

Leiden University

ICT in Business and the Public Sector

Can Children Spot AI? Distinguishing AI-Generated Images from Real Images in Primary Education

Name:	Boele Visser
Student ID:	s2934841
Date:	July 8, 2026
1st supervisor:	Natalia Amat Lefort
2nd supervisor:	Max van Duijn

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

Background: Generative AI is increasingly present in children's everyday lives, through tools such as ChatGPT, voice assistants, AI companions and AI-generated visual content. As children are exposed to more AI-generated content, it becomes important to assess whether they can critically evaluate it and distinguish AI-generated images from real ones. Previous research suggests that cognitive abilities, trust in AI, AI usage, and education may influence how people evaluate AI-generated content, but little is known about these relationships among children.

Aim: This study investigates how accurately children can distinguish AI-generated from real images and how this ability is influenced by fluid intelligence, AI usage, trust in AI, and gender. Additionally, the study evaluates whether a short AI literacy lesson improves children's visual AI-distinguishing accuracy.

Method: We conducted an experiment with 65 Belgian primary school students. Participants completed a questionnaire measuring AI usage, trust in AI, fluid intelligence, and their ability to correctly classify images as AI-generated or real. They then followed an interactive AI literacy lesson about algorithms, machine learning, ethical considerations in the usage of AI, and visual cues for recognizing AI-generated images. After the lesson, participants completed a second image classification task.

Results: Before the lesson, children performed slightly above chance, with an overall accuracy of 51.43%. Participants were better at identifying real images (60.26%) than AI-generated images (42.48%). Fluid intelligence, AI usage, and trust in AI did not significantly predict pre-experiment AI-distinguishing accuracy. However, AI usage was positively associated with trust in AI. Gender was not significantly related to either fluid intelligence or visual AI-distinguishing accuracy.
Contrary to our expectations, participants did not improve after the AI literacy lesson. Instead, post-experiment performance was significantly lower than pre-experiment performance. This decrease may be explained by cognitive fatigue and over-awareness of AI cues, leading to wrong classification. Neither fluid intelligence nor gender significantly influenced the extent of change in children's performance after the lesson.

Conclusion: Our study shows that children struggle to distinguish AI-generated images from real images, and that a short, standalone lesson is insufficient to improve this specific competency. As AI-generated content becomes more common, AI literacy education should be integrated more systematically into primary education, for example through regular digital literacy programs or in a media literacy class. Such education should go beyond explaining how AI works and include repeated practice with both real and AI-generated content, feedback on children's reasoning, and practical strategies for evaluating whether online content can be trusted.

Acknowledgements

I wish to express my deepest gratitude to my thesis supervisor, Natalia Amat Lefort, for her guidance, support, and valuable feedback throughout this thesis. Her insights and encouragement were instrumental during this process. I am also sincerely grateful to Max van Duijn, whose comments encouraged deeper reflection and helped to refine my ideas.

Furthermore, I would like to thank Sus Dams and his colleagues at Natuur en Wetenschap vzw for facilitating contact with schools and arranging participation in the experiment. I am also very grateful to the 78 students and their teachers who dedicated their time to participate. Their contributions were essential to this thesis.

I would also like to thank my family and friends for their continuous support, patience, and encouragement throughout my studies. Their support and numerous conversations helped me remain focused.

Finally, I wish to acknowledge all individuals who contributed to this thesis through their support, advice, or encouragement. Their contributions are greatly appreciated.

Contents

Abstract . . . 1
Acknowledgements . . . 1

1 Introduction **3**

2 Background and related work **5**
2.1 Background . . . 5
2.1.1 Definition of AI . . . 5
2.1.2 How generative AI systems produce artifacts . . . 5
2.1.3 AI ethics . . . 6
2.1.4 AI literacy . . . 7
2.1.5 AI-education tools . . . 7
2.2 Related work . . . 9

3 Theoretical Framework **10**
3.1 Fluid intelligence . . . 10
3.2 Fluid intelligence & distinguishing AI-generated artifacts . . . 11
3.3 Trust . . . 11
3.4 Gender, fluid intelligence, and distinguishing AI . . . 12
3.5 Impact of education . . . 12
3.5.1 Variables moderating lesson effectiveness . . . 13

4 Method **14**
4.1 Measures . . . 14
4.1.1 AI Usage & Trust in AI . . . 14
4.1.2 Fluid intelligence . . . 15
4.1.3 Ability to distinguish AI and real images . . . 16
4.2 Procedure . . . 16
4.2.1 Questionnaire . . . 16
4.2.2 Lesson . . . 18
4.3 Participants . . . 19
4.4 Analytical Strategy . . . 20
4.4.1 Compensation . . . 20
4.4.2 Hypothesis testing . . . 21
4.5 Ethics . . . 22

5 Results **23**
5.1 Descriptive statistics . . . 23
5.1.1 Fluid Intelligence . . . 25
5.1.2 AI distinguishing . . . 26
5.2 Hypothesis Testing . . . 29
5.2.1 Pre-experiment performance . . . 29
5.2.2 AI usage and trust . . . 29
5.2.3 Effects of gender . . . 30
5.2.4 Post-experiment performance . . . 30
5.2.5 Summary of results . . . 30

6 Discussion **32**
6.1 Contributions to existing literature . . . 32
6.2 Practical implications . . . 34
6.3 Limitations . . . 34
6.3.1 Sample . . . 34
6.3.2 Methodology . . . 35
6.3.3 AI Literacy . . . 36

7 Conclusions **37**
7.1 Answers to the research questions . . . 37
7.2 Contributions . . . 38
7.3 Future work . . . 38

A AI-generated Images **43**

B Example Surveys **47**

C Lesson Slides **62**

Chapter 1

Introduction

Generative artificial intelligence has rapidly become part of everyday life. Tools such as ChatGPT, Google Gemini, voice assistants, AI companions, and image generators are now widely available and increasingly used by both adults and children; in September 2025, OpenAI reported 700 million weekly active users for ChatGPT (Chatterji et al., 2025), whilst AI usage among children increased from 37.1% in 2023 to 77.1% in 2024 (Picton and Clark, 2024). These systems can generate text, speech, images, videos, and conversations that appear realistic and human-like (Vaswani et al., 2017; Dhariwal and Nichol, 2021). As a result, children are growing up in an environment where the boundary between human-made and AI-generated content is increasingly blurred (Kamali et al., 2025).

This development creates new opportunities, but also new risks. AI tools can support learning, creativity, and access to information, but they can also produce inaccurate, misleading, biased, or emotionally persuasive content (Yang et al., 2025). For children, these risks are especially relevant, as they are still developing the cognitive and critical skills needed to evaluate digital information (Nikolaeva et al., 2023). These skills are particularly important in evaluating images that may be AI-generated; they may appear realistic at first glance, but on closer inspection, they contain subtle errors in details such as hands, faces, text, shadows, or spatial consistency (Kamali et al., 2025). If children cannot critically evaluate content that might be AI-generated, they may be more likely to trust or share misleading content, with all the consequences it entails (Robb and Mann, 2025; Yang et al., 2025).

For this reason, AI literacy has become an important part of education. AI literacy encompasses not only understanding how an AI system works but also recognizing its limitations, critically evaluating its outputs, and reflecting on the ethical issues associated with its use (Long and Magerko, 2020; Ng et al., 2021). One specific component of AI literacy is the ability to distinguish between AI-generated content and real human-made content. While previous work has examined adults' ability to detect AI-generated text, fake news, and deepfakes (Chein et al., 2024; Sindermann et al., 2021; Somoray et al., 2025), less is known about how children perform on a visual AI-distinguishing task. It also remains unclear which individual characteristics influence this ability.

This thesis investigates children's ability to distinguish AI-generated images from real images. It also examines whether fluid intelligence, AI usage, trust in AI, and gender are related to this ability. These variables were selected because previous research suggests that the ability to distinguish AI-generated content is influenced by individual characteristics (Chein et al., 2024; Sindermann et al., 2021; Somoray et al., 2025). Fluid intelligence may be relevant because distinguishing AI-generated images requires reasoning and pattern recognition skills (Chein et al., 2024). Prior studies have found that fluid intelligence is related to the ability to distinguish AI-generated text from human-written text and to critically evaluate fake news (Chein et al., 2024; Sindermann et al., 2021). In addition, research on deepfake detection suggests that cognitive ability and analytical thinking influence people's ability to detect manipulated visual content (Somoray et al., 2025).

AI usage and trust in AI are included because children's experience with AI may shape their familiarity with AI-generated content, while trust in AI may affect how critically they evaluate such content. Previous research has found that more frequent AI usage is associated with higher trust in AI (Alasmari et al., 2025) and that higher trust in AI may reduce critical evaluation of AI output (Shaw and Nave, 2026). Finally, gender is included because prior research has reported mixed findings on gender

differences in fluid intelligence, technology attitudes, and trust in AI (Colom and Garcıa-López, 2002; Giofrè et al., 2022; Kaplan et al., 2023). Together, these variables allow this study to examine not only how well children distinguish AI-generated images from human-made images, but also which individual characteristics may help explain differences in this ability.

In addition, our study evaluates whether a short AI literacy lesson can improve children's visual AI-distinguishing accuracy. The lesson introduces children to basic AI concepts, including algorithms, machine learning, and the limitations of AI systems. It also provides practical cues for recognizing AI-generated images. By measuring children's performance before and after the lesson, this study investigates whether a short educational intervention can improve their ability to critically evaluate AI-generated images.

Taken together, these considerations show that children's ability to recognize AI-generated images may be studied not only in terms of overall accuracy but also in relation to individual characteristics and the role of AI literacy education. Therefore, the central research question of this study is:

[RQ] How well are children able to distinguish AI-generated images from real images, and how is this ability influenced by individual characteristics and AI education?

To answer the central research question, this thesis addresses five sub-questions that examine children's baseline performance, the role of individual characteristics, and the effect of an educational intervention:

RQ.1 How accurately do children distinguish AI-generated images from real images before receiving AI literacy education?

RQ.2 How are fluid intelligence, AI usage, and trust in AI related to children's AI-distinguishing accuracy?

RQ.3 How does gender influence fluid intelligence or AI-distinguishing accuracy?

RQ.4 How does an AI literacy lesson affect children's AI-distinguishing accuracy?

RQ.5 Does fluid intelligence or gender moderate the effect of the lesson on visual AI-distinguishing accuracy?

The remainder of this thesis is structured as follows. Chapter 2 discusses the background and related work on artificial intelligence, generative AI, AI literacy, and existing research on distinguishing AI-generated content. Chapter 3 presents the theoretical framework and discusses the relevant literature related to the research questions and hypotheses. Chapter 4 describes the study's method, including the study design, participants, measures, procedures, lesson content, analytical strategy, and ethical considerations. Chapter 5 presents the results. Chapter 6 discusses the findings in relation to existing literature, practical implications, and limitations. Finally, Chapter 7 concludes the thesis by answering the research questions, summarizing the main contributions, and outlining directions for future research.

Chapter 2

Background and related work

2.1 Background

2.1.1 Definition of AI

The European AI Act (European Parliament, 2023) defined an AI system as *"... a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*

Within the scope of this study, this includes generative AI systems such as ChatGPT and Google Gemini, image generators, AI companions such as Snapchat AI, and voice assistants. Those systems process user input and generate new text, speech, images, or responses. Based on this definition, a rule-based program that follows predetermined logic without learning falls outside the scope of AI in this study. We distinguish AI systems, which generate output based on a given input (European Parliament, 2023), and AI-generated artifacts, which are the products of an AI system.

This definition bounds the scope of the systems to which our results are applicable. However, our main focus is not on evaluating whether a given system is an AI system, but rather on children's ability to recognize and critically evaluate AI-generated (virtual) artifacts.

2.1.2 How generative AI systems produce artifacts

Based on the definition of AI systems in the previous section, this section explains how such systems produce outputs. A prominent recent development in artificial intelligence is the emergence of large language models (LLMs). Language models are computational models designed to process and generate human language by estimating the probability of sequences of words or tokens (Chang et al., 2024). Modern large language models are built on transformer architectures, which enable them to learn complex relationships between words and their contexts at scale (Vaswani et al., 2017). Examples of widely used LLM-based AI systems include ChatGPT and Google Gemini.

A second important development in generative AI is diffusion modeling, which is widely used for image generation (Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022). Diffusion models are trained by gradually adding noise to real images until the original image becomes unrecognizable. The model then learns the reverse process: removing a small amount of noise, step by step, to reconstruct a coherent image. After training, the model can start from random noise and gradually turn the noise into a new image that resembles patterns learned from the training data (Ho et al., 2020). Text-to-image AI systems such as DALL-E, Stable Diffusion, and Midjourney combine this denoising process with text-based conditioning, allowing the generated image to be guided by a written prompt (Ramesh et al., 2022; Rombach et al., 2022).

For example, when a user prompts an image generation system with *"an elephant wearing sunglasses surfing on the ocean"*, the model does not retrieve and copy an existing image from its training dataset. Instead, it generates a new image by gradually transforming noise into visual patterns that are statistically

associated with the prompt (Ho et al., 2020; Rombach et al., 2022). In early denoising steps, the model establishes global features such as composition, object placement, and lighting. In later steps, more detailed visual features, such as textures, facial features, and shadows, are introduced (Dhariwal and Nichol, 2021). The generated artifact is therefore the result of a probabilistic generation process rather than direct reproduction of one specific source.

However, this process also explains why AI-generated images can contain visual inconsistencies. Generative models do not possess an understanding of the physical objects or real world; instead, they learn statistical relationships between visual features and textual descriptions from training data (Ramesh et al., 2022). As a result, generated images may appear realistic at first glance while still containing errors in details such as hands, facial features, object boundaries, shadows, spatial relations, or written text (Kamali et al., 2025).

Thus, mistakes in AI-generated artifacts are the consequence of the underlying generative methodology. In an AI system that has a complete understanding of the world and its conventions, it would be impossible to distinguish AI-generated content from human-made content. Therefore, in our study, we focus specifically on AI-generated artifacts that contain subtle but observable errors. These images allow us to examine whether children can identify visual cues that indicate that an artifact was generated by AI. In Section 6.3.3, we discuss the limitations of this focus on distinguishable content.

2.1.3 AI ethics

In addition to the difficulty of detecting AI-generated artifacts, there are also ethical concerns regarding the usage of AI (Safdar et al., 2020). First of all, an AI is trained on training data, which might be biased. For example, a facial recognition system might have lower accuracy on faces with darker skin tones because Western training data consists of more light skin tones. Secondly, the process between input and output is unclear, a black box. There is no insight into why such a generative AI produces a given output for a given input. Finally, there is no guarantee that the output is factually correct; the task of the AI, in principle, is to predict the next word, not give a trustworthy output.

To limit the ethical concerns, the European Union introduced the AI Act (Edwards, 2021). The goal of the AI Act is to set rules for the development and use of AI within the European Union. Each AI system is placed in a risk-based category, with the risk defined as the impact it might have on people's rights. Based on this risk, AI systems could need a code of conduct (minimal risk), providing transparency on how the system works (limited risk, such as chatbots), a conformity assessment (high risk, such as employment software), or be forbidden within the EU (unacceptable risk, such as social scoring). This limits the unforeseen impact that AI (systems) have on society and thus on the lives of people.

The AI Act explicitly mentions children and young teens as vulnerable users in the light of AI: "In addition to those rights, it is important to highlight that children have specific rights as enshrined in Article 24 of the EU Charter and in the United Nations Convention on the Rights of the Child (further elaborated in the UNCRC General Comment No. 25 as regards the digital environment), both of which require consideration of the children's vulnerabilities and provision of such protection and care as necessary for their well-being." (European Parliament, 2023). Thus, the usage of AI by children and teens requires special attention.

This special attention is even more important because most children use generative AI. In the United Kingdom in 2024, 77.1% of 13- to 18-year-olds reported using generative AI in the past year, up from 37.1% in 2023 (Picton and Clark, 2024). In a survey in the United States by Common Sense Media (Robb and Mann, 2025), a nonprofit focusing on kids' online safety, 72% of children aged 13-17 have used an AI companion (a generative AI used as a 'friend') at least once, with 52% of the teens using an AI companion more than once a month.

However, teens do not solely have positive thoughts about those 'AI friends'. In the same survey (Robb and Mann, 2025), 1 in 3 teenagers who used an AI companion felt uncomfortable with something the AI said, and an equal number chose to discuss an important matter with their AI friend rather than with a real person. Some of those even found talking with their AI companion more satisfying than with real people.

This high number of personal interactions, not all of which are good, raises the question of whether children understand how generative AI works. In a study by a child-development-focused non-profit, 4-H (National 4-H Council, n.d.), 66% of questioned American 9-to-17-year-olds expressed some basic

knowledge about generative AI. A similar survey was conducted in the Netherlands (La Fors, 2023), in which 71% of children aged 4-16 had not heard of AI systems before taking the survey. Children who did hear about AI were primarily knowledgeable of specific brands like Siri and Alexa rather than understanding how those systems work. However, this research took place in 2022, so it is likely that more children have learned about AI and how it works since then.

In the same study by 4-H (National 4-H Council, n.d.), 72% of children want adults to be involved in learning with AI in a responsible manner. This raises the question of how education could be shaped to meet children's needs. In the next section, we discuss the concept of AI literacy and the developed frameworks for AI literacy education.

2.1.4 AI literacy

AI literacy is a broad concept that has been defined by various studies. Long and Magerko (2020) define AI literacy as *"a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace"*. Based on a review of 150 documents, they identify 17 competencies organized around five themes: what AI is, what AI can do, how AI works, how AI should be used, and how people perceive AI. The first competency Long and Magerko (2020) mention is the ability to recognize AI, to see if content (or more generally, technical artifacts) is AI-generated or human-made. The authors also discuss children's perception of AI, in which children first notice that an agent behaves in a human-like way and only then recognize that it is a computer. Besides that, children tend to attribute socio-emotional characteristics to AI agents more than adults do.

Ng et al. (2021) build on this concept by reviewing 30 empirical AI literacy studies. They classified all studies into four AI literacy dimensions, based on Bloom's Taxonomy: knowing and understanding AI, using and applying AI, evaluating and creating AI, and engaging with ethical issues related to AI. The majority of the existing research included in the review focused on the dimensions of knowing and using, while the higher-order skills of critically evaluating AI output and engaging with AI ethical issues receive less research coverage. Based on this review, Ng et al. (2021) propose a framework for AI literacy education that describes pedagogical approaches, learning content, and learning artifacts.

At the international level, the European Commission and the Organization for Economic Co-operation and Development (OECD) jointly developed a framework for AI literacy, called the AILit Framework (Organisation for Economic Co-operation and Development (OECD) and European Commission, 2025). This framework is more focused on the future: their definition of AI literacy includes durable skills and future-ready attitudes required to thrive in a world shaped by AI. Based on the competencies they mention, they also suggest ways to develop those competencies through AI education. For example, they state a competence about social and ethical impacts. For primary education, they suggest discussing why a voice assistant sometimes does not understand the command given. For secondary education, they propose a more thorough discussion by investigating how predictive AI is used to calculate credit scores and what the consequences of it are.

A collaboration of Dutch educational institutions also developed their own framework: the AI-GO Framework (Renkema et al., 2025). Existing frameworks state that the different domains (Knowing, using, evaluating, and ethics) are separate and add up to a level of AI literacy, whereas Renkema et al. (2025) found in their review that AI literacy is achieved only through interaction among these domains. Instead, they suggest a model where AI literacy is based on a combination of knowledge, skills, attitudes, and ethics.

These frameworks show that AI literacy is not limited to knowing what AI is, but also how an AI system works, how to use it appropriately, how to critically evaluate AI output, and how to reflect on the ethical and social consequences of AI systems. In the context of this study, the ability to distinguish AI-generated images from real images can therefore be seen as a specific component of AI literacy.

2.1.5 AI-education tools

In the previous section, the different frameworks showed the different competencies children need to achieve AI literacy. However, these frameworks do not yet explain how to develop those competencies in children. In this section, we discuss the existing AI education tools and how they support children in developing AI-generated knowledge, skills, and critical reflection.

Su and Yang (2022) conducted a literature review of AI educational tools in early childhood education (children up to age 8). They found that multiple AI tools and platforms, such as Zhorai and Teachable Machine, are effective for teaching AI and machine learning concepts to children, as well as 'physical' robotic tools like Lego Mindstorms and PopBots.

Zhorai (Lin et al., 2020) is a conversational agent aimed at children to educate them about machine learning concepts. Zhorai is an alien-like monster that responds to questions you ask it, just like Alexa and Siri. However, it is only trained on a small dataset. In the first module, children ask Zhorai questions about ecosystems. In the second module, children will teach Zhorai about animals. In the third module, children can observe machine learning when they ask Zhorai to guess which ecosystem it lives in. Finally, there is an ethical discussion about Zhorai and its learning, for example, about Zhorai knowing the grounded truth and spreading false information.

In a study of children aged 8-11 (Lin et al., 2020), participants reported having a much better understanding of how machine learning works. Besides that, observations of the children showed that they felt bad when Zhorai made a mistake, especially when the conversation was about people and culture. Teaching Zhorai new knowledge not only provides insights into how a conversational agent works, but also helps children understand why it works and what the ethical drawbacks might be.

Another digital tool created for AI education is LearningML (Rodríguez-García et al., 2021). This tool, of Spanish origin, creates insights into how a supervised ML is trained. First, students input training data (training), after which they can train a model (learning). With the trained model, they can input new text or images to test its accuracy (evaluation). The tool is easy to use but also offers the ability to dive deep into machine learning by manually tweaking hyperparameters such as epochs and learning rates. The tool even allows for the trained model to be integrated into a block-based programming tool (Scratch). The Teachable Machine (Carney et al., 2020) is a similar tool developed by Google researchers, but it lacks a Scratch-based building-block environment to play with.

Rodríguez-García et al. (2021) evaluated LearningML among children aged 10-16. They found that using this tool improves knowledge of machine learning through hands-on activities with LearningML, and that 84.4% of participants find LearningML a useful tool for learning about machine learning and AI. Based on a text network analysis, they also found that central words from before the experience are more focused on 'robots', while afterward the words 'problem' and 'solve' become central. Thus, it is an effective way to increase AI literacy, particularly in machine learning fundamentals, among children.

There are also 'physical' tools designed for AI education. A scientifically developed example of this is Popbots (Williams et al., 2019). A Popbot consists of a Lego body with blocks, sensors, and motors, a smartphone head, and a tablet as a programming interface. Using a block-based programming interface, programs could be made for the Popbot to perform. To explain AI, three different programs were made. To explain knowledge-based systems, a rock-paper-scissors game was implemented, in which children both set the rules (e.g., paper wins over rock) and decide what to do (e.g., play a happy sound when winning). For supervised machine learning, a program was developed in which children could classify foods as healthy or unhealthy. Based on k-nearest neighbors, it would give children suggestions on what is healthy and unhealthy, and why. Finally, the researchers created a program in which children provide musical instructions (music emotion, tempo, and chord progression), and the AI remixes the music to match the prompts.

After testing the setups in classrooms, Williams et al. (2019) found that understanding of AI concepts correlated with interactivity with the robot. However, the research had some limitations that limited the transferability to our research case. The research focused on children aged 4 to 8, and the researchers also found that the children's understanding and prior knowledge might have affected their answers. For example, when a classroom had just learned about fruits and seeds, children categorized a tomato in the same class as a banana (seeds inside) rather than the researchers' correct answer as a strawberry (red).

One overarching element in all listed tools is interactivity; children are not only listening to the teacher, but also exploring the subject themselves. Various researchers (Digout and Samra, 2023; Atanasov and Ivanova, 2019) found that interactivity in education enables meaningful learning. Thus, our lesson should include interactive elements to optimally teach the children about AI.

Additionally, research shows (Snow, 1990; Kroon, 2003) that it is important to teach children in their native language, as they learn better when educated in it. Thus, another requirement for our lesson is that it be available in the participants' native language.

2.2 Related work

Previous research has examined the extent to which people can distinguish AI-generated artifacts from human-made artifacts. Chein et al. (2024) investigated whether adults could distinguish AI-generated texts from human-written texts. Participants first completed several psychological measures, including an abbreviated version of Raven's Progressive Matrices, which measures fluid intelligence (see Section 4.1.2). They were then asked to classify general-interest and scientific news articles as either AI-written or human-written. In addition, participants were asked to identify which of two social media comments had been written by a human. The study found that fluid intelligence was significantly related to the ability to distinguish AI-generated content from human content. This suggests that the ability to distinguish AI-generated artifacts from human-made artifacts is affected by general reasoning abilities.

Another study examined people's ability to distinguish fake news from genuine news (Sindermann et al., 2021). Although fake news is not necessarily AI-generated, detecting false or misleading information also requires critical evaluation of the content. Sindermann et al. (2021) investigated how individual characteristics relate to the ability to detect fake news. In their online study, participants evaluated whether the provided news articles were true or false. They found that participants with higher fluid intelligence were less likely to believe fake news. The findings further support the idea that reasoning ability may play an important role in the ability to critically evaluate digital information.

The detection of visual AI-generated content has also been studied, particularly in relation to deepfakes. In a deepfake detection study, participants are typically asked to distinguish real human face images or videos from AI-generated or manipulated ones (Somoray et al., 2025). A literature review by Somoray et al. (2025) showed that human performance in deepfake detection is influenced by individual characteristics, such as cognitive ability and analytical thinking. Beyond deepfakes, Kamali et al. (2025) investigated whether participants could distinguish AI-generated images in a large-scale experiment. Their study showed that detection accuracy not only depends on the observer but also on properties of the image itself, such as scene complexity and visible artifact types. This indicates that distinguishing AI-generated images from human-made images is a complex task shaped by both visual characteristics and individual characteristics.

Together, these studies suggest that the ability to distinguish AI-generated content is related to cognitive abilities and other individual characteristics. However, most existing studies focus on adults, while much less is known about children. This is an important gap, since children are increasingly exposed to AI-generated content (Picton and Clark, 2024; Robb and Mann, 2025) while their critical evaluation skills are still developing (Nikolaeva et al., 2023). In addition, prior studies often examine only one or a limited number of relevant variables at a time. As a result, it remains unclear how factors such as fluid intelligence, AI usage, trust in AI, and gender jointly relate to children's ability to distinguish AI-generated images from human-made images. Addressing this gap is relevant to both research and education because it can provide insights into which children may be more vulnerable to misinterpreting AI-generated visual content and how AI literacy education can be designed to support children's development in the online world.

Chapter 3

Theoretical Framework

This chapter develops the theoretical foundation for understanding children's ability to distinguish AI-generated images from real images. Building on theories of fluid and crystallized intelligence, it examines how reasoning ability (fluid intelligence), trust in AI, prior AI usage, gender, and AI literacy education may influence visual AI-distinguishing ability. Together, these concepts provide the basis for the hypotheses we test in this study and how individual characteristics and educational intervention affect children's ability to distinguish AI-generated visual content from real images. We provide an overview of the theoretical framework in Figure 3.1.

3.1 Fluid intelligence

Cattell (1987) proposed a theory of fluid intelligence (g_f) and crystallized intelligence (g_c). Fluid intelligence is, according to Cattell, the ability to think logically and solve problems without using existing knowledge. This means that education can not improve fluid intelligence. The counterpart of fluid intelligence is crystallized intelligence, the ability to use knowledge and skills from your long-term memory.

Fluid intelligence and crystallized intelligence are factors of general intelligence (Cattell, 1963). The 'fluidity' in fluid intelligence refers to the fact that it applies to all kinds of problems, while the crystallized intelligence is 'crystallized' knowledge only relevant to a certain problem. For example, knowing the capitals of all European countries is a form of crystallized intelligence, while fluid intelligence takes the form of 'having the ability to'.

Later, Horn (1991) expanded this idea of intelligence to include four additional abilities. These abilities include visual perception (g_v), short-term memory (g_{sm}), long-term memory and retrieval (g_{lr}) and processing speed (g_s). Later, Horn added auditory processing to this model, and finally, in 1991, he added quantitative intelligence (g_q, the ability to process statistical data) and reading-writing skills (g_{rw}). Concluding, Horn identified to 8 abilities that describe the intelligence of a person, named the Cattell-Horn g_f-g_c theory.

Carroll (1997) divided these intelligence abilities into three different strata. The previously mentioned (broad) abilities, such as g_f, g_c, and g_v, are in the middle stratum. Under that are narrow abilities, over 70 of which have been identified, grouped together to form a broad ability. At the top level, Carroll suggests a general intelligence g, computed from the middle-stratum-level abilities.

These models were later integrated into the Cattell-Horn-Carroll (CHC) theory (Schneider and McGrew, 2012). Differences between the CHC and the three-stratum theory include slightly different definitions for abilities and other minor adjustments. In this model, the abilities could be grouped into Reasoning, Acquired Knowledge, Memory and Efficiency, Sensory, Motor, and Speed & Efficiency. In this thesis, we examine the influence of intelligence abilities on the ability to distinguish AI artifacts from human artifacts at scale. Testing for all abilities would not be possible in the setting of this research. Therefore, we focus on the abilities of reasoning (g_f, whether the participant is able to reason that something is an AI) and acquired knowledge (i.e., g_c, whether the participant becomes better after acquiring knowledge).

In the following sections, the review of prior research on fluid intelligence examines which variables affect these abilities and how they relate to distinguishing AI-generated artifacts from human-made content.

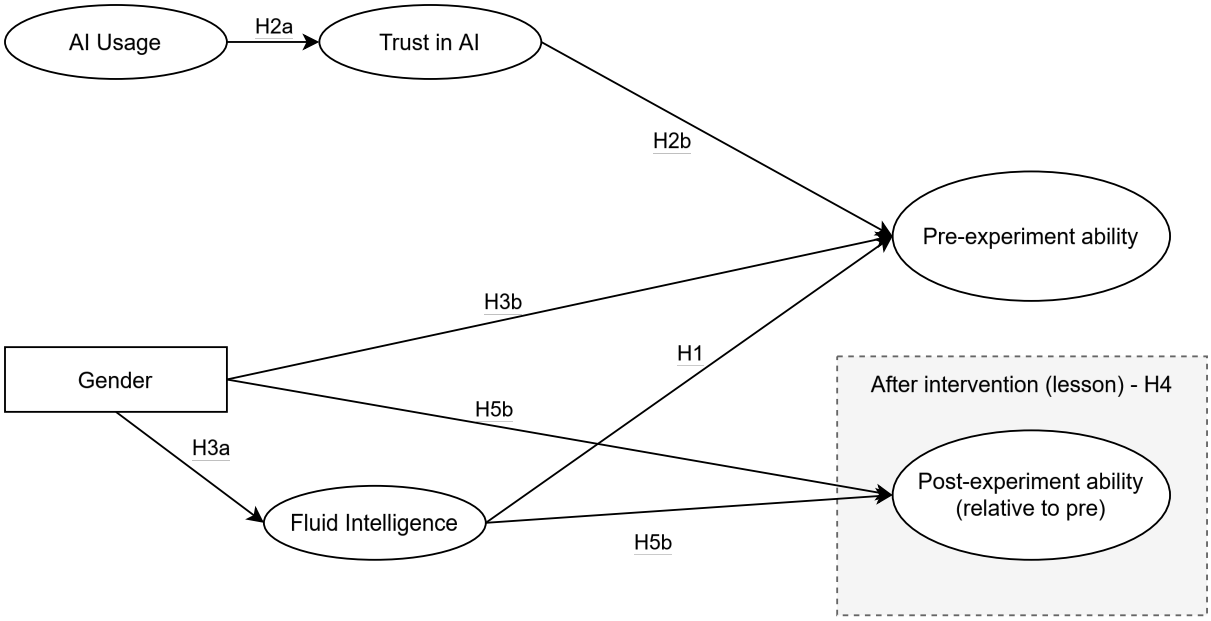


Figure 3.1: Overview of the theoretical framework

3.2 Fluid intelligence & distinguishing AI-generated artifacts

Chein et al. (2024) performed a study where participants between the ages of 18 and 34 were asked first to fill in tests of psychological characteristics, including the (abbreviated) Raven's Progressive Matrices Test. Then, the participants were asked to first categorize both general-interest and scientific news stories as human-written or AI-written. After this, participants were asked to choose between two social media comments to determine which one was written by a human. A statistically significant effect was found between the level of fluid intelligence from the abbreviated RSPM and the accuracy of distinguishing AI-generated content from human-made content.

Another study was about fake news. Sindermann et al. (2021) conducted an online study with 530 participants, in which they were asked how much they believed a news article that was either true or fake. They found that participants with higher fluid intelligence were less likely to believe fake news. While fake news is not necessarily AI-generated, it is an indication of spotting non-real texts.

To date, no study appears to have examined the effect of fluid intelligence on accuracy in visual (i.e., images) AI-distinguishing tasks. A literature review consisting of 40 independent studies (Somoray et al., 2025) showed that human performance in deepfake detection (distinguishing real human faces from AI-generated human-like faces) is influenced by person-level characteristics such as cognitive ability and analytical thinking. Based on the findings of Chein et al. (2024), where the effect of fluid intelligence was proven on textual tasks, and the aforementioned literature review, we formulate the following hypothesis:

H1 Higher levels of fluid intelligence are positively associated with children's accuracy in distinguishing AI-generated images from real images.

3.3 Trust

A meta-analysis by Kaplan et al. (2023) found that several human factors determine trust in AI. As humans gain more understanding of an AI system, trust in AI increases. The same effect is evident in task expertise: when a human has greater expertise in the task, they trust the AI system more. Additionally, gender was found to influence AI trust levels; males tend to trust AI more than females. Other drivers, found in a survey by Alasmari et al. (2025) were usage frequency (using AI more correlates with a higher trust) and age (older respondents had less trust in AI). On the other hand, Alasmari et al. (2025) found no differences in trust between genders. Since Alasmari et al. (2025) found the effect of AI usage among adult participants, we expect that the same effect happens among children, and we formulate the following hypothesis:

H2a AI usage is positively related to children's trust in AI.

In relation to trust in AI, Ladd and Shrivastava (2026) added an extra intelligence factor to the Cattell-Horn-Carroll theory, namely 'Skeptical Intelligence'. This intelligence is an individual's ability to use information (e.g., the output of generative AI) through rational reasoning. Using a survey of 400 respondents, they found that skeptical intelligence enabled the respondents to use AI for innovation. A similar construct is introduced by Shaw and Nave (2026), where "cognitive surrender" by an AI output may override intuitive and analytical thinking. In this study, Shaw and Nave (2026) found that higher trust in AI was associated with greater 'surrender' to AI output and thus a lower likelihood of critically evaluating the outcome.

Based on this less critical evaluation of AI output found by Shaw and Nave (2026), we expect that children will be less accurate in detecting if an image is AI-generated, and thus we formulate the following hypothesis:

H2b Higher levels of trust in AI are negatively associated with children's accuracy in distinguishing AI-generated images from real images.

3.4 Gender, fluid intelligence, and distinguishing AI

Gender is a relevant characteristic to consider in this study, since previous findings on gender differences in fluid intelligence remain mixed. Lynn and Irwing (2004) found a slight advantage for boys at age 15 (5 IQ points) and for boys among children aged 5-11 years (3.2 IQ points) on the Progressive Matrices, a test measuring fluid intelligence.

Colom and Garcıa-López (2002) conducted a study using a representative sample of Spanish high school students (18-23 years) and found that the results varied across different measures of fluid intelligence. In the PMA Inductive Reasoning Test, females outperform males. However, on the Raven Advanced Progressive Matrices (RAPM), males outperform females. In the Culture-Fair Intelligence Test, no differences were found. Colom and Garcıa-López (2002) conclude that there are no sex differences in fluid intelligence. A replication study in Bosnia and Herzegovina found that, while no differences between boys and girls exist between ages 12 and 16, boys aged 17 scored 1 standard deviation higher than girls of the same age (Đapo and Kolenović-Đapo, 2012).

In a meta-analysis of 79 studies by Giofrè et al. (2022) on gender differences in intelligence among school-aged children, there were negligible differences between females and males in fluid intelligence. For crystallized intelligence, there was a notable difference where males outperformed females.

In contrast, Kaufman et al. (2009) did not find a significant difference between men and women in samples ranging in age from 22 to 90 years. However, Kaufman et al. (2009) also found that men outperformed women in writing skills, with the opposite effect in mathematical skills.

Taken together, the literature does not provide consistent evidence for meaningful gender differences in fluid intelligence. Therefore, we expect no significant gender differences in either fluid intelligence. Since we expect that visual AI-distinguishing accuracy partly depends on reasoning ability and fluid intelligence, we also do not expect gender differences in visual AI-distinguishing accuracy.

H3a There is no significant difference between male and female participants in fluid intelligence.

H3b There is no significant difference between male and female participants in visual AI-distinguishing.

3.5 Impact of education

AI literacy education is expected to influence children's ability to distinguish AI-generated images from real ones by increasing crystallized knowledge about AI systems. While fluid intelligence refers to general reasoning ability, crystallized intelligence refers to acquired knowledge and skills. According to this definition, it can be expected that increasing acquired knowledge and skills would lead an individual to perform better on a task that directly involves that knowledge.

Existing research on AI education suggests that children can learn AI and machine learning concepts in age-appropriate ways. For example, the educational tools Zhorai and Teachable Machine have shown to help children understand that AI systems learn from data, that training data can influence model

output, and that AI systems can make mistakes (Lin et al., 2020; Carney et al., 2020). Those insights are relevant to visual AI-distinguishing tasks because recognizing AI-generated content requires children to understand that AI-generated images are not merely inauthentic but artifacts created by systems that learn statistical patterns from data.

AI literacy frameworks also support the expectation that AI literacy education improves AI-distinguishing ability, as they conceptualize AI literacy as more than just knowing what AI is. Long and Magerko (2020) defined AI literacy as *"a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace"*. Following this definition, recognizing AI-generated content falls in the category of critically evaluating AI technologies and is part of the concept of AI literacy. Similarly, Ng et al. (2021) emphasize that evaluating AI and engaging critically with AI artifacts is part of AI literacy. Following the vision of Renkema et al. (2025) that AI literacy does not consist of separate building blocks, but is an interaction of the different domains, we would expect that increasing the level of AI literacy as a whole increases the ability to correctly distinguish AI-generated images from real images:

H4 Receiving AI education results in higher accuracy in distinguishing AI-generated images from real images.

3.5.1 Variables moderating lesson effectiveness

In this theoretical framework, we discussed the effect of various variables on the accuracy prior to the experiment. However, it raises the question of whether those variables also influence the learning rate (i.e., the improvement of performance after the experiment).

In the original work of Cattell (1963), Cattell proposes the 'Investment Theory', which holds that fluid intelligence is invested in developing crystallized intelligence through education and experience. Fluid intelligence is then the 'ability to learn', while crystallized intelligence is the result of the learning ability over time. Thus, a higher level of fluid intelligence would lead to a higher level of crystallized intelligence after learning. This effect was demonstrated by Schweizer and Koch (2002), who found evidence in a sample of 104 participants for a relationship between fluid intelligence and learning. This effect was also observed in a subset of young adults (19- 23-year-olds). Also, on a rule-based learning task (Bui and Birney, 2014) and on the learning of English as a foreign language (Debatin et al., 2019), a relationship between fluid intelligence and learning rate was found. This effect was not only found in adults; it was also observed in a study (Thorsen et al., 2014) that followed over 9,000 children throughout compulsory school in Sweden (ages 6-16), providing support for the Investment Theory. Thus, we draw the following hypothesis:

H5a Fluid intelligence moderates the effect of AI literacy education on children's improvement in visual AI-distinguishing accuracy.

Given the earlier hypothesis that there is no difference in fluid intelligence levels between boys and girls, and our expectation that fluid intelligence and visual AI-distinguishing accuracy are associated, we expect no difference between genders in the increase in task accuracy after the experiment.

H5b Gender does not moderate the effect of AI literacy education on children's improvement in visual AI-distinguishing accuracy.

Chapter 4

Method

This chapter describes the methodological approach used to investigate children's ability to distinguish AI-generated images from real images and to examine the role of fluid intelligence, AI usage, trust in AI, gender, and AI literacy education in this ability. The following sections describe the measures, procedure, participants, analytical strategy, and ethical considerations of this study. In Figure 4.1, we provide an overview of this study's methodology.

4.1 Measures

4.1.1 AI Usage & Trust in AI

To measure AI usage, the most valid approach would be to rely on objective behavioral data, such as participants' actual online activity. However, due to privacy regulations and the nature of this study, collecting such data is neither feasible nor ethically allowed. Consequently, this study uses self-reported measures of AI usage.

AI usage is measured on an ordinal scale, with response categories ranging from never to seldom, monthly, weekly, and daily. This scaling approach captures variation in usage frequency while remaining accessible and understandable for students to report accurately.

To ensure a shared definition of AI, the questionnaire includes concrete examples of AI tools, such as ChatGPT, Google Gemini, the Snapchat AI companion, and voice assistants like Siri, Alexa, and Google Home. By providing these examples, we reduce ambiguity about what constitutes AI usage and increase the validity of this self-reported measurement.

Alongside the construct of AI usage, we also measure trust in AI. Ideally, trust would be assessed using behavioral measures, as they are generally considered more reliable. For instance, participants could be presented with AI-generated outputs, after which their reliance on these outputs is observed. However, due to time constraints in the survey, implementing such behavioral measures is not feasible. Capturing trust behaviorally would require a large number of tasks and observations, which is not feasible within the already short time allocated for the survey. Therefore, we rely on self-reported measures of trust in AI.

To form this construct, we ask 4 Likert-scale items (Q2, Q3, Q4, and Q5, see Table 4.1). These items are adapted from Marimon et al. (2025), Amat-Lefort (2026), and Schmidmaier et al. (2024), in order to

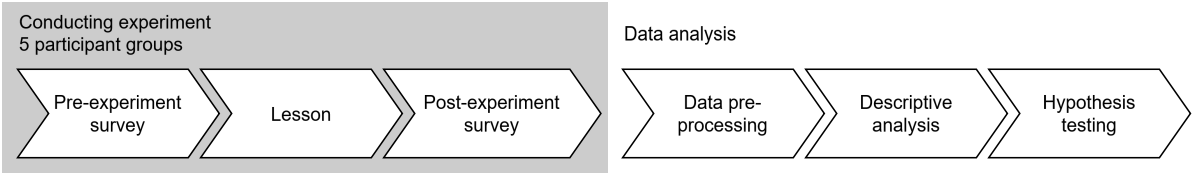


Figure 4.1: Overview of the methodology

	Question	Translation	Source
	Measuring AI usage		
Q1	How often do you use AI?	Hoe vaak gebruik je AI?	Proposed by the authors
	Measuring Trust in AI		
Q2	AI provides credible information.	AI geeft betrouwbare informatie	Adapted from Marimon et al. (2025)
Q3	I trust the information that AI provides.	Ik vertrouw de informatie die AI geeft	Adapted from Marimon et al. (2025)
Q4	I trust that AI is honest and transparent about what it can and cannot do.	Ik vertrouw dat AI eerlijk is en transparant over wat het wel en niet kan	Adapted from Amat-Lefort (2026)
Q5	Overall, I trust AI	Over het algemeen heb ik vertouwen in AI	Adapted from Schmidmaier et al. (2024)

Table 4.1: Ordinal and Likert-scale questions in the survey

minimize measurement bias and increase the validity of our construct. As not all participants are fluent in English, the items were translated into Dutch.

4.1.2 Fluid intelligence

There are various measurements available for the construct of (fluid) intelligence. In this subsection, we will discuss the most popular standardized measurements.

The Wechsler Intelligence Scales for Children (WISC) is a general intelligence test for children, measuring not only fluid and crystallized intelligence but also visual processing, short-term memory, and processing speed (Reynolds and Keith, 2017). The so-called Fluid Reasoning Index (FRI, fluid intelligence component) is derived from various subtests, such as matrix reasoning, in which the child selects the missing picture from a given matrix pattern, and figure weights, in which the child chooses shapes to keep the scale or scales in balance. The test includes both verbal and non-verbal tests (Reynolds and Keith, 2017). Given that our research takes place in the classroom, only non-verbal tests can be used; hence, we do not use this test.

The Woodcock–Johnson Tests of Cognitive Abilities (Schrank and Wendling, 2018) are also a set of general intelligence tests, but they are suitable for both children and adults, unlike the WISC. This test examines multiple factors of ability beyond the fluid and crystallized intelligence in the Cattell-Horn-Carroll theory. Since it assesses more abilities than we consider in our research, we do not use this test.

Finally, the most used test for defining fluid intelligence is the Raven's Progressive Matrices Test (RSPM). In this test, the subject is presented with a pattern, from which a small piece is missing. The subject should then select the correct element that completes the pattern. The original version, developed by Penrose and Raven (1936), consists of 5 sets of 12 questions, with each set increasing in difficulty. It is considered a measure of fluid intelligence (Penrose and Raven, 1936; Bilker et al., 2012).

The downside to all of these measurements is the length of the tests. All of those tests involve numerous questions; for example, the standard Raven's Progressive Matrices test consists of 60 different questions (Penrose and Raven, 1936). This is way too long to incorporate into our study design, especially since it is only an independent variable. Therefore, we searched for shorter tests with proven statistical scalability. Langener et al. (2022) developed a shorter version of the Raven's Progressive Matrices Test, consisting of 15 items selected from the original, with the items selected depending on the age group (9-12 and 13-15). They proved that those 15 questions are a statistically scalable subset of the longer version.

There are other shortened versions of the Raven's Progressive Matrices test, such as a version by Bilker et al. (2012), which consists of even fewer questions. However, those shortened versions have not been validated on children (Irrgang et al., 2019), while the version of Langener et al. (2022) has. Given our study participants' ages, it is beneficial to use a shorter form validated for their age group.

Our lesson is 50 minutes, and 15 questions solely about fluid intelligence would make the survey too long

to allow for effective teaching time. Instead, we used those 15 questions as a question bank, of which 5 questions were randomly selected.

4.1.3 Ability to distinguish AI and real images

To measure participants' AI distinguishing skills, there are multiple options. There is the option to provide two different images and ask them to choose which one is AI-generated (and thus implicitly which one is a real image). Another option is to provide only one image and ask if that image is AI-generated or real, an absolute judgment task. We believe that the most 'true' measurement is of an absolute judgment. In real-life situations, for example, social media or the news, only one post or article is shown, with no comparison possible. This measures whether someone can detect which features in an image are AI artifacts, rather than comparing which image looks the most 'fake'. On the other hand, this might introduce a bias because participants might default to AI or real when they have no clue, rather than selecting the most 'fake' image.

Because this absolute judgment design may allow participants to adopt a response bias toward either AI-generated or real images when uncertain, we not only examined overall accuracy but also distinguished between performance on AI-generated and real images. This allowed us to investigate whether participants were more likely to miss AI-generated images or to incorrectly classify real images as AI-generated, and whether this pattern changed after the lesson.

Further, the images need to be carefully selected to avoid introducing bias. Besides that, images should not be either too easy (obviously AI) or too hard (AI-generated images that are so perfect that they are indistinguishable from real images), so when pointed out, it should be understandable why an image is AI-generated. As far as the authors know, there is no standardized test available for distinguishing AI-generated images from real images. We found an example of a test that includes an after-test explanation of why an image is AI-generated. This test was created by Britannica Education (Britannica Education), a private educational publisher that provides learning resources for schools, students, and libraries. Besides that, they also provide a curated online encyclopedia for both children and adults.

The original version of this test consists of 10 categories, in which the test-taker is prompted to choose which of 2 images is the real image. To make this suitable for our research and to convert it into an absolute judgment, we used the images fromBritannica Education, but displayed them individually; each image was accompanied by two selection boxes to choose between AI-generated and real images. All images are visible in Appendix A

Because of the limited time available, we randomly selected 7 images as questions in the survey before the experiment and 7 images in the survey after the experiment.

Images are never prompted more than once, and within a survey (either before or after), only one image from a given category is chosen for inclusion. There is a chance that a participant was prompted for the other image from a category in the second survey, but given that we never mention categories or that, within a category, there is one real and one AI-generated image, we do not think this has any influence on participants. We did not limit the number of AI-generated or real images in a survey; it is possible that by chance, all images displayed beforehand are AI-generated or real.

4.2 Procedure

4.2.1 Questionnaire

The questionnaire started with a general introduction page. This page consisted of the following:

- An explanation of our research purpose: *This survey is to investigate how well you could distinguish real images and images made by AI.*
- A disclaimer: *It is not a test, and you will not get a grade. Your answer has no effect on the remainder of the workshop. Nobody knows that you gave these answers. If you want, you can stop at any time..*
- An explanation of the workshop: *In this workshop, you learn more about AI, and we see if you become better at distinguishing images.*

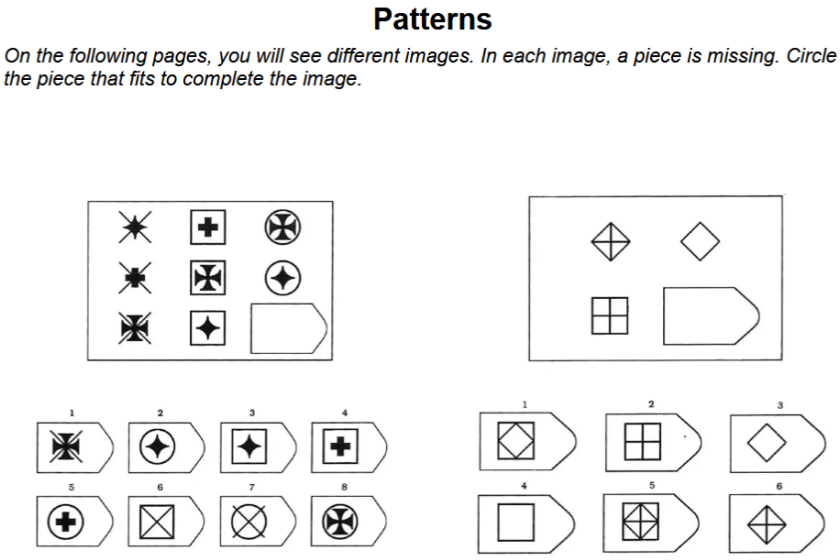


Figure 4.2: Example of a fluid intelligence question from the survey.

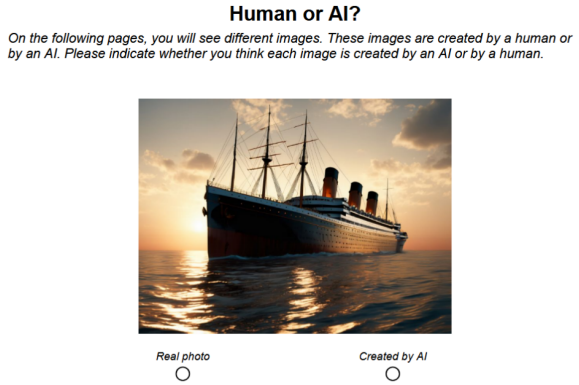


Figure 4.3: Example of AI/real images presentation in the survey.

After the explanation page, we ask participants whether they understand the purpose of our research and how their answers will be used. Then, the rest of the survey follows the measures:

- AI-usage question Q1
- Trust in AI questions Q2-Q6
- The 5 randomly selected fluid intelligence images (See Figure 4.2)
- The 7 randomly selected AI/real images (See Figure 4.3)

The prior survey ended with a box, in which participants were asked to draw a cat or a dog. This functioned as something to do when participants were already finished, while their classmates were not. The drawings were later used in the workshop to explain machine learning and classification models.

To introduce the survey, we quickly reviewed the questionnaire, explaining how the questions should be answered. We also mentioned the points that were on the explanation page.

In the last minutes of our workshop, we performed a second survey. This survey only consisted of the last 7 randomly selected AI/real images. No additional explanation about the survey was given.

Participants received two booklets: one from the prior survey and one from the posterior survey. Those two booklets had the same tag/code to match the prior and posterior surveys. This enabled us to assess



Figure 4.4: Main interface of Zhorai in the first phase: getting to know Zhorai.

4.2.2 Lesson

After the first questionnaire, participants attended a lesson on AI. The goal of the lesson is to increase their AI literacy, including their ability to distinguish AI, how AI and machine learning work, and the ethical constraints of AI.

We started with a simple explanation of what an algorithm is, as algorithms are the building blocks of AI. We illustrated the algorithms by giving all participants a series of steps that turned out to be dance moves to the Macarena. This served both as an entertaining way to introduce the topic and to break the ice for the interaction during the remainder of the lesson.

After this icebreaker, we moved on to the topic of machine learning. For this, we used Zhorai. As explained in Section 2.1.5, Zhorai is an online tool where students can talk to an AI that responds based on the knowledge it has gathered from their input. The learning effects of Zhorai were confirmed in a study by Lin et al. (2020), ensuring that the tool is not only entertaining but also educational.

Since Zhorai was originally developed in English and our participants' mother tongue is Dutch, we created a translated version of Zhorai. We placed the Google Chrome translation API between the model's input and output, changed the speech synthesizer API to Dutch, and translated all prewritten sentences (including the text on the website) into Dutch. This translated version is freely accessible on https://zhorai.nl/. We display the main interface of Zhorai in Figure 4.4.

During the lesson, we mainly followed Zhorai's original lesson plan. However, we incorporated the ethical questions into the tool's use, rather than afterward. For example, we asked participants to tell Zhorai many false facts about cows to show that a machine learning model processes input only and does not know which facts are correct or incorrect. Besides that, it shows that we can easily make Zhorai say false facts, even though Zhorai itself still thinks that the information given is correct.

Then we transitioned from voice/speech to image-based machine learning. For this, we used Google's Teachable Machine (see Section 2.1.5). We pre-trained a classification model from 100 images of cats and 100 images of dogs. In the last question of the survey, participants were asked to draw a cat or a dog. We asked participants to come to the front of the class to show their picture (if they wanted), and we scanned it with a webcam. Then the model identified whether the drawing was a cat or a dog and the percentage of certainty. This was followed by a discussion on how we could improve the model's performance (for example, by training it on drawings of cats and dogs rather than images of real animals).

This introduces how an AI learns to distinguish between cats and dogs. We dived a little deeper, with a brief explanation of how it works for a larger model like ChatGPT (by converting images to bitmaps and learning patterns), and noted that this is very hard for computers to do well. For example, we included Figure 4.5 in our lesson to illustrate that algorithms sometimes struggle to classify between two groups.



Figure 4.5: Sloth or pain au chocolat? By Ellwell1993 on Imgur.

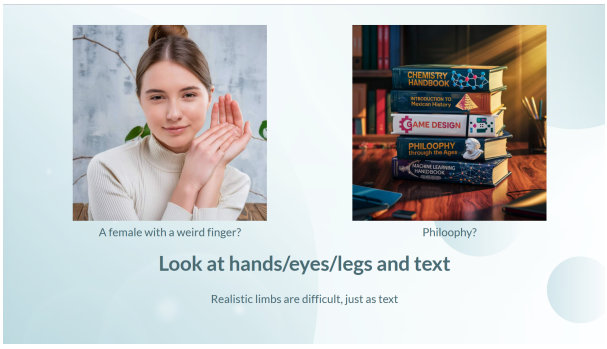


Figure 4.6: Screenshot of slide in our lesson, focusing on practical tips

Finally, we ended our lesson by explaining how they could distinguish AI-generated images from real ones. The tips were adapted from the same source as the AI images used in the survey. It included tips like timespatial awareness, limbs and the face, text, and lighting/shadows. See Figure 4.6 for a screenshot of a slide we used to explain the tips to the students.

After this explanation, the participants filled in the second booklet of the survey. At the end, we discussed the images they had seen to ensure that participants would know afterward which images were AI-generated and which were real.

The lesson slides are displayed in Appendix C, and have been published for use by other teachers via LessonUp, a Dutch platform for teachers.

4.3 Participants

The participants are school-going children. Participants are from various primary schools around Heist-op-den-Berg, Belgium. Those children attended a science day at a secondary school, where various workshops and lessons were held to interest them in STEM education. Our experiment was one of such lessons, given 5 times in different sessions during one day. In Chapter 5, we discuss the participants' demographics.

4.4 Analytical Strategy

The survey is conducted on paper. For each survey, we recorded a specific code. This code was the random key used to generate the chosen fluid intelligence and AI/real images. After we digitized the survey results, we matched the answers to the code and, in turn, to the generated images to automatically mark them as correct or incorrect.

In the survey, the ordinal variables were displayed from left to right (daily/strongly agree to never/strongly disagree). We converted those scores to a score from 1 to 5, where 5 corresponds to daily/strongly agree, to ensure that a higher score means a higher level of usage and trust.

4.4.1 Compensation

For the fluid intelligence task and the AI/real image classification task, items were randomly assigned to participants, meaning that not every participant completed the same set of items. Therefore, using only the proportion of correct answers would not provide a fair estimate of performance, as some participants may have received relatively easier or more difficult items. To account for these differences in item difficulty, we estimated the performance scores using a binomial generalized linear mixed model (GLMM).

A GLMM was appropriate for two main reasons. First, the outcome for each item was a binary variable (correct vs. incorrect), which violates the assumption of linearity in a standard linear regression model. A binomial generalized linear model allows for binary outcome variables. Secondly, responses are enclosed within both participants and items: each participant answered multiple items, and each item was answered by multiple participants. This creates a dependency in the data, as observations are not independent. A mixed model accounts for this clustering of the data by including random effects. Random intercepts for participants account for stable differences in overall ability, while random intercepts for items account for stable differences in item difficulty.

This results in a trained model with an overall model intercept corresponding to the average participant's ability. Besides that, each participant has a random effect (ability score), and each item has a random effect (item difficulty). This is formalized as:

$$logit(P(Y_{pi} = 1)) = \beta_0 + u_p + v_i$$

Where:

- Y_{pi} = response of participant p on item i (correct or incorrect)
- β_0 = overall intercept (average ability)
- u_p = participant random effect (ability score)
- v_i = item random effect (difficulty adjustment)

To obtain the true ability per participant, we added the participant random effect to the overall intercept (ability $\vartheta = \beta_0 + u_p$). This gives an ability score on a log-odds scale, where a score of 0 corresponds to a 50% chance of answering an average item correctly. A negative score corresponds to a lower-than-50% chance, a positive score to a higher-than-50% chance. The exact percentage can be found using the following formula:

$$\hat{p}_i = \frac{1}{1 + e^{-\vartheta}}$$

Where:

- ϑ = participant ability score from the GLMM (intercept + participant random effect)
- e = Euler's number
- $\hat{p}_i$ = expected probability of answering an average-difficulty item correctly.

For the fluid intelligence task, we trained the model using all available responses. This task was completed only before our lesson, so we could include all participant-item observations in a single model.

For the AI/real image classification task, however, participants completed the task twice: once before receiving the AI explanation and once afterward. To obtain a baseline measure of participants' prior classification ability, the model used to derive the ability estimates was trained only on responses from

the pre-explanation survey. This ensured that the estimated scores reflect the performance prior to an intervention or learning effect. Including the posterior responses in the same model would have led to a different baseline, making it impossible to see which improvements are due to our intervention.

For the post-explanation responses, participant scores were obtained by anchoring the item parameters from the pre-explanation model and then estimating participants' ability on that same scale. This means that the item difficulty values derived from the baseline data were seen as fixed reference values, while new participant ability parameters were estimated using the post-explanation responses.

Training a new model on the posterior data would have resulted in both shifts in participant and item difficulties, making direct comparison between the time points difficult. This anchoring of the item difficulty enables us to interpret the difference in prior and posterior scores as a change in participant performance, rather than a redefinition of the scale. Formally, for participant i answering item j after the explanation:

$$logit(P(Y_{pi} = 1)) = \beta_0 + u_p^{post} + v_i^{pre}$$

Where:

- u_p^{post} = estimated post-explanation ability of participant i
- v_i^{pre} = item random effect from the pre-explanation model.

And again, to receive the true ability, we add the fixed intercept to the participant's ability (ability $\vartheta^{post} = \beta_0 + u_p^{post}$).

4.4.2 Hypothesis testing

An exploratory structural equation model was initially considered to investigate the relationships between AI usage, trust in AI, fluid intelligence, and AI-distinguishing accuracy. However, this model was not retained for hypothesis testing because it was not identified. Therefore, the hypotheses were tested using regression-based analyses.

Trust in AI was specified as a latent variable measured by questionnaire items Q2-Q5. AI usage was specified as the answer to Q1, while fluid intelligence and the accuracy scores were represented by the theta score. The model tests whether AI usage predicts trust in AI and whether fluid intelligence, gender, and AI usage predict AI-distinguishing accuracy.

To test hypotheses H1, H2a, and H2b (fluid intelligence, AI trust, and AI usage $\rightarrow$ AI distinguishing ability), we fit a multiple linear regression model, with fluid intelligence, AI trust, and AI usage as independent variables and AI distinguishing ability as the dependent variable.

For hypothesis H2a (AI usage $\rightarrow$ trust in AI), we conduct a Spearman rank-order correlation. Since both AI usage and trust in AI are ordinal variables, we test whether a higher AI usage rank is associated with a higher trust-in-AI rank. To confirm those results, we also conducted a linear regression model with AI trust as the dependent variable and AI usage as the predictor. To see if there is a trend, rather than large differences between the highest and lowest ranks, we also conduct a linear regression with AI usage as a categorical predictor (with a unique term for each rank).

The effects of gender are tested using a Two-One-Sided Test (TOST) equivalence test. Our hypotheses state that there will be no differences in variables between males and females. A t-test to compare groups would only show that there is no significant difference; it does not prove that there is no difference. Schuirmann (1987) developed TOST, in which a bound is specified. This bound is often defined by Cohen's d, which expresses it as a multiple of the standard deviation ($d = 0.5$ means the bound is $pm0.5\sigma$). This particular bound is used to test two sub-hypotheses, one-sided tests $\Lambda \leq -d$ and $\Lambda \geq d$. If both sub-hypotheses can be rejected, we can claim that within the bounds, the differences between the two groups are equivalent. We test the hypotheses of no difference between gender groups in fluid intelligence and pre-experiment AI accuracy, with Cohen's d = 0.5.

To test if there is a difference between pre- and post-experiment performance, we conduct a paired-samples t-test. Since we know each participant's score before and after, we can use these pairs in our analysis.

To test the moderating effects of fluid intelligence and gender in the difference between pre- and post-experiment accuracy, we fitted two separate linear mixed-effects models. All data is converted to a long

format; instead of two columns pre_score and $post_score$ in the wide data format, each participant has two rows with an entry for $time$ (pre/post) and score, as well as all other variables. This allows us to use the repeated measures in our mixed model.

Since we have scores for all participants pre- and post-experiment, we use the post-experiment score as the dependent variable. In our model to test fluid intelligence, we include the interaction term $time \cdot fluid_intelligence$; to test for gender, we include the interaction term $time \cdot gender$. Both models include the control variables AI usage and trust in AI, and the latter also includes fluid intelligence. Since we use repeated measures for each participant, we add the random effect $(1 \mid participant_key)$ to model within-participant dependence. In an ideal situation, we would have used a random term $(1 + time \mid participant_key)$ to account for differences in learning rate across participants, but given the size of our sample, we do not have enough data for the linear mixed model to converge.

4.5 Ethics

Our experiment is anonymous and does not harm our participants in any way. The researchers do not know the participants themselves. The results of the questionnaires are assessed only after all data have been collected, never during the experiment itself. Answers to the questionnaire do not affect the treatment in the experiment, nor do they affect grades. Participants were free to stop at any time, without consequences. Participants received a Leiden University stress ball as a gift after their participation, regardless of their performance/commitment. The researchers did not receive any compensation from the schools for giving the lesson.

Ethical approval is pending from the ethics committee of the Leiden Institute for Advanced Computer Science, Leiden University.

Chapter 5

Results

On the 9th of February, we conducted the experiment. Divided into 5 groups, 78 participants attended the workshop (participants per group: 10, 16, 19, 21, 12). All participants were from the 5th year of primary education in Belgium; whilst participants' ages were not asked for, students in the 5th year are usually around 10-11 years old.

Data from 12 students were excluded because they either did not answer all questions, provided more than one answer to a question, or did not answer in the prescribed format. One student was a non-native Dutch/Flemish speaker and received help. It is unclear whether they understood the questions correctly; thus, their data was also excluded. This resulted in the data of 65 students being used in the analysis.

Of those 65 students, 32 said they are male; 33 said they are female.

5.1 Descriptive statistics

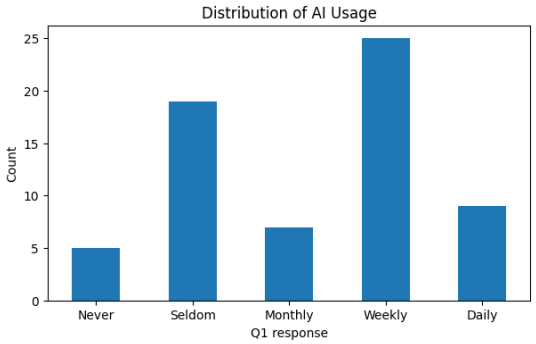
The experiment started with a survey. In Table 5.1, we show the mean and variance of the ordinal questions asked, as described in Figure 5.1. Note that the higher the number, the higher the AI usage and the higher the trust in AI. In Figure 5.1a, we show the distribution of the AI usage score. Almost all students have used AI, with over 50% using it at least weekly. Looking at Figure 5.1b, we see a difference between genders: 65.6% of male students reported using AI at least weekly, compared to 39.4% of female students. The averages are also different (Male: $\mu = 3.50, \sigma = 1.14, n = 32$; Female: $\mu = 2.94, \sigma = 1.27, n = 33$).

The other ordinal questions measured trust in AI. In Figure 5.2, we show the distribution of answers across all trust questions. We see that on almost all questions, the participants answered neutral. For Q4 and Q5, we see that most students agreed or strongly agreed with the questions' statements. There were no major differences in answers between genders.

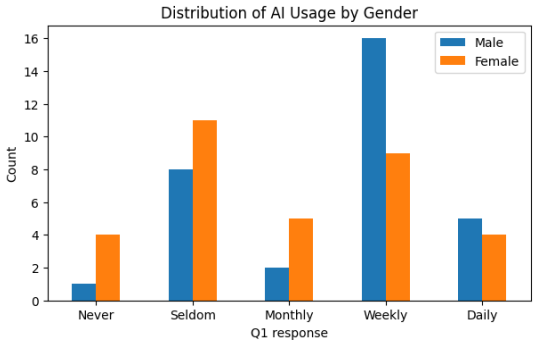
These four questions form our 'Trust in AI' scale ($\alpha = 0.757$). By calculating the average for each student, we got the resulting scores, as displayed in Table 5.1. There were slight differences between male ($\mu = 3.406, \sigma = 0.564, n = 32$) and female ($\mu = 3.136, \sigma = 0.583, n = 33$) trust scores.

	Question	Mean μ	SD σ
Q1	"How much do you use AI?"	3.215	1.231
↪	AI Usage Score	"	"
Q2	"AI gives trustful information"	3.077	0.669
Q3	"I trust the information AI gives"	3.185	0.788
Q4	"I trust that AI is fair and transparent about what it can and cannot do"	3.446	0.771
Q5	"In general, I trust AI"	3.369	0.840
↪	AI Trust Score	3.269	0.585

Table 5.1: Mean and variance of ordinal questions

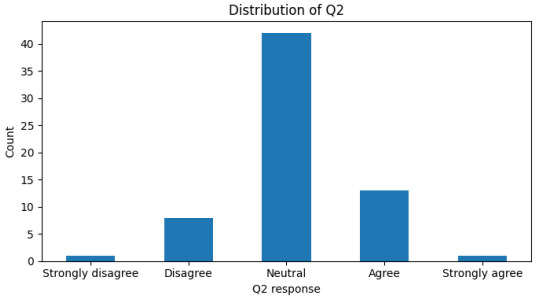


(a) Total distribution

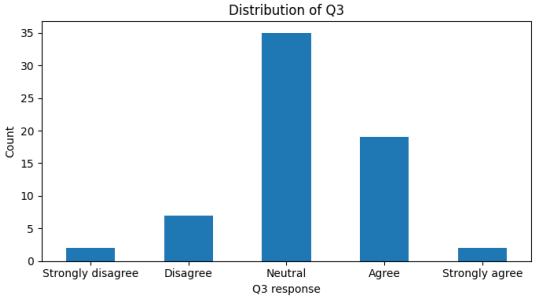


(b) Distribution split over gender

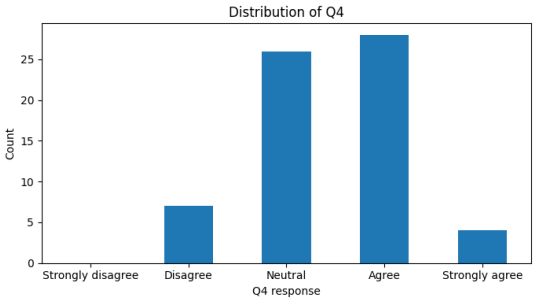
Figure 5.1: Distribution of AI Usage



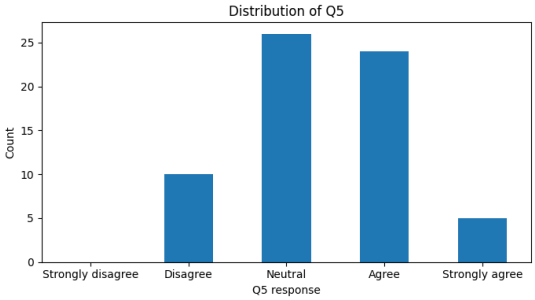
(a) Q2: "AI gives credible information"



(b) Q3: "I trust the information AI gives"



(c) Q4: "I trust that AI is fair and transparent about what it can and cannot do"

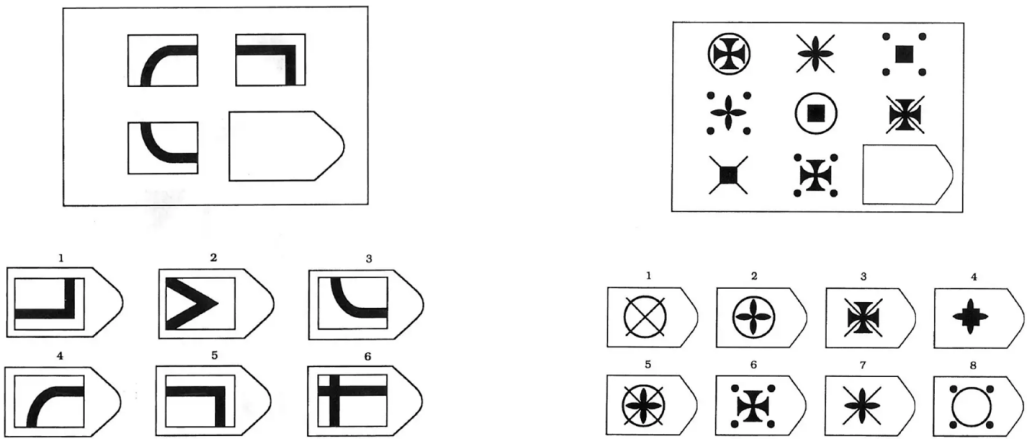


(d) Q5: "In general, I trust AI"

Figure 5.2: Distribution of trust question answers

	n	% Correct
A12	20	60.00%
B5	26	61.54%
B9	17	41.18%
B10	28	92.86%
B11	23	82.61%
C8	23	56.52%
C9	23	56.52%
D4	24	37.50%
D5	22	77.27%
D6	22	68.18%
D7	20	45.00%
D9	17	70.59%
D10	24	54.17%
E1	21	61.90%
E5	15	46.67%
Total	325	61.85%

Table 5.2: Overview of all Fluid Intelligence questions



(a) B5: The easiest question

(b) D10: The hardest question

Figure 5.3: Easiest and hardest Fluid Intelligence questions

5.1.1 Fluid Intelligence

After the ordinal questions, participants were asked to correctly identify the missing pattern. From a pool of 15 questions, participants were randomly asked to answer 5. In Table 5.2, the percentage of students who chose the correct option for each question is shown. The hardest question was D10 (Figure 5.3b, while the easiest question for our participants was B5 (Figure 5.3a. When looking at individual students, most got the majority of the questions correct ($\mu = 3.092, \sigma = 1.476$). The detailed distribution is displayed in Figure 5.4.

Since not all participants answered the same questions, and some questions were harder than others, we need to adjust participants' scores based on the difficulty of the questions they answered. As explained in Section 4.4.1, we estimated a binomial generalized mixed model (GLMM) to account for variability across students and questions.

The model's intercept was positive (posterior $\mu = 0.648$, posterior $\sigma = 0.0.135$, indicating that, on average, participants performed better than randomly guessing the correct answer. Variability among students was substantial, with an estimated standard deviation of 1.35 (95% CI $[1.13, 1.60]$), suggesting notable differences in the ability to correctly answer fluid intelligence questions. For the questions themselves, variability had an estimated standard deviation of 0.82 (95% CI $[0.57, 1.18]$). Thus, the differences between participants are larger than the differences between questions.

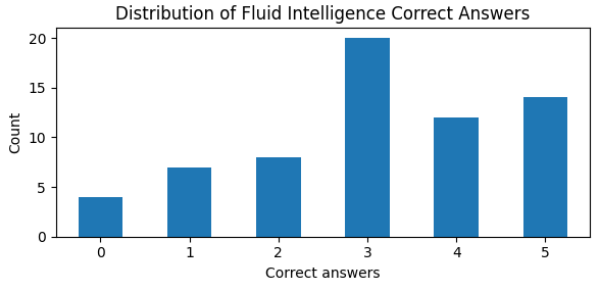


Figure 5.4: Distribution of correct Fluid Intelligence questions

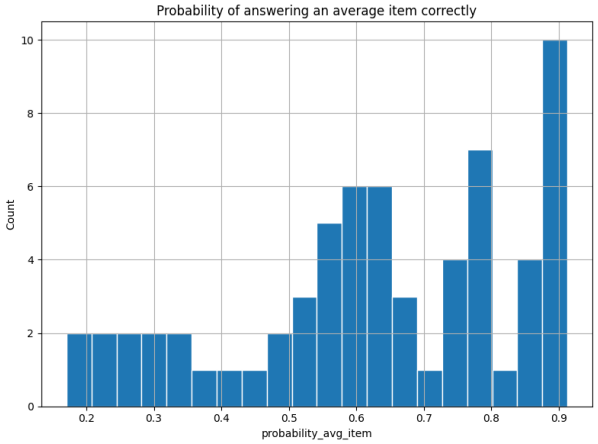


Figure 5.5: Probability of correctly answering an average item

With this model fitted, we extracted the model's random effect for each participant, added it to the fixed intercept, and calculated their estimated fluid intelligence. This resulted in an average fluid intelligence ability of $\mu = 0.652$ ($\sigma = 1.042$), where the ability is on a logit scale. When converted to a linear scale, we receive the probability that the student answered an average item correctly. The average student had a probability of $\mu = 0.631$ ($\sigma = 0.221$) of answering an average item correctly. In Figure 5.5, we display a histogram of the probabilities. Differences between male and female students on this probability were negligible (male: $\mu = 0.634$, female: $\mu = 0.629$).

5.1.2 AI distinguishing

After the fluid intelligence task, students were given 7 images and asked to determine whether each was AI-generated or real. After the lesson, they were again asked to classify 7 (different) images. In Table 5.3, we present a table of all images, showing the percentage of students who correctly classified each image. In total, 48.46% of the images were correctly classified, making overall performance worse than guessing. When splitting between AI-generated and real images, we can see that participants correctly identified real images about 18 pp better than AI-generated images, before, after, and in total.

When looking at individual images, we can see that participants struggled the most with the AI-generated images of Mount Fuji, a pangolin, and the Titanic (all total scores $< 20\%$). The easiest-to-distinguish AI-generated images were images of papyrus (80.00%), women protesting for suffrage (62.22%), and the Northern Lights (61.36%).

For real images, we can see that the hardest images to classify were images of Mount Fuji (total 34.04%), a pangolin (36.00%), a space shuttle, and the Titanic (both 40.91%). The easier images were of a group of children (90.70%), papyrus (85.37%), and women protesting for suffrage (76.00%).

Comparing the scores of the real and AI-generated images within a theme of images, we see that students structurally performed worse than guessing on the themes of Mount Fuji/Significant Landmarks (AI: 16.67%, Real: 34.04%), Pangolin/Endangered Species (AI: 12.77%, Real: 36.00%), the Space Shuttle/Space Exploration (AI: 16.67%, Real: 34.04%) and the Titanic/Historical Events (AI: 13.33%, Real:

	n Before	% Correct	n After	% Correct	pp Difference	n Total	% Correct
City-AI	14	28.57%	28	32.14%	3.57 pp	42	30.95%
City-Real	27	59.26%	20	50.00%	-9.26 pp	47	55.32%
Fuji-AI	28	7.14%	20	30.00%	22.86 pp	48	16.67%
Fuji-Real	24	41.67%	23	26.09%	-15.58 pp	47	34.04%
Lights-AI	19	73.68%	25	52.00%	-21.68 pp	44	61.36%
Lights-Real	24	54.17%	24	58.33%	4.17 pp	48	56.25%
Pangolin-AI	21	4.76%	26	19.23%	14.47 pp	47	12.77%
Pangolin-Real	24	50.00%	26	23.08%	-26.92 pp	50	36.00%
Papyrus-AI	23	91.30%	22	68.18%	-23.12 pp	45	80.00%
Papyrus-Real	21	76.19%	20	95.00%	18.81 pp	41	85.37%
People-AI	31	48.39%	19	26.32%	-6.28 pp	42	30.95%
People-Real	18	88.89%	25	92.00%	3.11 pp	43	90.70%
Shuttle-AI	23	34.78%	19	26.32%	-8.47 pp	42	30.95%
Shuttle-Real	19	42.11%	25	40.00%	-2.11 pp	44	40.91%
Suffrage-AI	22	68.18%	23	56.52%	-11.66 pp	45	62.22%
Suffrage-Real	25	84.00%	25	68.00%	-16.00 pp	50	76.00%
Titanic-AI	20	20.00%	25	8.00%	-12.00 pp	45	13.33%
Titanic-Real	24	54.17%	20	25.00%	-29.17 pp	44	40.91%
Wild-AI	25	48.00%	18	27.78%	-20.22 pp	43	39.53%
Wild-Real	23	56.52%	22	72.73%	-5.93 pp	45	64.44%
Total	455	51.43%	455	45.49%	-6.48 pp	910	48.46%
AI Images	226	42.48%	225	36.00%	-5.48 pp	451	39.25%
Real images	229	60.26%	230	54.78%	-5.93 pp	459	57.52%

Table 5.3: Overview of all answers on the AI distinguishing task

40.91%).

On the other hand, students performed better than guessing on the themes of Northern Lights/Natural Phenomena (AI: 61.36%, Real: 56.25%), Papyrus/World History (AI: 80.00%, Real: 85.37%), and Suffrage/Influential Events (AI: 62.22%, Real: 76.00%).

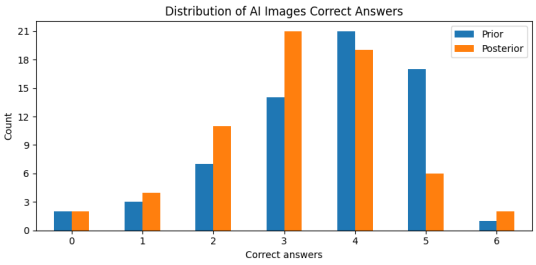
For the theme of the group of children/People, differences between AI-generated and real images were the largest (AI: 30.95%, Real: 90.70%), indicating that they viewed both images as containing real people.

When comparing before and after the lesson, we see that the percentage of correctly distinguished images is lower after the lesson than before, both overall and across all subgroups. When looking at individual images, we can see that only the AI-generated image of Mount Fuji (22.86 pp), the AI-generated image of a pangolin (14.47 pp), and a real image of a piece of papyrus were classified better. There were no themes where, after the lesson, both AI and real images were classified better.

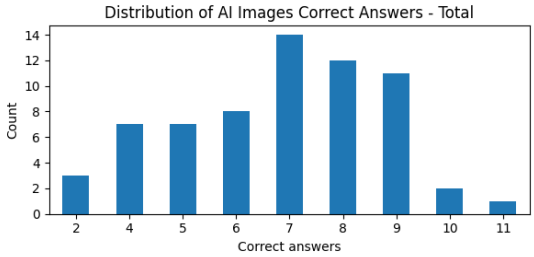
Looking at student-level, most students got 3, 4 or 5 answers correct prior to the experiment ($\mu = 3.600$, $\sigma = 1.321$, see Figure 5.6a), while after the experiment, the majority only had 3 or 4 questions correct ($\mu = 3.185, \sigma = 1.273$, see 5.6a). In total, students correctly distinguished on average $\mu = 6.785$ ($\sigma = 2.019$) images (see Figure 5.6).

Again, since not all students answered the same questions, we need to adjust for question difficulty. We estimated a binomial generalized mixed model (GLMM) to model the differences in variability across students and images. To estimate the model, we used only the data from the images that were asked in the survey before the experiment. This resulted in a small positive intercept (posterior $\mu = 0.064$, posterior $\sigma = 0.1056$), indicating that, on average, students gave slightly better answers than if they had been randomly guessing. Variability between students differed, with an estimated standard deviation of 0.404 (95% CI $[0.339, 0.481]$). Between images, variability had an estimated standard deviation of 1.182 (95% CI $[0.866, 1.613]$).

Using this fitted model, we extracted the model's random effect for each student and calculated their estimated prior AI-distinguishing ability. This resulted in an average prior ability of $\mu = 0.064$ ($\sigma = 0.165$), where the ability is on a logit scale. Converted to a linear scale, we receive the probability that



(a) Correctly answered AI questions prior and posterior to the experiment



(b) Total correctly answered AI questions

Figure 5.6: Distribution of correctly answered distinguishing questions

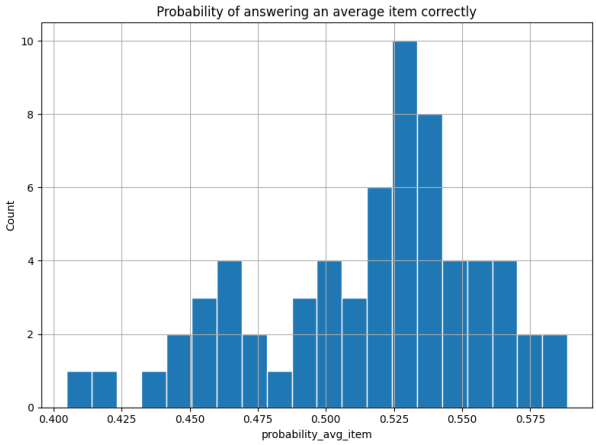


Figure 5.7: Probability of correctly answering an average item prior to the experiment

the students classified an average item prior to the experiment correctly. The average student had a probability of $\mu = 0.516$ ($\sigma = 0.041$) of answering an average item correctly prior to the experiment (See Figure 5.7).

From the model, we also received the difficulty per item. Using these difficulty scores, we calculated the posterior scores per person. This resulted in an average posterior ability of $\mu = -0.357$ ($\sigma = 1.368$), where the ability is on a logit scale. Converting to a linear scale again yields the probability of correctly answering an average item. An average student had a probability of $\mu = 0.445$ ($\sigma = 0.206$) of answering an average item posterior to the experiment correctly (see Figure 5.8). In Table 5.4, we specified the probabilities of correctly answering an average item before and after the explanation for each experimental group.

Group	P(X=correct) before	P(X=correct) after	Change
A	0.502	0.423	-7.90pp (-15.73%)
B	0.507	0.409	-9.74pp (-19.22%)
C	0.515	0.437	-7.81pp (-15.17%)
D	0.529	0.457	-7.44pp (-14.09%)
E	0.522	0.509	-1.37pp (-2.62%)

Table 5.4: Probability of an average item answered correctly, per experiment group

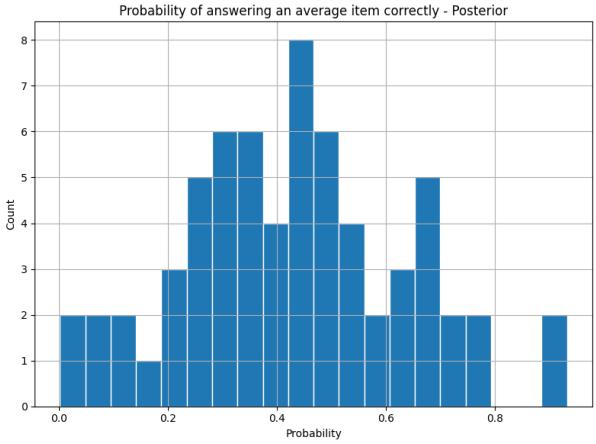


Figure 5.8: Probability of correctly answering an average item posterior to the experiment

5.2 Hypothesis Testing

As an exploratory analysis, a structural equation model was estimated. However, the model was not identified, as indicated by negative degrees of freedom ($df = -2$) and the inability to compute standard errors. Therefore, the SEM results were not interpreted and are not reported as evidence for the hypotheses. Instead, the hypotheses were tested using regression-based analyses.

5.2.1 Pre-experiment performance

Hypothesis 1 stated that participants with higher fluid intelligence would demonstrate greater accuracy in distinguishing visual AI.

To test this hypothesis, a multiple linear regression analysis was conducted, with pre-experiment AI-distinguishing ability as the dependent variable and fluid intelligence, trust in AI, and AI usage as predictors. The overall model was not statistically significant, $F(3, 61) = 1.702, p = .176$, and explained 7.7% of the variance in pre-experiment performance ($R^2 = .077$, adjusted $R^2 = .032$).

In this model, Fluid Intelligence was not a significant predictor of AI-distinguishing ability when controlling for trust in AI and AI usage, $b = 0.030, SE = 0.020, t(61) = 1.53, p = .131$. Although the regression coefficient was positive, indicating an association between higher fluid intelligence and higher estimated task performance, this effect did not reach statistical significance.

Trust in AI (H2b) was also not a significant predictor of performance, $b = -0.010, SE = 0.036, t(61) = -0.27, p = .792$, suggesting that there is no meaningful relationship between trust in AI and pre-experiment AI-distinguishing ability.

AI usage showed a negative but non-significant trend, $b = -0.030, SE = 0.017, t(61) = -1.71, p = .092$. This indicates that participants who reported more frequent AI use have a lower suggested ability to distinguish AI. However, there is no statistical significance.

5.2.2 AI usage and trust

To test if AI usage predicted trust in AI (H2a), a Spearman rank-order correlation was conducted. The analysis showed a significant positive association between AI usage and trust, $\rho = .316, p = .010$. This indicates that participants with higher self-reported AI usage tended to report higher levels of trust in AI.

A supplementary regression analysis was conducted with trust as the dependent variable and AI usage as the predictor. The model was statistically significant, $F(1, 63) = 6.64, p = .012$, and explained 9.5% of the variance in trust ($R^2 = .095$, adjusted $R^2 = .081$). AI usage was a significant positive predictor of trust, $b = 0.147, SE = 0.057, t = 2.58, p = .012$, showing a similar positive effect.

When AI usage was entered as a categorical predictor, the overall model was not statistically significant, $F(1, 63) = 6.64, p = .012$, although the highest usage category showed a tendency toward higher trust.

This suggests that the relationship is better described as a general upward trend rather than by large differences between individual categories.

5.2.3 Effects of gender

To test if gender has an influence on the level of fluid intelligence (H3a), we conducted an independent-samples Welch's t-test. The test showed that fluid intelligence scores did not differ significantly between male participants and female participants, $t(62.90) = 0.17, p = .886, 95\%$CI [-0.56,0.48].

Because our hypothesis stated that there would be no difference, we conducted a Two One-Sided Tests (TOST) equivalent test with the equivalence bounds of Cohen's $d = \pm 0.50$. The equivalence test was significant, $t(62.90) = -1.85, p = .035$, indicating that any gender difference was statistically equivalent to zero within the predefined bounds.

An independent-samples Welch's t-test showed that pre-experiment AI-distinguishing accuracy did not differ significantly between male and female participants, $t(60.82) = -1.28, p = .204, 95\%$CI [-0.134,0.029].

To test hypothesis H3b, which stated that there would be no gender difference in accuracy, we conducted a Two One-Sided Tests (TOST) equivalence test with equivalence bounds of Cohen's $d = \pm 0.50$. The equivalence test was not significant, $t(60.82) = -0.73, p = .234$. Therefore, we could not conclude statistical equivalence.

These results indicate that, although no significant gender difference was detected, the data were insufficient to show that male and female students perform equivalently within the predefined bounds. Therefore, Hypothesis H3b was not supported.

5.2.4 Post-experiment performance

To examine if students' post-experiment AI-distinguishing ability scores differed from their pre-test ability scores, we conducted a paired-samples t-test. Results showed a significant difference between scores before and after the lesson, $t(64) = -2.5424, p = .013$, 95% CI [-0.75, -0.09]. The mean difference was -0.42, indicating that post-experiment AI-distinguishing ability scores were, on average, 0.42 points lower than pre-experiment scores.

Given that our hypothesis expected a higher score, we cannot support H4. However, we conclude that there is a significant effect, but in the reverse direction.

We conducted a linear mixed-effects model to investigate if fluid intelligence moderated the changes in AI-distinguishing accuracy (H5a), while accounting for repeated measurements within participants by including a random intercept for participant (*key*). Trust in AI and AI usage were included as control variables. The random intercept variance for participants was relatively small ($\sigma = 0.184$), suggesting that there is limited between-person variability beyond the fixed effects.

The interaction between time (before or after the experiment) and fluid intelligence was not significant, $b = -0.028, SE = 0.162, t = -0.17$, indicating that fluid intelligence did not moderate changes in performance over time. Thus, we reject H5a.

We fitted another linear mixed-effects model with gender as a moderating factor of the changes in accuracy (H5b), while still including a random intercept for participants. The interaction between time and gender was not significant, $b = 0.360, SE = 0.331, t = 1.09$. The 95 % CI [-0.287, 1.008] is wide, suggesting that there might be meaningful effects. Thus, we see no significant effect, but we cannot confirm that there is none.

5.2.5 Summary of results

In Table 5.5, we display a summary of our hypotheses, tests, main findings, and conclusions.

	Expected effect	Test	Main finding	Conclusion
H1	Higher fluid intelligence → higher pre-experiment AI-distinguishing accuracy	Multiple linear regression	Fluid intelligence did not significantly predict pre-experiment accuracy	Not supported
H2a	Higher AI usage → higher trust in AI	Spearman correlation	AI usage was positively related to trust in AI	Supported
H2b	Higher trust in AI → lower pre-experiment AI-distinguishing accuracy	Multiple linear regression	Trust in AI did not significantly predict accuracy	Not supported
H3a	No gender difference in fluid intelligence	TOST equivalence test	Gender groups were equivalent within the predefined bounds	Supported
H3b	No gender difference in AI-distinguishing accuracy	TOST equivalence test	Equivalence could not be confirmed	Not supported / Inconclusive
H4	Post-experiment accuracy > pre-experiment accuracy	Paired-samples t-test	Post-experiment accuracy was significantly lower than pre-experiment accuracy	Opposite effect found
H5a	Fluid intelligence moderates improvement after the lesson	Linear mixed-effects model	No significant moderation effect was found.	Not supported
H5b	Gender moderates improvement after the lesson	Linear mixed-effects model	No significant moderation effect was found, but absence of an effect could not be confirmed	Not supported / Inconclusive

Table 5.5: Summary of hypothesis results

Chapter 6

Discussion

This study investigated children's ability to distinguish AI-generated images from real images and examined whether this ability was related to fluid intelligence, AI usage, trust in AI, gender, and participation in an educational lesson about AI. The study was conducted among 65 students from the fifth year of Belgian primary education. Participants completed a pre-test, consisting of questions about AI usage, trust in AI, fluid intelligence, and visual AI-distinguishing ability. They then followed a lesson about AI and machine learning, after which their visual AI-distinguishing ability was measured again.

The results showed that children's performance on the AI-distinguishing task was relatively low. For all image classification responses, the participant correctly classified only 48.46% of the images, slightly below chance. Participants were better at correctly identifying real images (57.52%) than AI-generated images (39.25%). This indicates that participants perceived most images as real.

Contrary to our expectations, participants performed significantly worse after an AI literacy lesson. Before the lesson, participants answered 51.43% of all questions correctly; afterward, only 45.49% were answered correctly. Participants made more mistakes both in recognizing AI-generated images (42.48% before, 36.00% afterward) and in recognizing real images(60.26% before, 54.78% afterward).

We found a significant relationship between AI usage and trust in AI: children who reported using AI more frequently also reported higher trust in AI. We also found no significant gender differences in the level of fluid intelligence, and an equivalent test supported the conclusion that boys and girls are statistically equivalent within the predefined bounds.

Our findings also confirm the existing literature indicating that gender does not affect fluid intelligence. After conducting a TOTS equivalence test, we found no difference between male and female participants in fluid intelligence.

We did not have any statistical evidence of a relationship among fluid intelligence, trust in AI, AI usage, and AI-distinguishing ability. For AI-distinguishing ability, we did not find a significant gender difference, but we could not confirm equivalence between the two genders; thus, we do not have a confirmed 'no-effect'. We draw the same conclusion for gender as a moderator of the change in AI-distinguishing ability: we did not find a significant difference between gender groups, but we cannot confirm equivalence. Finally, fluid intelligence did not significantly moderate the change in AI distinguishing ability.

6.1 Contributions to existing literature

Our study contributes to the growing body of literature on AI literacy among children. Previous research has shown that children increasingly use generative AI (Picton and Clark, 2024), also in the form of an AI companion (Robb and Mann, 2025). However, their views AI companions are not entirely positive (Robb and Mann, 2025), as 1 in 3 teenagers felt uncomfortable with AI output. Existing work on AI literacy focuses mainly on understanding basic machine learning concepts and reflecting on ethical issues (Lin et al., 2020; Carney et al., 2020). Our study adds to that literature by focusing on visual AI-distinguishing ability: the ability to identify whether an image is real or AI-generated.

Additionally, our study found that children struggle substantially to distinguish AI-generated images

from real ones, with overall performance below chance. This finding suggests that children are not yet equipped to recognize AI-generated visual content, confirming the European Union's view that children and young teens are vulnerable users of AI systems (European Parliament, 2023).

We did not find support for a relationship between fluid intelligence and visual AI-distinguishing ability. Chein et al. (2024) found that a higher level of fluid intelligence increases textual AI-distinguishing ability, and Somoray et al. (2025) found that personal characteristics influenced deepfake detection. Based on these studies, we expected that children with higher levels of fluid intelligence would better distinguish AI-generated visual content. One possible explanation is that visual distinguishing requires more specific knowledge than general reasoning ability. For example, recognizing unrealistic shadows, distorted text, and incorrect body parts might depend more on specific knowledge about AI-generated images than on abstract reasoning and thus on fluid intelligence.

We confirmed the effect reported by Alasmari et al. (2025) that AI usage and trust in AI are positively related among children. However, we did not have enough data to research the direction of causality in this relationship. Both directions are possible: children who use AI more might trust it more because they use it, but children who trust AI more could also be more willing to use it. Nevertheless, our findings suggest that frequent AI usage and trust in AI are related. Given that we did not find statistical evidence for further relationships between AI usage and trust in AI on one hand and fluid intelligence and AI distinguishing ability on the other hand, we cannot say whether positive effects, such as familiarity supporting more confident and effective use, or negative effects, such as higher trust making children less critical of AI-generated outputs, are dominant. Given that we do not find any further relationships, our study does not prove or disprove the suggested cognitive surrender by Shaw and Nave (2026).

One of the main contributions of this study is its challenge to the assumption that AI literacy education automatically improves children's ability to distinguish AI-generated images from real ones. Previous work on educational tools such as Zhorai (Lin et al., 2020), LearningML (Rodríguez-García et al., 2021), and PopBots (Williams et al., 2019) has shown that using these tools increases knowledge of AI and machine learning. However, our findings suggest that conceptual understanding of how an AI system works does not necessarily improve performance in distinguishing AI-generated images from real images. In fact, participants performed significantly worse after the lesson than before it.

These findings suggest that within AI literacy, the ability to distinguish visual AI is a distinct skill. Our lesson covered important AI literacy topics, including algorithms, machine learning, and ethical considerations. These topics are relevant to existing AI literacy frameworks (Organisation for Economic Co-operation and Development (OECD) and European Commission, 2025; Renkema et al., 2025), which emphasize understanding, using, evaluating, and critically reflecting on AI (Ng et al., 2021). However, the decrease in post-experiment scores suggests that knowing how AI works is not sufficient to identify AI-generated content.

A possible alternative explanation is cognitive exhaustion. Sievertsen et al. (2016) found that Danish students' test performance decreased by 0.9% for every hour in the day, while a 30-minute break improved performance by 1.7%. Since our lesson was 50 minutes, some decrease in performance afterward is expected. However, the decrease is substantially larger than this across all experiment groups (Table 5.4). Although the lunch break before group D may partly explain its higher performance compared with group C, cognitive exhaustion does not account for the increasing scores among groups A, B, and C. Thus, cognitive exhaustion may have contributed to the results but cannot fully account for the observed pattern.

Our findings highlight the role of item-level difficulty in visual AI-detection tasks. The results show large differences across image categories, with participants performing worse on images of pangolins and the Titanic but relatively well on images of papyrus and women's suffrage. This suggests that performance depends on the specific visual content shown, and that perhaps a single measure of accuracy/ability is not enough to capture the difficulty of a question. Due to randomization of pre- and post-experiment items, both tests had equal difficulty within the sample, but the results show differences in difficulty across image categories.

Finally, this study contributes by showing that AI-detection instruction may change children's distinguishing strategies. Before the lesson, participants may have relied more on intuition and their existing knowledge in judging whether an image is real or AI-generated. The lesson then provided explicit cues for recognizing AI-generated images, such as irregularities in limbs, faces, text, lighting, and time-spatial consistency. Rather than improving accuracy, this might have made participants more suspicious of all

images, including real images. The lesson may have reduced what Shaw and Nave (2026) describe as "cognitive surrender" to AI output, but it shifted children too far in the opposite direction. Instead of accepting images too readily as real, they may have become hyper-skeptical and over-applied the detection cues taught in the lesson. This could explain why post-experiment performance decreased, particularly why children also classified real images as AI-generated more often.

6.2 Practical implications

Our findings have several practical implications for schools, teachers, parents, policymakers/governments, developers of AI education materials, developers of AI systems, and, of course, children themselves.

For schools, our results suggest that AI literacy education does not necessarily improve children's AI detection skills, especially when it is offered as a single lesson. A single lesson might increase awareness, but it might not be enough to improve the ability to distinguish AI-generated content from real content. Moreover, as generative AI techniques are likely to continue improving, education should not only rely on teaching children fixed visual cues for detecting AI-generated content. Students need a varied curriculum in AI education across multiple lessons and subjects, in which they learn not only how AI works and what AI-generated content may look like, but also how to critically evaluate digital content.

For parents, our findings show that children may not readily recognize AI-generated images, even as exposure to AI-generated content increases. Parents should therefore not assume that children are digitally literate in the context of AI. We believe that parents should have ongoing conversations with their children about AI and its applications, including AI-generated images, deepfakes, social media posts, and fake news. Rather than only warning children about the risks of AI, parents should help children navigate the digital space and help children develop a habit of asking critical questions about all (online) information.

For policymakers and governments, our results suggest that AI literacy is a much broader term than solely understanding how AI works. AI literacy means not only understanding how an AI system works but also skills such as critically evaluating AI output, recognizing AI-generated output, and understanding AI's limitations. Therefore, educational materials and curricula should distinguish between different subfields of AI literacy goals, such as conceptual understanding, ethical awareness, practical AI use, and critical evaluation of online content.

Developers of AI systems should prioritize transparency when designing platforms for children. Our research found that it is not clear to children whether images are AI-generated or real images. Platforms that display AI-generated content should clearly label it as such with an age-appropriate label to help children understand which content might be AI-generated. The same argument also applies to AI chatbots, such as Snapchat AI. Developers should make it very clear that such a chatbot is artificial intelligence, since, based on our research on AI-generated visual content, children cannot distinguish a chatbot from its output alone.

Finally, for children themselves, our findings suggest that it is important to understand that AI-generated content is not always easy to recognize. Children should not trust everything they see online and should ask critical questions about the content they encounter. Besides that, children should not be ashamed that they can not see that the content is AI-generated. They should know that it is a common issue and that they can talk to adults about what they see online.

6.3 Limitations

Our study has several limitations that should be taken into account when interpreting our findings. We split those into limitations of the sample and limitations of our methodology.

6.3.1 Sample

First, the experiment's sample size was limited. Although 78 children participated, only 65 students were included in the final analysis. This limits the statistical power of the analysis. Several tests found a positive correlation, but we failed to statistically prove the effect.

Besides that, we only had children aged around 10 years old. That means all of our findings are only confirmed in children aged 10 years. Therefore, we could not compare results between different age groups or with adults.

Also, our experiment consisted only of Belgian (Flemish) children from a specific region. Research comparing Belgian and Dutch children is very limited; the only research on digital skills dates to 1999, when Beentjes et al. (1999) found that Belgian (Flemish) parents are stricter about media use than Dutch parents and that Flemish parents talk more to their children about media use. However, the digital landscape has changed significantly since 1999, so it is unclear whether those findings remain applicable and whether our study's results are directly transferable to the situation in the Netherlands. On a larger scale, it is even less clear if our findings are transferable to other countries.

6.3.2 Methodology

During the survey, participants were asked to report their AI usage and trust in AI on a Likert(-like)-scale. Children completed the survey in a classroom setting and provided a self-reported measure. This does not fully capture their real AI usage or trust in AI, as a behavioral measure would.

In addition, children might have struggled to correctly understand the questions being asked. Teachers filled in a mock survey when their students were filling in the survey, and one teacher commented that especially the third question about trust *("I trust that this chatbot is honest and transparent about what it can and cannot do.")* was not easy to understand for children aged 10 years old. However, looking at Figure 5.2, we do not see an obvious difference between Q4 and Q5. A supplementary analysis did not show that Cronbach's alpha increased when Q4 was removed; in fact, it was lower ($\alpha = 0.757$ with Q4, $\alpha = 0.678$ without Q4).

Our AI-distinguishing task used a limited set of images. Images were from an educational quiz and included categories such as landmarks, animals, historical events, people, and nature. However, this set may not represent the full range of AI-generated images that children can encounter online. Our task also focused solely on still images; direct generalization of our findings to other content types, such as videos, text, audio, and social media posts, is therefore not possible.

The methodology and images used in the AI-distinguishing task may not reflect how children encounter AI-generated content in everyday online environments. Participants were asked to make a binary judgment about whether a still image is AI-generated, without surrounding context, captions, comments, or information about the source. This matters because children's interpretations of pictures are not based solely on the image's literal content. Vivaldi and Allen (2021) found that children can interpret pictures flexibly depending on the context in which they are presented. Thus, it is likely that children use the image's context to evaluate its content. These are not only visual cues, but also who posted the content, whether the caption seems trustworthy, and how other users respond to the content. Therefore, the task used in this study may have measured a narrower form of AI detection than the kind of critical evaluation skills children need online. Besides that, the images themselves may not reflect the types of images children will encounter online, whereas the visual cues and examples during the lesson are more like content they might encounter on social media. As a result, it is possible that the lesson was more effective than the measured accuracy suggests, but this effect on the critical evaluation of online content encountered was not captured well by the specific binary image-classification task used in this study.

Another limitation concerns the match between the lesson content and the outcome measure. Although the post-test measured children's ability to distinguish AI-generated images from real images, the lesson itself focused more broadly on AI and machine learning. Participants learned about algorithms, machine learning, training data, classification, and ethical issues, while only the final part of the lesson focused directly on visual cues for recognizing AI-generated images. Therefore, the lesson may have improved children's general understanding of AI without providing sufficient focused practice on the specific skill measured by the AI-distinguishing task. The expected improvement in accuracy (following Hypothesis 4) may have been more likely to appear if the lesson had been centered more strongly on direct detection skills, repeated examples, and guided practice with feedback. The mismatch between the broader AI literacy intervention and the narrower post-test task may partly explain why accuracy did not improve after the lesson.

Related to that, the measure of the AI-distinguishing task focused primarily on accuracy, but not on reasoning or response bias. Students did not explain why they classified an image as AI-generated or

real. Thus, we do not know whether incorrect answers were caused by guessing, misunderstanding of the task, overthinking, overconfidence, or the specific visual cues students used in their reasoning. Moreover, accuracy alone does not distinguish between children's ability to discriminate AI-generated from real images and their tendency to choose one response category over the other. Since participants performed better on real images than on AI-generated images, some of the observed performance pattern may reflect a default tendency to classify uncertain images as real.

Finally, the setting might influence the findings. Students came to a science day at a school different than their everyday school, like a field trip. Students might have been excited for all the workshops, or they might be distracted by a new environment. Therefore, it could be that students in their home classroom are more focused and answer the questions differently.

6.3.3 AI Literacy

In this study, we treat the ability to distinguish AI-generated images from real images as one component of AI literacy. The result suggests that a short AI literacy lesson did not improve children's classification accuracy, but the question arises whether classification itself should be an end goal. The performance of generative models is likely to improve further, and current cues that indicate an image is AI-generated, such as hands, facial features, shadows, and spatial relations described by Kamali et al. (2025), might not be present in the future. Therefore, there is a risk that educating children to look for current visual cues will become outdated as generative models advance.

Rather than assuming adults know the signs and visual cues of AI-generated content and simply need to educate children, AI literacy may need to shift from a binary "yes/no" detection model to a broader critical evaluation model. In such a model, children should learn to ask where content comes from, whether the source is trustworthy, whether it aligns with children's existing knowledge, whether it may have been manipulated or generated, and what the consequences are of sharing or believing it. We believe this is a better approach for the future, when real and AI-generated images will no longer be distinguishable.

Chapter 7

Conclusions

This thesis investigated how well children can distinguish AI-generated images from real images and how this ability is influenced by individual characteristics, such as fluid intelligence, AI use, trust in AI, gender, and AI education. We conducted an experiment with 65 Belgian primary school students, who completed a pre-test, followed an AI lesson, and then completed a post-test on visual AI distinguishing ability. In this section, we summarize our findings by answering our research questions, discussing our contributions to the literature, and posing opportunities for future work.

7.1 Answers to the research questions

[RQ.1] How accurately do children distinguish AI-generated images from real images before receiving AI literacy education?
Our results show that children struggled with this task. Pre-experiment classification performance for AI-generated images vs. real images was 51.43%, slightly above chance. However, when split between AI-generated and real images, children were better at recognizing real images (60.26%) than AI-generated images (42.48%), suggesting that they more often interpreted images as real rather than AI-generated. This indicates that children are not able to recognize AI-generated visual content.

[RQ.2] How are fluid intelligence, AI usage, and trust in AI related to children's AI-distinguishing accuracy?
Contrary to our expectations, we did not find statistical evidence that fluid intelligence, AI usage, and trust in AI are significantly related to children's AI-distinguishing accuracy. For fluid intelligence, we found a positive regression coefficient, but our study does not statistically demonstrate this effect. We found a statistical relationship between AI usage and trust in AI, suggesting that familiarity with AI and trust in AI are connected, although the direction of causality cannot be determined from our study.

[RQ.3] How does gender influence fluid intelligence or AI-distinguishing accuracy?
The results of our study confirm our expectation that there are no significant gender differences in fluid intelligence, and an equivalence test supported the conclusion that the fluid intelligence scores between boys and girls are equivalent within the predefined bounds. For AI-distinguishing accuracy, we also found no significant effect; however, the equivalence tests did not confirm that boys and girls perform equivalently. Thus, our study cannot conclude that gender has no influence on AI-distinguishing accuracy.

[RQ.4] How does an AI literacy lesson affect children's AI-distinguishing accuracy?
Our hypothesis was that children would achieve higher accuracy after receiving an AI literacy lesson. However, children's accuracy did not improve after the lesson; instead, post-experiment ability scores were significantly lower than pre-experiment scores. This suggests that a single short lesson on AI and machine learning is insufficient to improve children's ability to distinguish AI-generated images from real ones. One possible explanation is that our lesson mainly focused on understanding AI and machine learning, while only the last part addressed how to distinguish AI-generated images from real ones. Another possible explanation is that the lesson led children to overthink the task and become more suspicious of all images.

[RQ.5] Does fluid intelligence or gender moderate the effect of the lesson on visual AI-

distinguishing accuracy?
Our findings showed that fluid intelligence did not significantly moderate the change in accuracy between the pre-test and post-test. Similarly, gender did not significantly moderate the change in AI-distinguishing accuracy. However, the wide confidence interval does not completely rule out any moderating effects of gender on changes in accuracy.

7.2 Contributions

A major contribution of this study is the finding that children struggle to distinguish AI-generated images from real ones. Since children are increasingly exposed to AI-generated content, our study suggests that they should be considered vulnerable users, as they may struggle to recognize AI-generated outputs when encountered online.

Additionally, we did not find evidence for the expected relationship between fluid intelligence and visual AI-distinguishing ability. While our findings are not significant, this suggests that recognizing AI-generated images might depend less on general reasoning ability and more on specific knowledge. For example, distinguishing distorted text and unrealistic shadows may require targeted instruction and repeated practice rather than abstract reasoning alone.

We did find a significant positive relationship between AI usage and trust in AI. This indicates that children who use AI more also tend to trust AI more. Although our study cannot determine the direction of causality (whether usage leads to trust or trust leads to usage), the finding shows that AI usage and trust in AI are related and warrant further investigation.

Finally, our study shows that a single AI literacy lesson is insufficient to improve practical AI detection skills. While the lesson might have contributed to general understanding of AI and machine learning, the results suggest that conceptual understanding and practical skills are not the same. AI education should therefore distinguish among different learning goals, such as understanding how AI works, using AI responsibly, critically evaluating online information, and recognizing AI-generated content in practice.

7.3 Future work

As future work, we suggest conducting the same study with a larger, more diverse sample. Our study focused on 10-to 11-year-old children from one region in Belgium, which limits the generalizability of the findings. When running the same experiment across different age groups, the findings can be compared to see how age groups differ in their ability to distinguish AI-generated content. The experiment can be repeated across different countries to examine differences in AI literacy. Lastly, the experiment can be run across different school levels (e.g., high schools) to see whether education level influences AI literacy.

Furthermore, our research focused solely on quantitative data collection, giving students the option to choose between an AI-generated image and a real image, rather than asking them to explain their choice. We were unable to see how different participants decided whether a given image was AI-generated or real. This information, for example, gathered through interviews, think-aloud tasks, or written explanations, could provide insights into which images are the hardest to distinguish and how educational material could be better designed to assist learning to distinguish AI-generated content. Such future work could also examine whether the binary classification between real and AI-generated images is the most useful educational goal, or whether AI literacy should focus more on broader skills such as evaluating the source, context, credibility, and possible consequences of content.

In addition, future research should use a more diverse set of AI-generated content. Our study focused on still images from a limited number of categories. However, online content does exist in many other forms, such as videos, audio, text, and social media posts. Future research should investigate whether our findings on children's difficulty with AI-generated image classification also apply to other media formats.

Lastly, we suggest investigating the long-term effects of both our learning intervention and children's scores. We measured the learning rate immediately after the lesson, whilst it is more important to know if they retain the learned skills. At a societal level, measuring the long-term trend in AI-distinguishing scores would indicate whether current education is sufficient or whether revisions are needed.

Bibliography

A. A. Alasmari, R. F. Alruwaili, R. F. Alotaibi, I. K. Youssef, and S. A. Asklany. Demographic influences on trust in artificial intelligence across cognitive domains: A statistical perspective. *Plos one*, 20(11): e0331003, 2025.

M. C. C. A. P.-d.-A. F. M. Amat-Lefort, Natalia; Yazan. From chatbots to confidants: A cross-cultural study of llm adoption for emotional support, 2026. URL `https://arxiv.org/abs/2604.25525`.

V. Atanasov and A. Ivanova. A framework for measurement of interactivity of digital learning resources. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 649–654. IEEE, 2019.

J. Beentjes, L. d'Haenens, T. Van Der Voort, and C. M. Koolstra. Nederlandse en vlaamse kinderen en jongeren als gebruikers van interactieve media. *Tijdschrift voor communicatiewetenschap*, 27(2), 1999.

W. B. Bilker, J. A. Hansen, C. M. Brensinger, J. Richard, R. E. Gur, and R. C. Gur. Development of abbreviated nine-item forms of the raven's standard progressive matrices test. *Assessment*, 19(3): 354–369, May 2012. doi: 10.1177/1073191112446655.

Britannica Education. Real or ai quiz: Can you tell the difference? URL `https://britannicaeducation.com/blog/quiz-real-or-ai/`.

M. Bui and D. P. Birney. Learning and individual differences in gf processes and raven's. *Learning and Individual Differences*, 32:104–113, 2014. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2014.03.008. URL `https://www.sciencedirect.com/science/article/pii/S1041608014000545`.

M. Carney, B. Webster, I. Alvarado, K. Phillips, N. Howell, J. Griffith, J. Jongejan, A. Pitaru, and A. Chen. Teachable machine: Approachable web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–8, 2020.

J. B. Carroll. The three-stratum theory of cognitive abilities. 1997.

R. B. Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1, 1963.

R. B. Cattell. *Intelligence: Its structure, growth and action*. Elsevier Science Publishers, 1987.

Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, and et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, Mar 2024. doi: 10.1145/3641289.

A. Chatterji, T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman. How people use chatgpt, Sep 2025. URL `https://www.nber.org/papers/w34255`.

J. Chein, S. Martinez, and A. Barone. Human intelligence can safeguard against artificial intelligence: individual differences in the discernment of human from ai texts. *Scientific Reports*, 14(1):25989, 2024.

R. Colom and O. Garcıa-López. Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, 32(3):445–451, Feb 2002. doi: 10.1016/s0191-8869(01)00040-x.

T. Debatin, B. Harder, and A. Ziegler. Does fluid intelligence facilitate the learning of english as a foreign language?—a longitudinal latent growth curve analysis. *Learning and Individual Differences*, 70:121–129, 2019. ISSN 1041-6080. doi: https://doi.org/10.1016/j.lindif.2019.01.009. URL `https://www.sciencedirect.com/science/article/pii/S1041608019300159`.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

J. Digout and H. E. Samra. Interactivity and engagement tactics and tools. In *Governance in Higher Education: Global Reform and Trends in the MENA Region*, pages 151–169. Springer, 2023.

L. Edwards. The eu ai act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*, 1:25, 2021.

European Parliament. Artificial Intelligence Act: Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Texts adopted, P9_TA(2023)0236, June 2023. URL `https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf`. Accessed: 2026-06-01.

D. Giofrè, K. Allen, E. Toffalini, and S. Caviola. The impasse on gender differences in intelligence: A meta-analysis on wisc batteries. *Educational Psychology Review*, 34(4):2543–2568, Sep 2022. doi: 10.1007/s10648-022-09705-1.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

J. L. Horn. Measurement of intellectual capabilities: A review of theory. *Woodcock-Johnson technical manual*, pages 197–232, 1991.

M. Irrgang, M. Dorenkamp, P. Reohr, and P. Vik. C-28 raven's progressive matrices: Validation of a short form. *Archives of Clinical Neuropsychology*, 34(6):1057–1057, Jul 2019. doi: 10.1093/arclin/acz034.190.

N. Kamali, K. Nakamura, A. Kumar, A. Chatzimparmpas, J. Hullman, and M. Groh. Characterizing photorealism and artifacts in diffusion model-generated images. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2025.

A. D. Kaplan, T. T. Kessler, J. C. Brill, and P. A. Hancock. Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2):337–359, 2023.

A. S. Kaufman, J. C. Kaufman, X. Liu, and C. K. Johnson. How do educational attainment and gender relate to fluid intelligence, crystallized intelligence, and academic skills at ages 22-90 years? *Archives of Clinical Neuropsychology*, 24(2):153–163, Mar 2009. doi: 10.1093/arclin/acp015.

S. Kroon. Mother tongue and mother tongue education. In *World Yearbook of Education 2003*, pages 47–60. Routledge, 2003.

K. La Fors. 2022 ai register of children in the netherlands: Mapping children's awareness, ethical and social sense-making and imaginaries of artificially intelligent systems via meaningful participation. *2022 AI Register of Children in The Netherlands*, 2023. URL `https://ris.utwente.nl/ws/portalfiles/portal/307143534/Full_Report._2022_AI_Register_of_Children_in_the_Netherlands.pdf`.

T. Ladd and P. Shrivastava. Skeptical intelligence: Refining critical thinking for ai-powered innovation. *Available at SSRN 6060134*, 2026.

A. M. Langener, A.-W. Kramer, W. van den Bos, and H. M. Huizenga. A shortened version of raven's standard progressive matrices for children and adolescents. *British Journal of Developmental Psychology*, 40(1):35–45, 2022.

P. Lin, J. Van Brummelen, G. Lukin, R. Williams, and C. Breazeal. Zhorai: Designing a conversational agent for children to explore machine learning concepts. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13381–13388, 2020.

D. Long and B. Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16, 2020.

R. Lynn and P. Irwing. Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, 32 (5):481–498, Sep 2004. doi: 10.1016/j.intell.2004.06.008.

F. Marimon, M. Mas-Machuca, and A. Akhmedova. Trusting in generative ai: Catalyst for employee performance and engagement in the workplace. *International Journal of Human–Computer Inter-*

action, 41(11):7076–7091, 2025. doi: 10.1080/10447318.2024.2388482. URL https://doi.org/10.1080/10447318.2024.2388482.

National 4-H Council. Artificial Intelligence (AI) Use & Understanding Survey, n.d. URL https://4-h.org/about/research/teen-perspectives/artificial-intelligence/. Accessed: 2026-06-01.

D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao. Conceptualizing ai literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2:100041, 2021.

E. I. Nikolaeva, I. A. Kalabina, T. K. Progackaya, and E. V. Ivanova. Ground rules for preschooler exposure to the digital environment: A review of studies. *Psychology in Russia: State of the Art*, 16 (4):37–54, 2023. doi: 10.11621/pir.2023.0403.

Organisation for Economic Co-operation and Development (OECD) and European Commission. Empowering learners for the age of AI: AI literacy framework for primary and secondary education. Review draft, OECD / European Commission, 2025. URL https://ailiteracyframework.org/wp-content/uploads/2025/05/AILitFramework_ReviewDraft.pdf. Accessed: May 2026.

L. S. Penrose and J. C. Raven. A new series of perceptual tests: preliminary communication. *British Journal of Medical Psychology*, 1936.

I. Picton and C. Clark. Children and young people's use of generative ai to support literacy in 2024. Technical report, National Literacy Trust, London, June 2024. URL https://nlt.hacdn.org/media/documents/Children_and_young_peoples_use_of_AI_to_support_literacy_in_2024_vFSWEyC.pdf. Accessed: 2026-06-01.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

M. Renkema, E. van den Boom-Muilenburg, I. Friso-van den Bos, H. Theelen, I. Wopereis, and K. Schildkamp. Ai-go! een raamwerk voor ai-geletterdheid in het onderwijs (ai-go raamwerk. 2025.

M. R. Reynolds and T. Z. Keith. Multi-group and hierarchical confirmatory factor analysis of the wechsler intelligence scale for children—fifth edition: What does it measure? *Intelligence*, 62:31–47, May 2017. doi: 10.1016/j.intell.2017.02.005.

M. B. Robb and S. Mann. Talk, trust, and trade-offs: How and why teens use ai companions. Technical report, Common Sense Media, San Francisco, CA, 2025.

J. D. Rodríguez-García, J. Moreno-León, M. Román-González, and G. Robles. Evaluation of an online intervention to teach artificial intelligence with learningml to 10-16-year-old students. In *Proceedings of the 52nd ACM technical symposium on computer science education*, pages 177–183, 2021.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

N. M. Safdar, J. D. Banja, and C. C. Meltzer. Ethical considerations in artificial intelligence. *European Journal of Radiology*, 122:108768, Jan 2020. doi: 10.1016/j.ejrad.2019.108768.

M. Schmidmaier, J. Rupp, D. Cvetanova, and S. Mayer. Perceived empathy of technology scale (pets): Measuring empathy of systems toward the user. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

W. J. Schneider and K. S. McGrew. The cattell-horn-carroll model of intelligence. 2012.

F. A. Schrank and B. J. Wendling. The woodcock–johnson iv. *Contemporary intellectual assessment: Theories, tests, and issues*, 383, 2018.

D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, 15(6):657–680, 1987.

K. Schweizer and W. Koch. A revision of cattell's investment theory: Cognitive properties influencing learning. *Learning and Individual Differences*, 13(1):57–82, 2002. ISSN 1041-6080. doi:

https://doi.org/10.1016/S1041-6080(02)00062-6. URL https://www.sciencedirect.com/science/article/pii/S1041608002000626.

S. D. Shaw and G. Nave. Thinking—fast, slow, and artificial: How ai is reshaping human reasoning and the rise of cognitive surrender. 01 2026. doi: 10.31234/osf.io/yk25n_v1.

H. H. Sievertsen, F. Gino, and M. Piovesan. Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences*, 113(10):2621–2624, 2016.

C. Sindermann, H. S. Schmitt, D. Rozgonjuk, J. D. Elhai, and C. Montag. The evaluation of fake and true news: on the role of intelligence, personality, interpersonal trust, ideological attitudes, and news consumption. *Heliyon*, 7(3):e06503, 2021. ISSN 2405-8440. doi: https://doi.org/10.1016/j.heliyon.2021.e06503. URL https://www.sciencedirect.com/science/article/pii/S240584402100606X.

C. E. Snow. Rationales for native language instruction. *Bilingual education: Issues and strategies*, pages 60–74, 1990.

K. Somoray, D. J. Miller, and M. Holmes. Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, 2025(1):1833228, 2025.

J. Su and W. Yang. Artificial intelligence in early childhood education: A scoping review. *Computers and Education: Artificial Intelligence*, 3:100049, 2022. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2022.100049. URL https://www.sciencedirect.com/science/article/pii/S2666920X22000042.

C. Thorsen, J.-E. Gustafsson, and C. Cliffordson. The influence of fluid and crystallized intelligence on the development of knowledge and skills. *British Journal of Educational Psychology*, 84(4):556–570, 2014.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

R. A. Vivaldi and M. L. Allen. Beyond literal depiction: Children's flexible understanding of pictures. *Journal of Experimental Child Psychology*, 210:105208, 2021.

R. Williams, H. W. Park, L. Oh, and C. Breazeal. Popbots: Designing an artificial intelligence curriculum for early childhood education. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9729–9736, 2019.

A. Yang, L. Jarrett, and F. Gallagher. The family of teenager who died by suicide alleges openai's chatgpt is to blame, 2025. URL https://www.nbcnews.com/tech/tech-news/family-teenager-died-suicide-alleges-openais-chatgpt-blame-rcna226147.

N. Đapo and J. Kolenović-Đapo. Sex differences in fluid intelligence: Some findings from bosnia and herzegovina. *Personality and Individual Differences*, 53(7):811–815, Nov 2012. doi: 10.1016/j.paid.2012.05.036.

Appendix A

AI-generated Images

In this appendix, we include all AI-generated and real images that we have used in the surveys. The order of the images was random, and thus not the order in which we display them here. The images are from an online quiz from Britannica Education.



(a) AI-generated image



(b) Real image

Figure A.1: Category: Geography (Hong Kong at Night)



(a) AI-generated image



(b) Real image

Figure A.2: Category: Significant Landmarks (Mount Fuji, Japan)



(a) AI-generated image



(b) Real image

Figure A.3: Category: Natural Phenomena (Aurora Borealis, the Northern Lights)



(a) AI-generated image



(b) Real image

Figure A.4: Category: Endangered Species (Pangolin)



(a) AI-generated image



(b) Real image

Figure A.5: Category: World History (Ancient Egyptian Papyrus)



(a) AI-generated image



(b) Real image

Figure A.6: Category: Significant Landmarks (Mount Fuji, Japan)



(a) AI-generated image



(b) Real image

Figure A.7: Category: People (Students Doing Art Projects)



(a) AI-generated image



(b) Real image

Figure A.8: Category: Space Exploration (Shuttle Launch)



(a) AI-generated image



(b) Real image

Figure A.9: Category: Influential Events and Figures (Women's Suffrage Movement)



(a) AI-generated image



(b) Real image

Figure A.10: Category: Historical Events (The Titanic)



(a) AI-generated image



(b) Real image

Figure A.11: Category: Wildlife and Nature (Poisonous Frogs)

Appendix B

Example Surveys

Survey

Code: beae

This survey is to see how well you can distinguish between real images and AI-generated images.

- It is not a test and you do not receive a grade
- What you fill in has no impact on the rest of the workshop
- Nobody knows that you have given these answers
- If you wish, you can stop at any time

During this workshop, you will learn more about AI, and we will see if you can better distinguish between real and AI-generated images.

I understand the purpose of this questionnaire and how my responses will be used.

◯ *Yes*

General Questions

1) What is your gender?

◯ *Male*

◯ *Female*

◯ *Other*

AI Usage

Please indicate how often you use AI. Think of ChatGPT, Gemini, but also Snapchat AI, Google Home, Siri, Alexa, etc.

2) How often do you use AI?

Daily	*Weekly*	*Monthly*	*Rarely*	*Never*
◯	◯	◯	◯	◯

Trust in AI

Please indicate how much you trust AI.

3) AI provides credible information

Completely Agree	*Agree*	*Neutral*	*Disagree*	*Completely Disagree*
◯	◯	◯	◯	◯

4) I trust the information that AI provides

Completely Agree	*Agree*	*Neutral*	*Disagree*	*Completely Disagree*
◯	◯	◯	◯	◯

5) I trust that AI is honest and transparent about what it can and cannot do

Completely Agree	*Agree*	*Neutral*	*Disagree*	*Completely Disagree*
◯	◯	◯	◯	◯

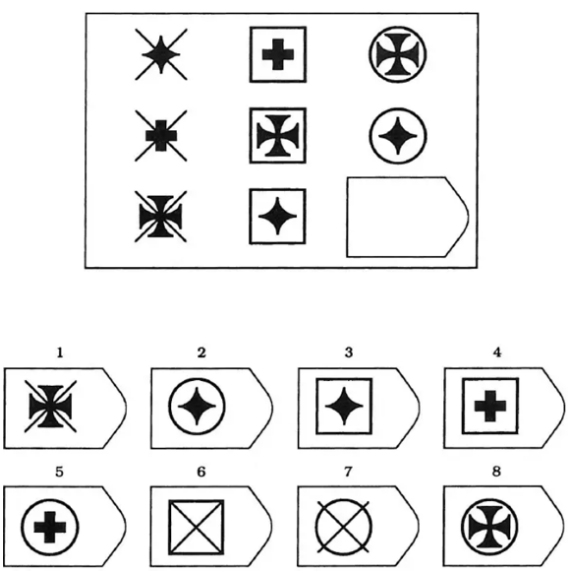
6) Overall, I trust AI

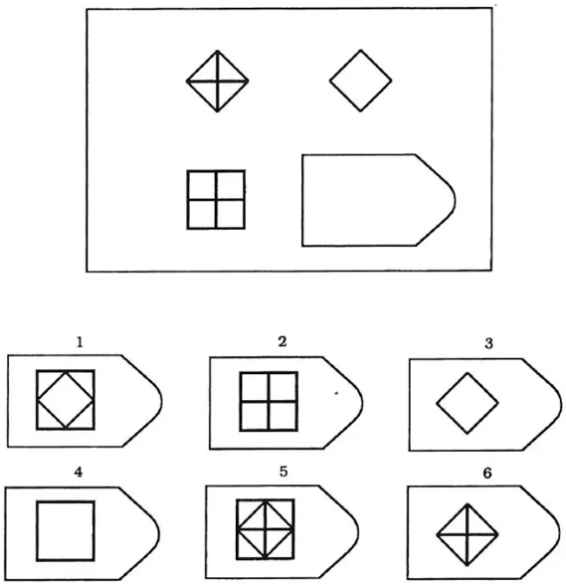
Completely Agree	*Agree*	*Neutral*	*Disagree*	*Completely Disagree*
◯	◯	◯	◯	◯

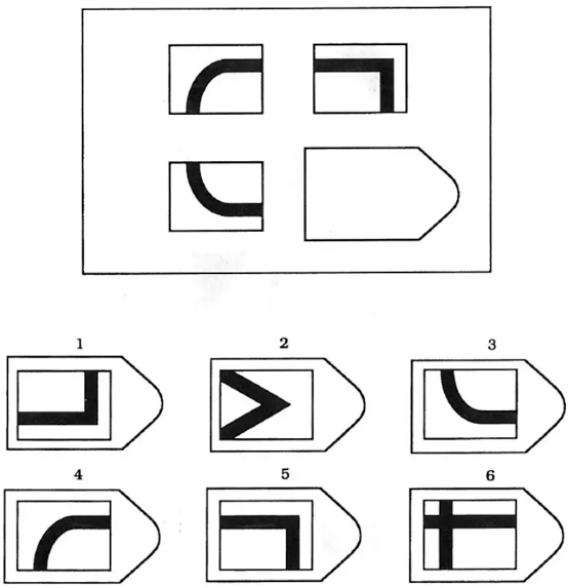
beae

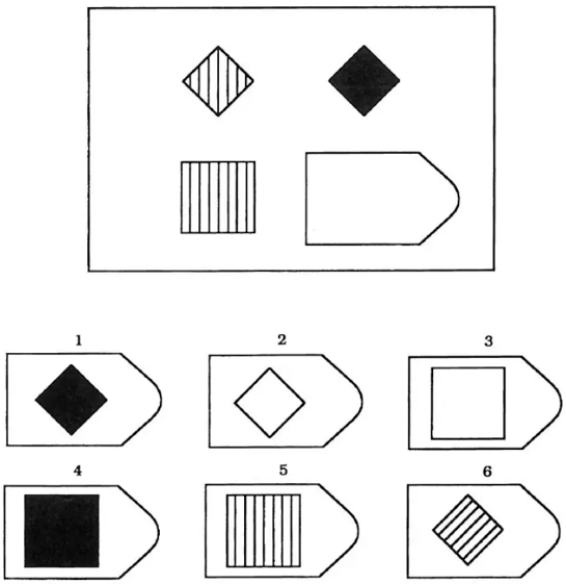
Patterns

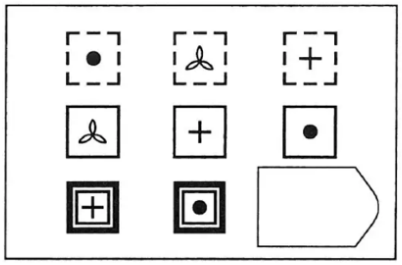
On the following pages, you will see different images. In each image, a piece is missing. Circle the piece that fits to complete the image.

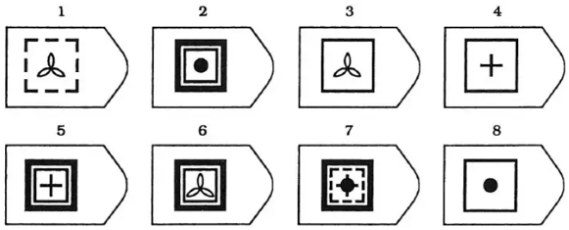












beae

Human or AI?

On the following pages, you will see different images. These images are created by a human or by an AI. Please indicate whether you think each image is created by an AI or by a human.



Real photo ◯ *Created by AI* ◯



Real photo ◯ *Created by AI* ◯



Real photo ○ *Created by AI* ○



Real photo ○ *Created by AI* ○

beae



Real photo

Created by AI



Real photo

Created by AI

beae



Real photo

Created by AI

beae

Draw a dog or a cat

Make sure you have answered all previous questions. Draw a dog or a cat in the space below, at your choice. You have the time until the timer is up.

Survey

Code: beae

Human or AI

On the following pages, you will see different images. These images are created by a human or by an AI. Please indicate whether you think each image is created by an AI or by a human.



Real photo ◯ *Created by AI* ◯



Real photo ◯ *Created by AI* ◯



Real photo ◯ *Created by AI* ◯



Real photo ◯ *Created by AI* ◯

beae



Real photo

Created by AI



Real photo

Created by AI

beae



Real photo

Created by AI

beae

Appendix C

Lesson Slides

AI:
Hoe werkt het en hoe herken je het?

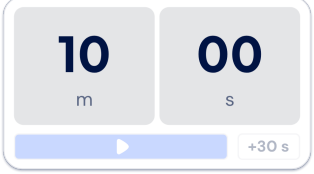


Enquête

Deze vragenlijst is om te kijken hoe goed je echte plaatjes en door AI gemaakte plaatjes uit elkaar kunt houden.

- Het is **geen toets** en je krijgt **geen cijfer**
- Wat je invult heeft **geen invloed** op de rest van de workshop
- **Niemand** weet dat jij deze antwoorden hebt gegeven
- Als je wilt, mag je **op elk moment stoppen**

Vul alle vragen in van boekje 1.
De laatste vraag is een tekening,
deze vraag is voor als je tijd over hebt.



Wat is een algoritme?

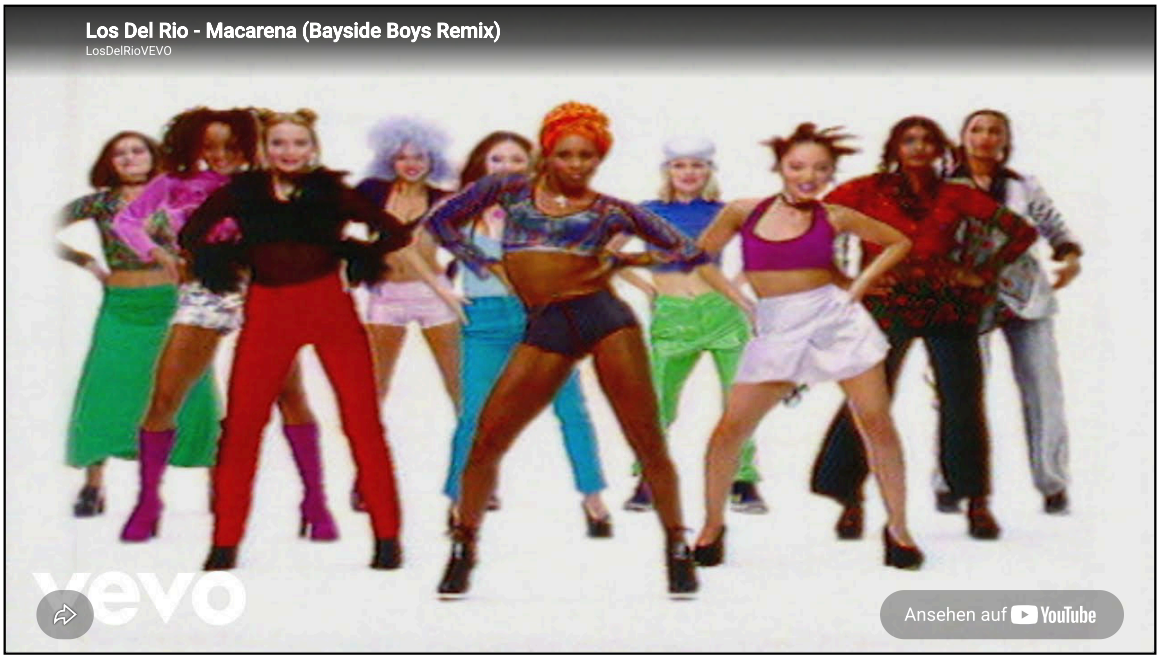
AI-plaatjes maken is 1 groot algoritme, maar wat is een algoritme?

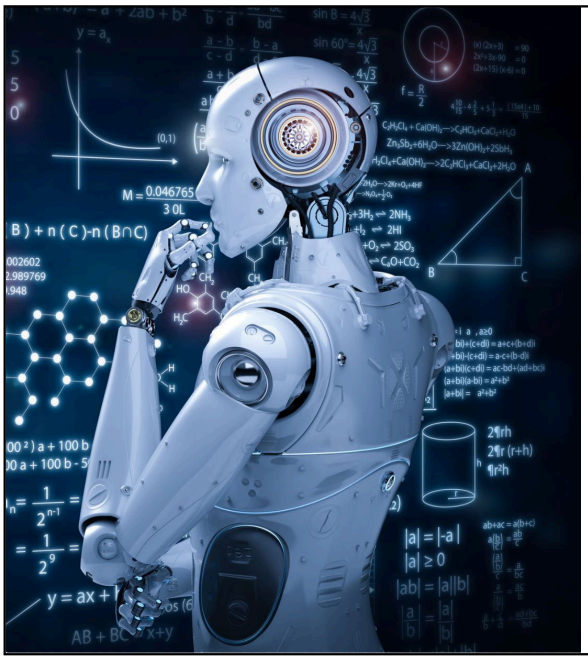
Algoritme

Jullie mogen gaan staan, en de volgende stappen uitvoeren:

- Strek je armen
- Draai je handen om
- Kruis je handen op je schouders
- Zet je handen op je wang
- Kruis je handen op je heupen
- Zet je handen op je heupen

Komt dit bekend voor?





Machine Learning

Een algoritme is dus een reeks stappen.

De machine **'leert'** doordat wij als mensen informatie geven en dat te 'vertalen' naar machinetaal

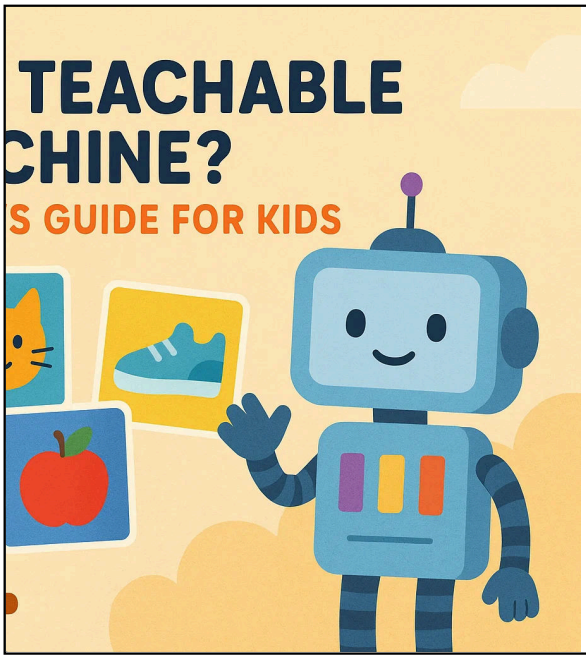
- Wij gaan alien Zhorai iets leren over de aarde!



Hoe heeft Zhorai Nederlands geleerd?

Weet Zhorai over wat goed en fout is?

Hoe kunnen we Zhorai fouten laten maken?



AI en plaatjes

De machine '**leert**' doordat wij als mensen informatie geven.

- Dit was nog steeds met tekst, hoe werkt het met plaatjes?
- Dat gaan we bekijken met jullie tekeningen!

https://teachablemachine.withgoogle.com/models/jxKIbjhq7/



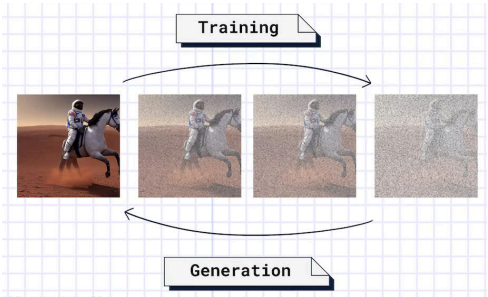
Maar we kunnen het AI best lastig maken...

Van herkennen naar maken

Tijdens de training worden
de patronen herkend
(net als bij de honden en katten)

Maar nu proberen we die patronen
'na te maken' op basis van wat we
weten

Maar dat werkt niet altijd goed...



Hoe herken je AI plaatjes?



De paus met een Balenciaga-jas

Lijkt het nep? Dan is dat waarschijnlijk ook zo

Klopt het met tijd & plek? Is het logisch?



Een vrouw met een rare vinger?



Philoophy?

Kijk naar handen/ogen/benen en naar tekst

Realistische beelden zijn lastig, net als tekst



Schaduw voor én achter?



Kun je zo dicht bij flamingo's komen?

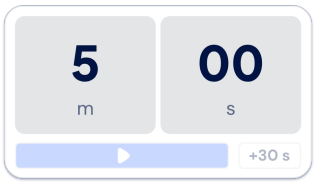
Klopt de achtergrond?

Enquête

Vul de vragen in boekje 2 in.

- Het is **geen toets** en je krijgt **geen cijfer**
- Wat je invult heeft **geen invloed** op de rest van de workshop
- **Niemand** weet dat jij deze antwoorden hebt gegeven
- Als je wilt, mag je **op elk moment stoppen**

Ben je klaar? Lever dan beide boekjes in



AI:
Hoe werkt het en hoe herken je het?

Bedankt voor het meedoen!
Laat het blaadje op je tafel liggen