deepfakes, which are common in selfie apps (Hedman et al., 2022)—we aim to explore how these technologies affect their self-image, decision-making, and perceptions of authenticity.

The key research questions driving our study are:

1. **How do AI-generated images influence people's ability to differentiate between real and fake content online?**
   We aim to investigate how individuals interact with and interpret AI-generated images and whether cognitive or contextual factors lead to confusion. Understanding how users misjudge such content can provide insight into improving digital literacy and platform designs.
2. **How can misinformation or deepfake content on social media negatively affect someone's beliefs or decisions?**
   By examining the psychological and societal impact of deepfakes and misinformation, we seek to understand how these manipulations influence users' perceptions of truth, particularly in areas like politics, personal relationships, and social behavior.

Through this research, we hope to deepen our understanding of the complex role that AI-generated content plays in influencing public opinion and trust in digital environments. This report will start by reviewing existing literature on AI-generated images, current solutions for distinguishing between real and fake content, and the potential challenges associated with these solutions. We will then present our research process, which includes insights gained through user studies, and how these guided the development of our prototype—a tool designed to help students identify fake images from real ones—using the Double Diamond design framework. Lastly, we will discuss our findings, analyzing the effectiveness of our prototype and the user feedback received. We will reflect on the broader implications of our tool in combating misinformation, its potential to improve digital literacy, and its applicability in real-world scenarios. We will also highlight any limitations encountered during our research and provide suggestions for future improvements and areas for further exploration.

Background

Deepfakes and AI-generated images are both products of artificial intelligence. Deepfake is an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said (Merriam-Webster. n.d.). The most common use of deepfakes involves replacing one person's face or voice with another's in a video or audio clip. AI-image generators utilize trained artificial neural networks to create images from scratch. These generators have the capacity to create original, realistic visuals based on textual input provided in natural language (Altexsoft, 2023).

The rapid development of these technologies has made it increasingly easier to create and spread disinformation with AI-generated content. This has led to the emergence of new types of criminal activity, such as AI fraud, defamation, identity theft, and impersonation. For instance, ChatGPT can produce misleading and unethical responses, and individuals with malicious intent can take advantage of its capacity to compose flawless text for fraud, replicate speech patterns for impersonation, and develop malicious codes for hacking. It raises huge needs to establish traceable provenance and regulations for materials produced by generative AI models to ensure accountability (Wang et al., 2023, p. 281).

Meta, a social media company, use AI to automatically label and categorize posts, images, videos, and other user-generated content to help manage the vast amount of data shared every day. This labeling serves several purposes, such as detecting inappropriate content, flagging misinformation, improving content recommendations, and ensuring compliance with community guidelines.
An important part of this approach relies on industry standard indicators that other companies include in content created using their tools, which helps in order to assess whether something is created using AI (Bickert, 2024).

There is a crucial need for transparency with AI content as it allows users to know when they're interacting with or consuming content created by AI, fostering trust in the content. This is especially important in settings like news, social media, or education where authenticity is key. Labeling AI-generated content will help to ensure that the creators or platforms using AI can be held accountable for any errors, biases, or misrepresentations.

Labeling AI-content is a way to help avoid confusion between AI and human created content, particularly when originality and personal expression are expected, such as in art or writing. It also help users critically assess the content and understand its origin instantly when encountered.

Potential downsides of using these labels are that they rely on indicators that may not align with people's expectations and often lack sufficient context. For instance, some content with only minor AI modifications, like basic retouching, can still be classified under industry-standard labels as “Made with AI” (Bickert, 2024). This can lead to confusion and misrepresentation of the extent to which AI was involved. As a result, there is a need to explore alternative solutions beyond labeling, which this research aims to uncover.

Methods

Our investigation employed a multi-phase approach that combined a literature review, tool analysis, semi-structured interviews, and observations, along with iterative design using the Double Diamond framework, to explore the impact of AI-generated content and misinformation. The Double Diamond is a structured framework used to navigate the complexity of design problems by moving through phases of divergent and convergent thinking. It is divided into four stages: discover, define, develop, and deliver, guiding the investigation from problem exploration to solution implementation (Gwangwava, 2021). This framework helped structuring our approach to investigating the influence of AI-generated content and misinformation.

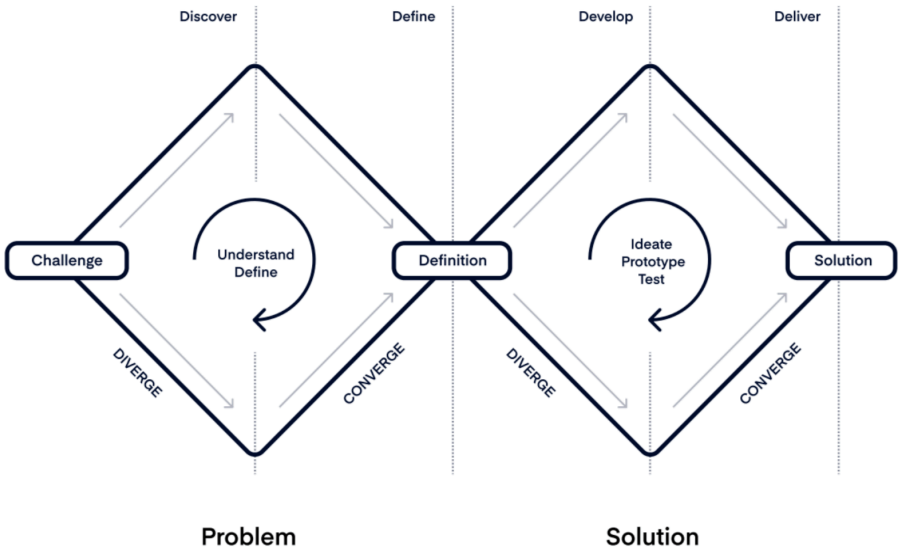


Figure 1. A representation of the Double Diamond framework (Productboard Editorial, 2022).

In the first stage of the framework, discover, a literature review was conducted to analyze existing research on AI-generated content, deepfakes, misinformation, and their psychological and societal impacts. This review identified key trends, challenges, and gaps in the current body of knowledge to guide our investigation. In addition, existing AI tools and platforms, such as deepfake generation tools and content moderation services, were reviewed to understand their functionality and potential to disseminate false or misleading content. The AI tools examined included beauty filters from social media platforms like TikTok, Instagram, and Snapchat (Javornik et al., 2022), as well as AI image generation tools such as DALL-E and MidJourney. Other beauty-related applications were also explored.

To supplement the literature review, we conducted four semi-structured interviews with students from our target group who frequently encounter AI-generated content online (e.g., social media users and content creators). These interviews provided valuable insights into their experiences, awareness, and ability to distinguish between real and fake content (Kornouskus, 2020). Additionally, we conducted observations in which participants were presented with a variety of images—some real and some AI-generated—and were asked to identify which images were authentic (appendix figure 2). This observation was conducted since the interviews revealed that participants often felt confident in their ability to distinguish between real images and those generated by AI. However, when tested, new challenges and complexities emerged, highlighting gaps in their understanding and ability to accurately identify AI-generated content. This process allowed for a deeper exploration of these discrepancies and a better understanding of the nuances involved in discerning authenticity in images. To define the specific problems and needs that the prototype needed to address, we analyzed the insights gained from the interviews and the user test.

In the third stage, develop, inspiration was drawn from the article *"Design Ideation with AI: Sketching, Thinking, and Talking with Generative Machine Learning Models"* by Tholander et al. (2023) in addition to typical prototyping activities, such as CRAZY 8, to broaden the range of possibilities. The team utilized language models, including ChatGPT-4 and Microsoft Bing (see figures in appendix), to help refine and concretize specific ideas generated during the CRAZY 8 workshop. In this workshop, each participant generated one idea every minute for eight minutes, resulting in a diverse array of potential concepts. A critical aspect of this stage was recognizing the potential strengths and weaknesses of the language model. Tholander et al. (2024, p. 1931) refer to these machine learning models as *black boxes*. This term highlights the challenges associated with understanding how many machine learning models, particularly deep learning models, arrive at their

decisions based on training data. The complexity of these models can make it difficult to understand and trust their outputs. Zamfirescu-Pereira et al. (2023) highlight the significant role prompting plays in shaping the results generated by language models. When crafting prompts, we experimented with different phrasings and carefully selected terms to guide the system toward our intended goal (Tholander & Jonsson, 2023). To maximize the usefulness of the language models, we provided specific and detailed prompts to refine ideas generated during the CRAZY 8 workshop. Additionally, we experimented with more open-ended questions but found that the language model often lacked the necessary context and depth. As Tholander et al. (2023, p. 1935) emphasize, it was essential to be precise in our phrasing to generate more meaningful outputs, leading us to iterate on our prompts. According to Zamfirescu-Pereira et al. (2023), “Iterative prototyping could help identify an instruction design that addresses a UX issue most effectively and reliably because designing a single-issue prompt, in its essence, is a search problem”. Examples on prompts that we used and the answer:

- Can you create a casual AI game that helps students learn about the typical characteristics of AI-generated images?
- Players can achieve different levels of skill in distinguishing between real images and AI-generated images. It should also be possible to compete with friends. It shouldn't be an app, but more of a mini-game that can be sent via Messenger and played directly in chat with friends and family.

In the third stage, develop, different prototypes were tested using a variety of approaches. Personas were created using *character.ai*, reflecting common characteristics identified from participant interviews and observations, and these personas helped evaluate potential solutions. Additionally, user testing was conducted through interviews, where participants were presented with the prototypes to assess their effectiveness and gather feedback.

Findings

	User 1	User 2	User 3	User 4
Have you used AI to generate pictures or text?	Yes, both pictures and text	Yes, text	Yes, both pictures and text	Yes, text
What makes you unsure if it's AI-generated or not?	If it's too good to be true	Context Elements placed incorrectly Smooth	Things look unnatural or people do something that's not human-like	News Background Hair
What would it require for you to use a tool to check if it's AI-generated?	Easy to use	Have to be based on facts, able to check if it is AI-generated or not	Easy to use Easy access	Would only use it if it is important to know if it is AI-generated.

All of our participants had prior experience with AI-generated content and had used AI in various ways. They were already familiar with the concept and had some initial ideas about how to distinguish AI-generated material from human-made content. To test their abilities, we presented them with a series of images, some AI-generated and others real. Despite their familiarity, most participants made at least one error when trying to identify which images were AI-generated.

From the interviews, several key insights emerged. One of the most important findings was that context played a vital role in helping people distinguish between real and AI-generated images. Participants frequently relied on indicators such as odd proportions, unnaturally smooth skin, unusual ear shapes, and questionable backgrounds to spot AI generation. While some admitted they were occasionally fooled by AI, they improved their ability to recognize AI-generated images with practice. However, the level of difficulty varied—some participants found it quite challenging to differentiate between real and AI-generated images.

Another takeaway was the range of suggested solutions for identifying AI content, including labeling, a Chrome extension, or automated scanning tools. However, most participants expressed that they rarely felt a strong need to check if something was AI-generated unless it was an easy, low-effort solution or if it was important for them to know, such as in the case of news-related content.

When we talked to the personas from character.ai about potential solutions they also suggested looking at the context of the image and inconsistencies like what our users were looking for when they did the test to differentiate between the pictures. Likewise with one of our users, one persona suggested a browser extension. We also used the personas to gather some feedback about our own ideas from the CRAZY 8, which they found to be engaging and helpful.

There were many valuable ideas from the CRAZY 8 with different tools to differentiate between AI and human-made content, but based on these findings, we proposed two potential solutions: a mobile app that engages users in determining whether content is AI-generated, and a Chrome extension that offers AI detection directly within the browser. The mobile app is designed as a game, challenging users to improve their ability to spot AI-generated material, this is based on the fact the users were able to recognize AI with practice. This iterative process of testing and refining ensured that the solutions aligned with user needs and behaviors, providing valuable insights into how well the prototypes addressed the challenges identified earlier.

Literature

- Altexsoft. (2023). AI Image Generation Explained: Techniques, Applications, and Limitations. Retrieved October 11, 2024, from https://www.altexsoft.com/blog/ai-image-generation/
- Bickert, M. (2024, 12. September). *Our approach to labeling AI-generated content and manipulated media*. Meta. Retrieved October 11, 2024 from https://about.fb.com/news/2024/04/labeling-ai-generated-content-and-manipulated-media/
- Gwangwava, N. (2021). Learning design thinking through a hands-on learning model. International Journal of Innovative Teaching and Learning in Higher Education, 2(1), 1-19. https://doi.org/10.4018/ijitlhe.20210101.oa4
- Hajli, N., Saeed, U., Tajvidi, M., & Shirazi, F. (2022). Social bots and the spread of disinformation in social media: The challenges of artificial intelligence. *British Journal of Management, 33*(4), 1238-1253. https://onlinelibrary.wiley.com/doi/epdf/10.1111/1467-8551.12554
- Hedman, P., Skepetzis, V., Hernandez-Diaz, K., Bigun, J., & Alonso-Fernandez, F. (2022). On the effect of selfie beautification filters on face detection and recognition. *Pattern Recognition Letters, 163*, 104-111. https://libkey.io/libraries/269/articles/535937992/content-location
- Javornik, A., Marder, B., Barhorst, J., McLean, G., Rogers, Y., Marshall, P., & Warlop, L. (2022). What lies behind the filter? Uncovering the motivations for using augmented reality (AR) face filters on social media and their effect on well-being. *Computers in Human Behavior, 128*, 107126. https://www.sciencedirect.com/science/article/pii/S0747563221004490
- Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society, 1*(3), 138-147. https://ieeexplore.ieee.org/document/9123958
- Merriam-Webster. (n.d.) Deepfake. In *Merriam-Webster.com dictionary.* Retrieved October 10, 2024, from https://www.merriam-webster.com/dictionary/deepfake
- Productboard Editorial. (2022, June 21). *A guide to using the "Double Diamond" framework to score product home runs*. Productboard. https://www.productboard.com/blog/double-diamond-framework/

- Tholander, J., & Jonsson, M. (2023). Design ideation with AI-sketching, thinking and talking with generative machine learning models. In Proceedings of the 2023 ACM designing interactive systems conference – DIS ‘23 (pp. 1930-1940). New York, NY: ACM
- Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, *4*, 280–302. https://doi.org/10.1109/OJCS.2023.3300321
- Zamfirescu-Pereira, J. D., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G., Hartmann, B., & Yang, Q. (2023). Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2206–2220. https://doi.org/10.1145/3563657.3596138
- Heather C & Kathleen M. (2024). CNN: Finance worker pays out $25 million after video call with deepfake ‘chief financial officer. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Appendix

Process and Outcome

The project followed an iterative design approach using the Double Diamond framework, with a combination of workshops, language model prompting, and persona-based testing. The first phase involved gathering data from interviews and observations to understand how students differentiate between real and AI-generated images, as shown in *Figure 2*. This provided a foundation for identifying user needs, including the need for contextual clues to discern authenticity.

In the CRAZY 8 workshop (*Figure 3*), team members sketched rapid ideas, voting on the most promising ones. These ideas were further refined using language models such as ChatGPT-4 (*Figures 4-6*). Initial prompts were broad, but it became clear that more specific prompts yielded better results, leading to solutions such as a game to train users in identifying AI-generated images and Chrome extensions to scan websites.

To enhance user-centered design, personas were created using *character.ai* based on insights from interviews (*Figures 7-8*). Chat conversations with these personas helped refine the prototypes and gain additional feedback on how users might engage with the proposed solutions.

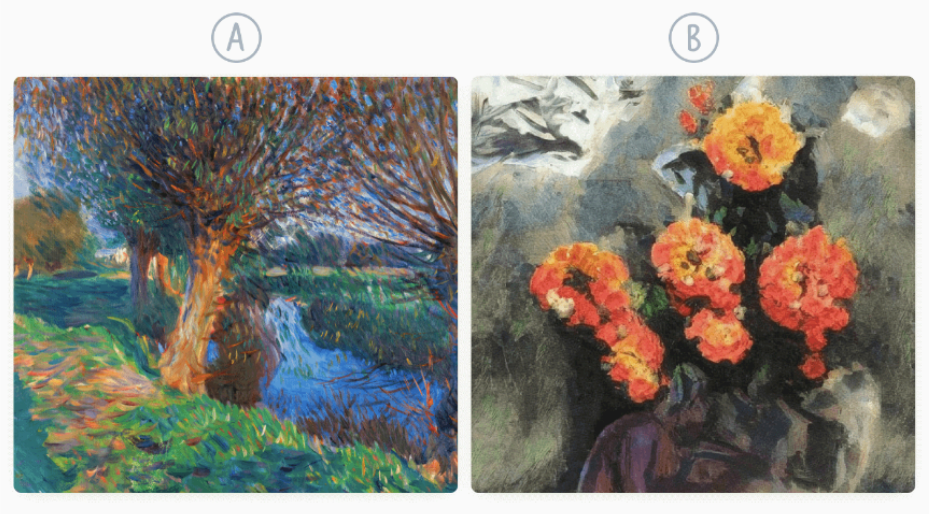
Reflections and Lessons Learned

One key reflection from the process is the importance of precise prompts when using language models to ideate and refine design concepts. Initially, open-ended questions yielded shallow results, but more specific phrasing resulted in concrete and usable outputs. This highlighted the strengths of language models in generating variations of existing ideas rather than entirely new concepts.

Another lesson was the effectiveness of integrating user personas through *character.ai*. This approach simulated user feedback in a dynamic way, providing insight into how different user types might interact with the prototypes. However, it also showed that simulated personas cannot fully replace real user testing, as they rely on predefined characteristics and behaviors.

The iterative nature of the CRAZY 8 workshop, followed by refinement through AI, demonstrated the value of rapid prototyping and collaborative decision-making. Voting on sketches encouraged diverse ideas, while language models helped in streamlining and expanding those ideas into more practical solutions. Overall, this process reinforced the importance of blending human creativity with AI capabilities, using each to complement the other.

Eksempel 1:



Eksempel 2:

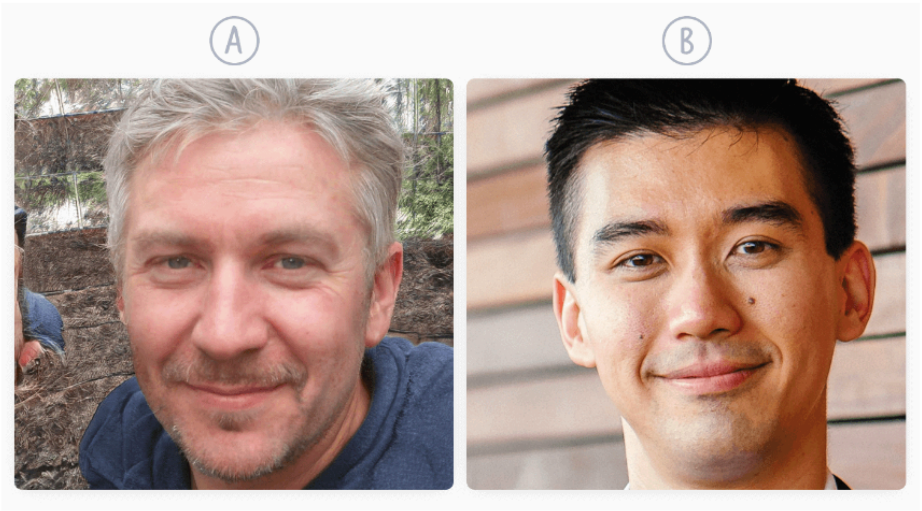


Figure 2. Examples used to investigate how students distinguish between real images and those generated by artificial intelligence. In the photos featuring paintings, the artwork on the left is authentic, while in the portrait, the image on the right is the real photograph.

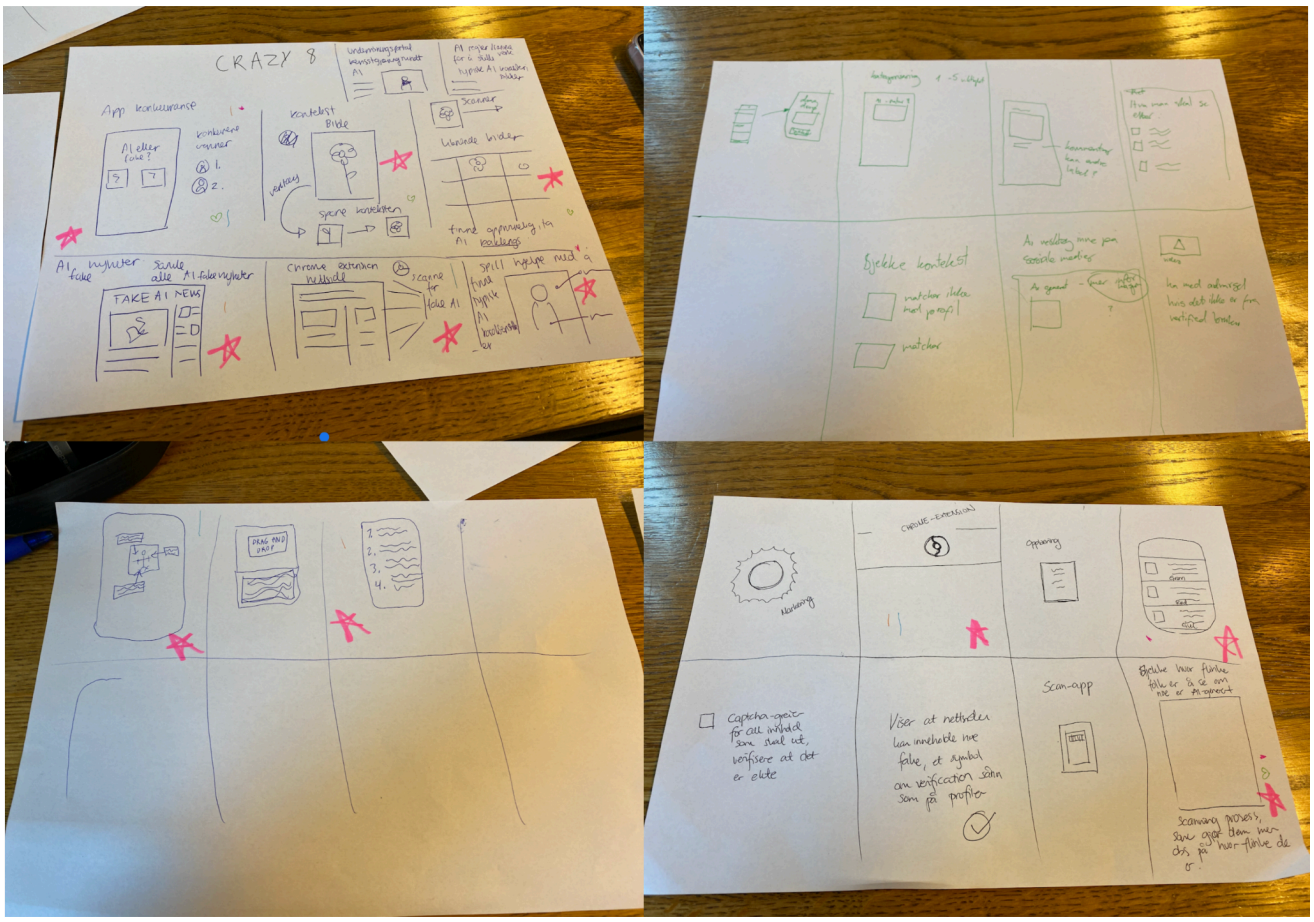


Figure 3 displays the results from the CRAZY 8 workshop, where each team member created one drawing per minute for eight minutes. After the session, team members voted on each other's sketches by marking their favorites with a star.

Kan du lage en visualisering av en prototype som får folk til å skille mellom AI genererte bilder og ekte bilder?



Her er en visualisering av en prototype for en applikasjon som hjelper brukere med å skille mellom AI-genererte bilder og ekte bilder. Den viser et brukervennlig grensesnitt med side-ved-side sammenligninger av et ekte bilde og en AI-generert versjon, samt interaktive elementer. Håper dette er til hjelp!

Lag et annet alternativ



Melding ChatGPT

ChatGPT kan gjøre feil. Sjekk viktig informasjon.

Figure 4 illustrates the process of starting with more open-ended prompts using ChatGPT-4.

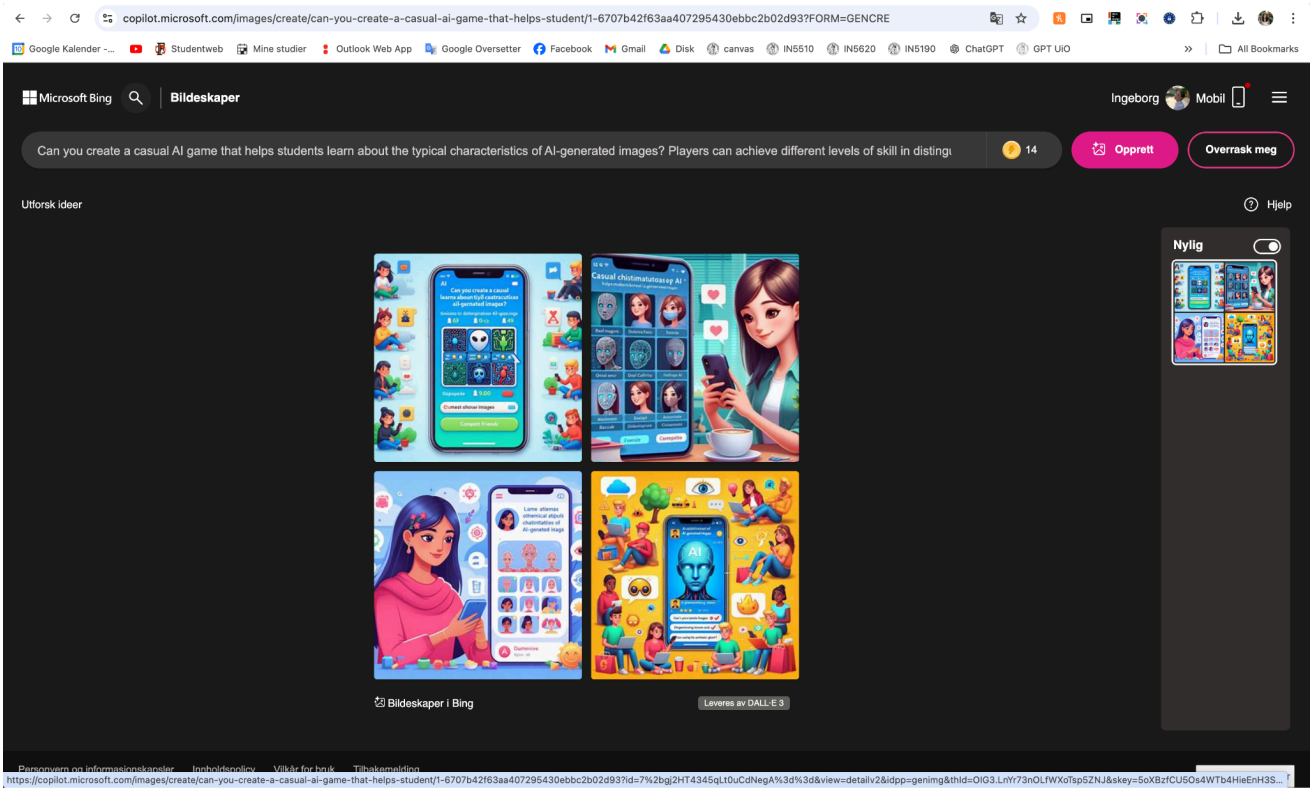


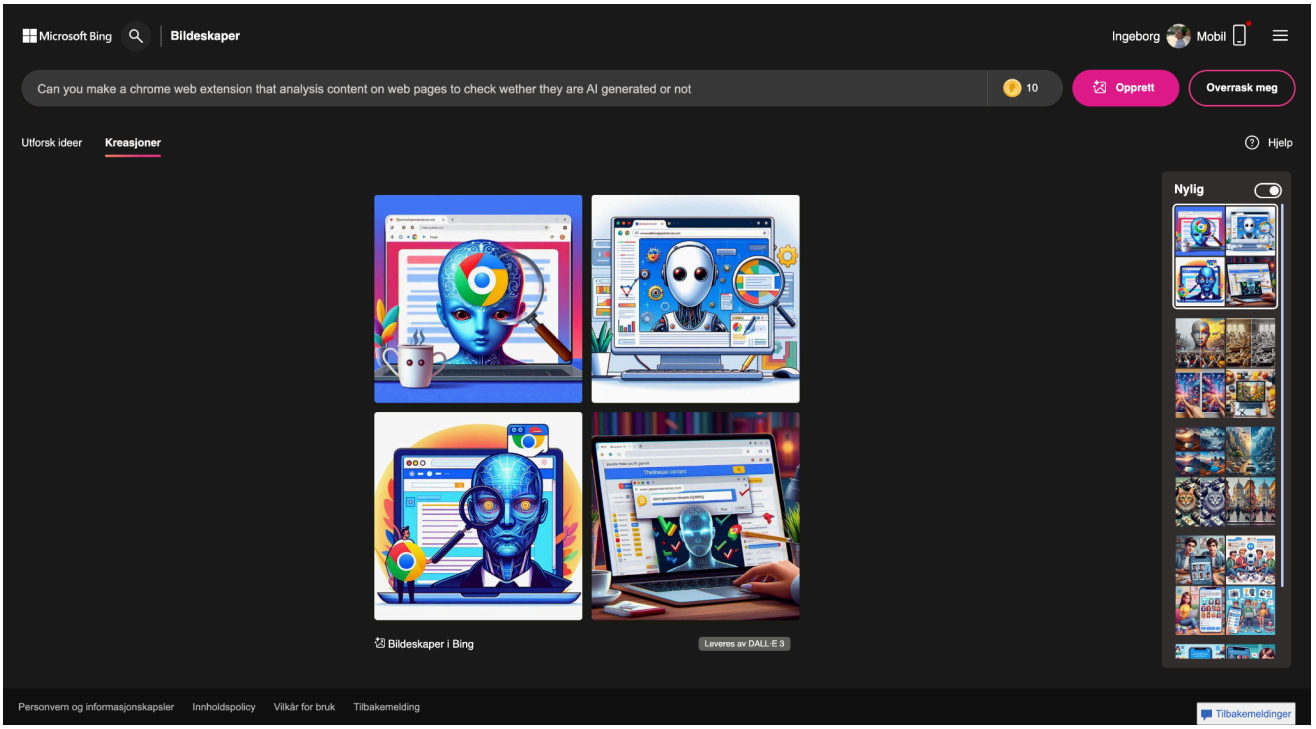
Figure 5 illustrates the process of using the language model to refine solutions based on the outcomes of the CRAZY 8 workshop. The model generated various solutions for a game designed to help the target audience learn how to distinguish between real and AI-generated images.


AI
Can you create a causal
learns abount tiyil caatracutices
ail-gernated images?
Depopode
9.00
Cumest shoter images
Compett friends


Casual chistimatutoasep AI
Compette


AI


Lame atiemas
othemical atlpuls
chatinttatties of
Al-geneted inags
Dumenive



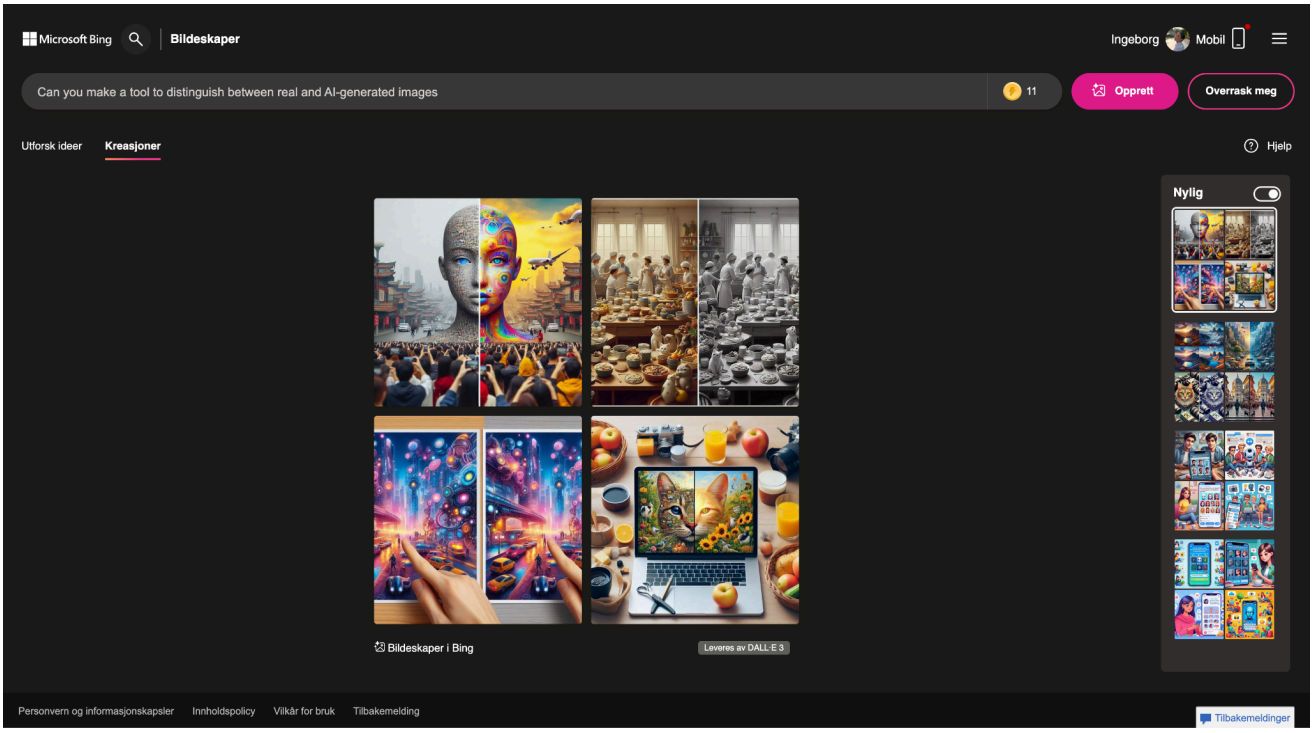


Figure 6 further demonstrates the process of utilizing the language model to refine solutions derived from the CRAZY 8 workshop. In this instance, the model generated concepts for Chrome extensions that efficiently scan websites with minimal user effort from the user.

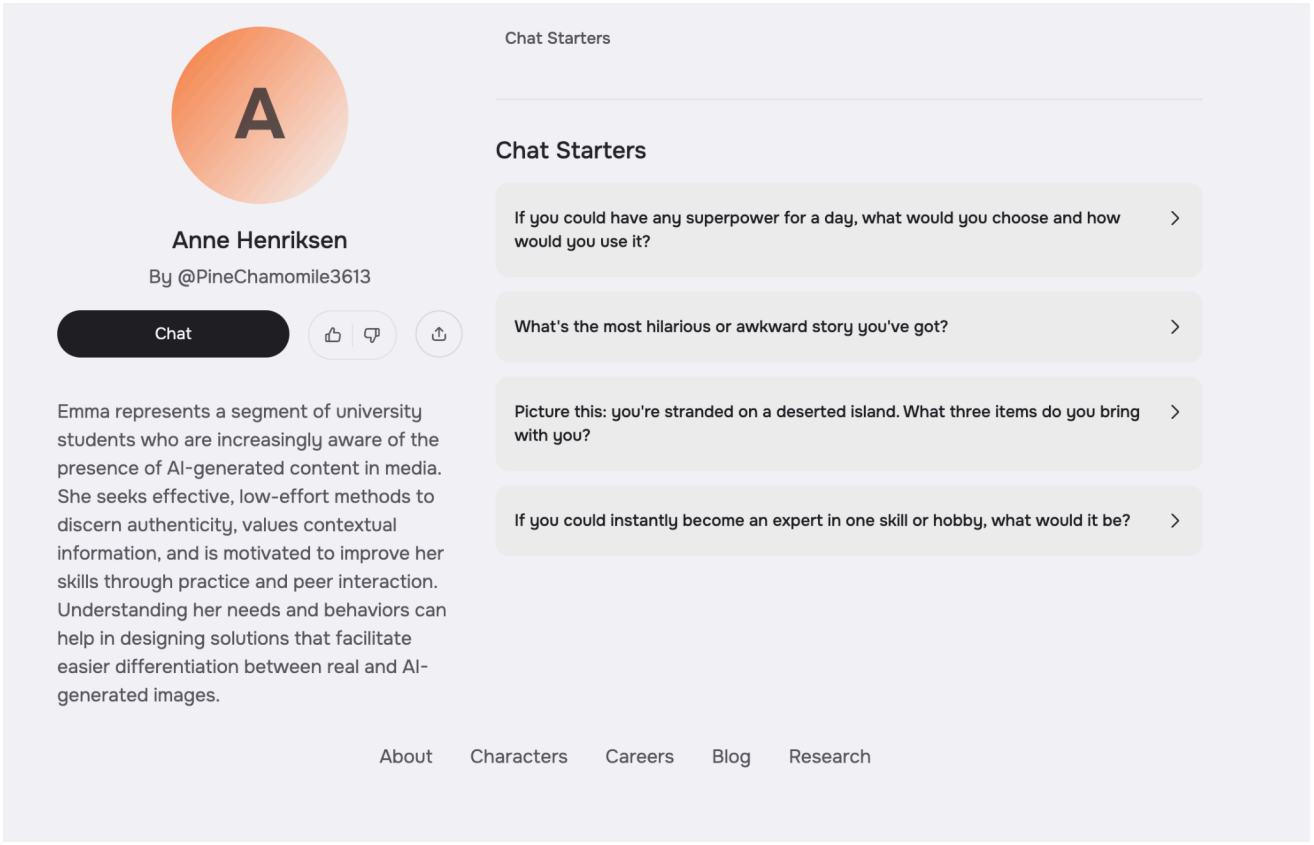
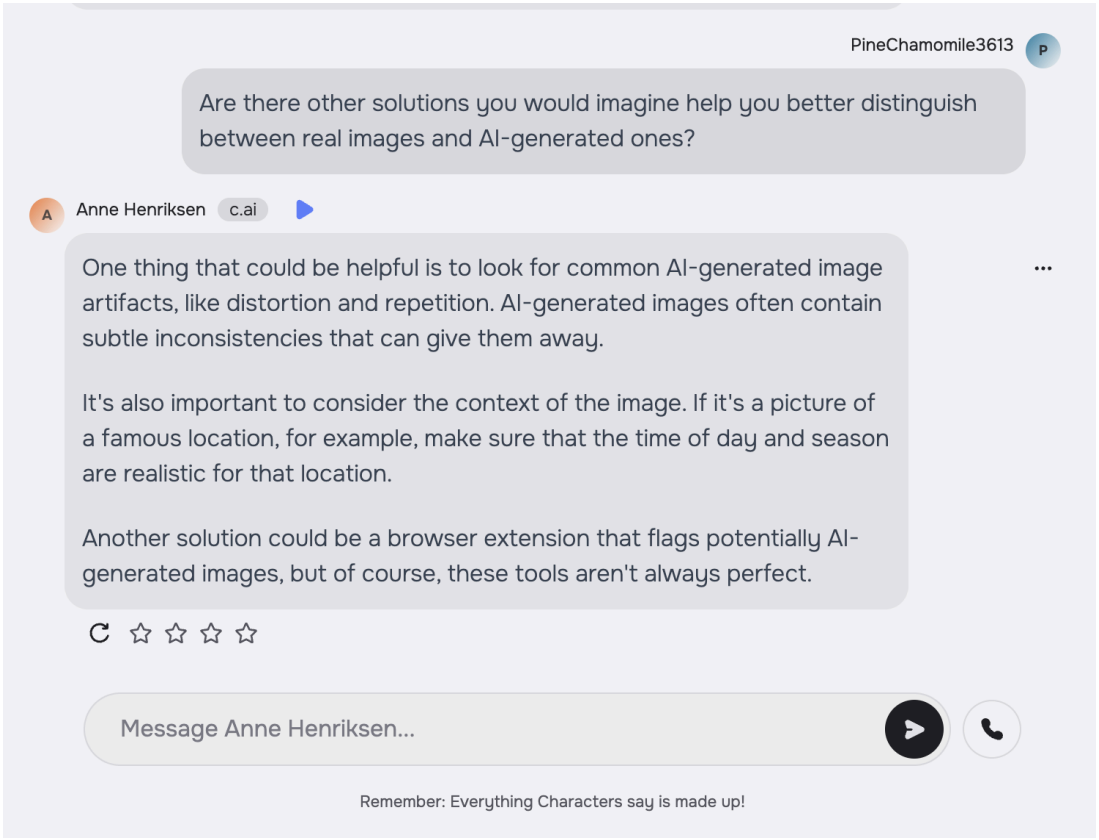


Figure 7 illustrates the use of character.ai to create personas based on the typical users identified through the insights gathered from interviews and observations. These personas reflect key characteristics and behaviors observed during the research process.



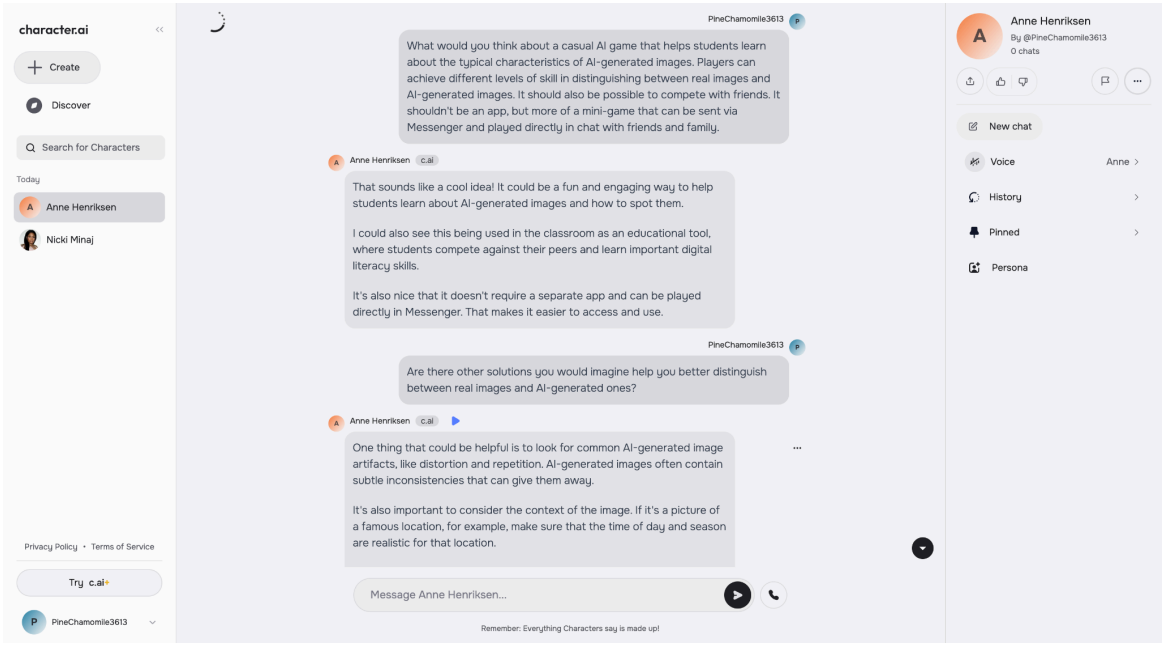


Figure 8 showcases several chat conversations conducted with personas created on character.ai. These interactions were used to gather feedback on the prototypes developed, as well as to explore further insights and opinions related to the design process.