Learning to Detect AI-Generated Images: Socio-Cognitive or Visual Processing Factors?

Lisa Bevers

S3458938

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number 11

Supervisor: dr. Ben Gutzkow

Second evaluator: prof. dr. Ernestine H. Gordijn

In collaboration with: Tess Walvius (s5490146), Megan Keane (s4697510), Maaike de Kruijf (s4351517), Moritz Noß (s4385829), Merwe Oosterhof (s5201632)

January 23rd, 2026

Declaration of AI use

2. AI used for background/self-study only

“I acknowledge the use of ChatGPT to generate materials for background research and self-study in the drafting of this assessment.”

Abstract

Generative Artificial Intelligence (GAI) is improving rapidly, creating increasing difficulty in distinguishing human made images from GAI images. This research explores whether inductive learning can improve GAI detection and whether individual differences in socio-cognitive skills, such as Theory of Mind (ToM), or visual processing skills, such as visual disembedding, contribute to detection accuracy. In an online experiment (N = 267) participants had to identify whether an image of a portrait expressing an emotion (happy, scared or angry) was GAI or human made, with a brief inductive training in the experimental condition. Participants additionally performed a ToM test (Reading the Mind in the Eyes) and a visual disembedding test (Leuven Embedded Figures Test). Results showed that inductive learning strongly improved GAI image detection. ToM showed marginally significant positive association with detection, while visual disembedding was not significant. No interaction was found between inductive learning and ToM. These findings indicate that inductive learning can improve GAI detection accuracy, independent of individual differences in ToM. After accounting for learning, ToM contributes to GAI detection, suggesting the socio-cognitive factors to be influential. This interpretation is further supported by the absence of a relationship between visual disembedding with both ToM and GAI detection accuracy. The results suggest the potential of inductive learning as a scalable intervention to improve GAI detection, however, future research is required to determine the longitudinal effects of training in order to improve humans ability to recognize GAI images from reality.

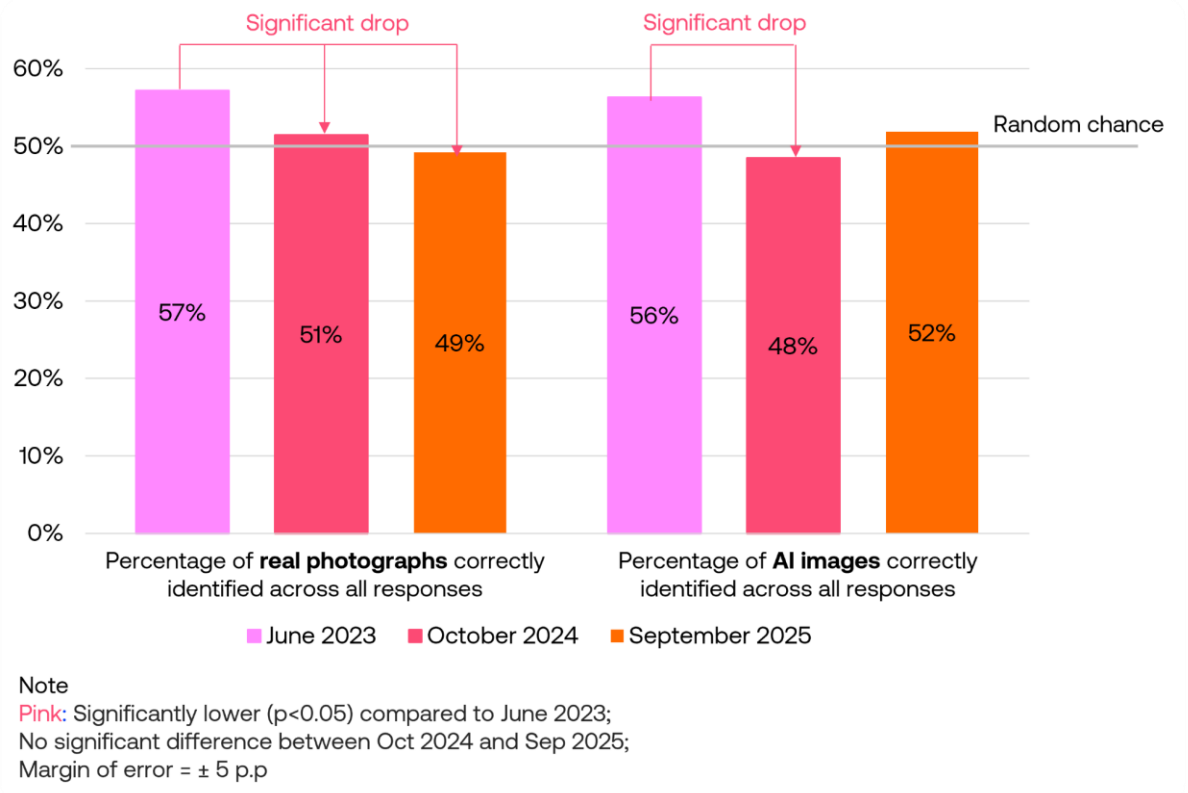
Keywords: Generative Artificial Intelligence, Inductive Learning, Theory of Mind, Visual Disembedding.

Learning to Detect AI-Generated Images: Socio-Cognitive or Visual Processing Factors?

In the span of a few years, Generative Artificial Intelligence (GAI) has reached a level of realism that challenges humans' ability to distinguish artificial creations from reality. Although the opportunities of the creation of GAI content through the widely accessible OpenAI platforms seem endless (Cao et al., 2023; Islam, 2024), the scale of the technological shift introduces new risks. Already in August 2023, an undetected GAI image was accidentally shown during a Dutch news broadcast of the NOS (Haan, 2023, August 25th). The GAI images have become increasingly more difficult to distinguish, introducing an increasingly dangerous trend of the inability of society to keep up with the rapid growth of GAI.

Figure 1

Identification accuracy for AI vs human made images from 2023 to 2025. Reproduced from Conjointly (Lee, 2025, September 25th)



The prominent danger of GAI images is the inability to distinguish GAI from reality, which is becoming increasingly more difficult. Figure 1, reproduced from a longitudinal survey study done by (Lee, 2025, September 25th), shows a worrying trend of recognition of GAI images around chance

level. The danger of GAI is this inability to detect the difference between human made and GAI images, leading to issues of deception and bias (Tredinnick & Laybats, 2023). Deception is one of the biggest dangers with the inability to distinguish GAI images, as it can convincingly create deceptive content (Islam, 2024). Deepfake images can be created and used for financial fraud, political manipulation through non-consensual images and harassment (Romero-Moreno, 2025). In addition, images created by GAI are biased as they are created based on existing prejudice. As GAI uses the information that is available online, the societal biases are translated to GAI content and further spread online (Islam, 2024). Due to the lack of detection, even prominent news channels start sharing fake images, like the NOS (Haan, 2023, August 25th), leading to misinformation and bias being publicized to society.

The continuous improvements of GAI become increasingly dangerous as long as humans are not capable of detecting GAI from human made images, which raises the issue whether humans can learn to improve detection accuracy. A crucial step is to establish whether detection accuracy can be improved through learning from experience or whether this is a fundamental limitation exceeding human perceptual capacities. Additionally, some people seem more vulnerable to this deception than others, suggesting that individual differences play a role in GAI detection. This introduces the possibility that people may differ in the extent in which they benefit from learning. However, a clear indication of who can learn to recognize GAI images remains undetermined. Exploring this is crucial in developing potential training methods to protect individuals from the dangers of GAI. This research adopts both a cognitive perspective and a perceptual perspective to examine the effect of learning from experience to improve the recognition of GAI images, questioning: To what extent does inductive learning improve GAI images recognition and is this influenced by socio-cognitive and visual processing skills? In the following section existing research on the detection of GAI will be discussed, followed by a theoretical framework of the socio-cognitive and visual processing skills.

Detection Accuracy

The core assumption of this research is that inductive learning facilitates detection of GAI images. Inductive learning is based on providing examples of specific observation, in this case GAI and human made images, allowing effective learning through abstracting similarities and differences

(Prince & Felder, 2006). Kornell and Bjork (2008) describe inductive learning as a process constantly allowing for learning and indicating that more exposure generally leads to better performance. In their study, they successfully used inductive learning paradigms to see if participants could recognize different artists that created a painting, indeed showing improved performance after a training session that facilitated rapid inductive learning. Thus, previous research does indicate that training might improve detection accuracy, however, no experiments exist testing the potential improvements of training. On the contrary, the detection of GAI images has been argued to exceed the limits of human perception, suggesting the need for AI fake-image detection techniques (Chai et al., 2020). Lu et al. (2023) compared the human capabilities to AI detection techniques, finding that humans misclassify 38.7% of the images, while detection techniques only miss 13% of the time. The human misclassification was only partly due to the inability to recognize GAI images, but also due to reduced trust in their own judgements, failing to recognize human-made images. Contrary to these earlier arguments, however, we argue that just as people can learn to recognize different painters (Kornell & Bjork, 2008), they can learn to recognize GAI models.

There is support for the notion that differences in experience with GAI appear to be relevant for detection ability. Lu et al. (2023) reported that experience with the knowledge of GAI or GAI creating showed a 0.7% increase in identification of non-GAI images and even a 3.7% increase for GAI images. Furthermore, people with a professional background of at least 2 years working in the visual field (e.g. design, photography, filmmaking), show better detection of GAI images (Velásquez-Salamanca et al., 2025). Contrary, the use of social media relates to more GAI text being undetected (Chein et al., 2024). An important difference to note is that social media does not necessarily disclose the source of the content, being exposed rather than potentially trained in recognition. Overall, these findings suggest improvement through exposure including examples of labelled GAI images. Therefore, inductive learning, as a general learning process based on labelled experiences, is likely to have benefits in improving recognition.

Perception of Images

Research on GAI image detection has demonstrated that human faces are better detected, indicating the influence of perception of emotional cues of images may facilitate GAI detection. A

large-scale survey by Roca et al. (2025) found that the human portraits had the highest overall success rate of 65%, while the average of GAI detection was 62%. Nature images scored the lowest, with 59%, although they did not explicitly mention whether this difference was significant. Even compared to AI detection techniques with images of different content, humans show improved performance on identifying the source of images containing humans (Lu et al., 2023). The importance of emotion recognition is in line with previous knowledge of evolution, as the priority was to recognize emotions to ensure (social) survival (Ferretti & Papaleo, 2019; Meletti, 2016). These findings highlight the relevance of human faces and emotional content in GAI image detection.

The perception and evaluation of GAI images have potential to influence the detection rate. The perceptions contain subtle differences from reality, which can invoke feelings of uncanniness which can aid in detection (Rapp et al., 2025). When the image is recognized as GAI content, the perception shifts based on the underlying evaluation and potential concerns a person has about GAI. Sun et al. (2022) found a difference in aesthetic appreciation, arguing that GAI art fails to capture a social significance from the limited semantic prompts. Moreover, there is a negative bias to GAI art that works bidirectionally (Grassini & Koivisto, 2024). Recognized GAI artworks are more likely to be evaluated worse, while when a human made artwork is evaluated less positively, it is more often determined to be a GAI artwork. These findings suggest that the perceptual and evaluative processes can interfere with accurate detection, beyond objective image characteristic alone.

Individual Differences

Individual differences contribute to the accurate detection of GAI, but this has not yet been explored with GAI images. Chein et al. (2024) looked into the differences between people in GAI text recognition. They found that people with higher fluid intelligence, such as abstract thinking and recognition of patterns, are better at distinguishing GAI from human-written text, while executive functioning and empathy did not show a significant correlation. These findings indicate that detection is not uniform across individuals and motivate exploration of underlying processes in detection abilities of GAI images.

To account for these differences, two explanatory perspectives can be distinguished. The first is a socio-cognitive perspective, relating to how humans interpret the world through processing social

information allowing them to understand others. The second is a perceptual processing perspective, involving the selection of stimuli and translating these to meaningful information to interpret. The following sections explore the contribution of socio-cognitive factors and visual processing factors with regards to the individual differences in GAI detection.

Theory of Mind

Theory of Mind (ToM) is a socio-cognitive theory, describing the understanding of other's minds such as emotions, intentions and perceptions of others (Carlson et al., 2013; Zimmer et al., 2025). In the context of GAI images, ToM may aid in the understanding of improved detection of human-likeness, emotional expression and the intention of the images. A study by Lee et al. (2014) found that ToM influences facial emotion recognition. ToM, combined with perception of emotion, influences deception detection (Stewart et al., 2019). Judgment of authenticity relies on evaluating whether it is intentional, coherent and socially meaningful. This is often lacking in the GAI images according to previous findings on perception of GAI as previously discussed (Rapp et al., 2025; Roca et al., 2025; Ting et al., 2023). Therefore, detecting subtle cues through ToM is a likely contributor to individual differences in detection accuracy of GAI images.

Beyond contributing to detection, ToM can be a potential benefit for inductive learning as it can provide better inferences and interpretation of emotional expression. This will help to identify intentional and mechanical patterns, similar to dual-process theories of how inductive processes are strongly influenced by reflective processing (Stephens et al., 2020). This creates the potential for a greater benefit from inductive learning to improve GAI detection in people with a higher score in ToM-based tests. A common measure for ToM are visual tests, such as the Reading the Mind in the Eyes Test (RMET), where a participant has to look at pictures of only eyes and ascribe an emotion most fitting to them (Olderbak et al., 2015). Crucially, this ability of inferring emotional states from minimal cues in ToM relies on visual information as well, suggesting not only socio-cognitive reasoning, but also a sensitivity to small visual irregularities.

Visual Disembedding

Visual disembedding is a perceptual processing mechanism referring to how individuals attend to visual information, specifically the ability to focus on distinctive images and ignore

distracting contextual information (Van der Hallen et al., 2015). Individuals with better visual disembedding skills are better at detecting relevant details and irregularities. In the context of GAI recognition, visual irregularities are assumed to be a key element of determining whether the image is unrealistic. Examples of the visual irregularities can be the amount of details and blurriness of the pictures (Lu et al., 2023). Individuals with better visual disembedding skills, therefore have potential to be better at detecting GAI images.

Visual disembedding does not rely on interpreting emotional or social information, making it a distinct perceptual process. Visual disembedding is commonly measured with the Embedded Figures Test (EFT), which captures the ability to extract simple forms from complex shapes, indicating a better ability to disembed information from context (F. Happé, 2013). Conceptually, this can be contrasted with the RMET, as both infer information from details in visual input, while the mechanism differs. Research on Autism Spectrum Disorder (ASD) illustrates that ToM and visual disembedding have opposing performance expectancies, namely reduced performance on ToM tasks and improved visual disembedding (Brewer et al., 2017; Cribb et al., 2016; Hutchins et al., 2012; Van der Hallen et al., 2015). Although this study focuses on a non-specific sample, this distinction highlights that socio-cognitive and perceptual processing skills can vary independently and in opposing directions. Therefore, the relation between socio-cognitive and perceptual processing mechanisms will be explored as potential underlying contributors for differences in GAI detection accuracy.

Hypothesis

This research examines to what extent inductive learning improves GAI image detection accuracy and whether this is associated with individual differences in socio-cognitive mechanisms or visual processing mechanisms. Specifically, ToM and visual disembedding are investigated as potential contributors to the accuracy of GAI detection. Based on this framework, the hypotheses are as follows:

1. Inductive learning hypothesis: inductive learning will improve participants' ability to detect GAI images.
2. ToM hypothesis: participants with higher ToM scores will perform better in detecting GAI.

3. ToM-Visual-disembedding-link hypothesis: ToM and visual disembedding will show a significant amount of shared variance in GAI image detection, indicating that part of the ToM effect in GAI image detection is explained by perceptual mechanisms, rather than purely socio-cognitive mechanisms. Controlling for visual-processing ability may significantly reduce the effect of ToM as noted in Hypothesis 2.
4. Interaction hypothesis ToM: the positive effect of inductive learning on GAI detection will be stronger for individuals with higher ToM scores, suggesting that ToM benefits abstracting similarities and differences in GAI images containing emotion.

Answering these questions will give concrete suggestions for the underlying mechanisms that explain individual differences in GAI detection and whether training can prevent the ethical dangers of undetected GAI images.

Methods

Participants

A total of *N*=292 participants completed the study. All participants had to be at least 16 years of age to participate in this study. Some participants were recruited through the SONA platform belonging to the University of Groningen (*N*= 242), consisting mostly of first-year psychology students, who received course credits for their participation. Other participants were recruited through the personal networks of the researchers (*N*= 50); these individuals did not receive any compensation for participating. All participants received an informed consent form, which they read and signed prior to participation in the study. This consent form explained their rights as participants in a research study and how their data would be handled.

To ensure data quality, a mandatory completion time was used. As the test was timed, longer duration meant a gap between the training and the test condition, therefore long durations likely reflect disengagement or interruption rather than deeper processing. Exclusions were made based on durations outside of the interquartile range (IQR) method. Participants that completed the study outside 1,5xIQR were excluded from analysis, as this is a robust method to detect extreme values. Based on these criteria, one participant was excluded for taking too little time and 24 participants

were excluded for taking too long to complete the study. After applying these criteria a total of *N*=267 participants were left in the final sample. Of these participants, *N*=148 were left in the control condition and *N*=119 participants remained in the training condition. The study was deemed low risk by the guidelines set up by the Ethics Committee and therefore exempt from full ethical review. Data collection commenced on November 6th of 2025 and ended on November 16th of 2025 under code PSY-2526-S-0030.

Design

This study used a mixed experimental design and was conducted online using the Qualtrics platform. The broader design included one between-subjects factor (training vs. control) and three within-subjects factors, namely facial emotion, attractiveness and image type. Participants were randomly assigned to either the experimental condition (training) or the control condition (no training). For this thesis, the variables of interest include the training condition, facial emotion (happy, angry and scared) and image type being either GAI or human made. Although attractiveness was part of the design, this is not explored in this research. The dependent variables measured were detection accuracy, theory of mind and visual disembedding. Confidence ratings were measured as well, but are not examined in this paper. Participants were asked to complete the study online, with the option to do it in English or Dutch. The data analyses were done with JASP (version 0.95.3).

Materials and Procedure

The experimental design was set up to examine whether inductive learning improves people's ability to distinguish GAI from human made portraits, and how individual differences in socio-cognitive and perceptual abilities affect their detection accuracy. All participants completed a test where they had to determine which picture was GAI or if it is human made. The human made pictures came from the Chicago database (Ma et al., 2021). This database was chosen as all faces were labelled and categorized for the emotions, gender, race and attractiveness, the latter based on a ranking by a large sample (*N*=100). All pictures in the database were unified, meaning the background and all jewellery was removed. The GAI images were generated through Midjourney version 7 (Midjourney, 2025) and Dall-e version 3 (OpenAI, 2025). These models were chosen because they are

easily accessible and are widely used so they would generate representable pictures. Example prompts that were used can be found in appendix A. The decision on which pictures to include was made on a joint agreement (*N*=6) based on photorealism, expressed emotion and not overly attractive. The pictures were matched to comparable images from the Chicago database, to relatively match the gender, race and attractiveness. Both AI models on occasion added jewellery or a coloured background. Using prompts to remove or alter the image did not result in similar pictures to the real images. Therefore the AI images were manually edited to remove the background and remove jewellery in a similar manner as the real photographs.

The test contained 48 unlabelled pictures, 24 GAI and 24 human made images. These were further divided into 12 males and 12 females, each with the 3 types of emotions. For each emotion, happy, angry and fear, there were 4 pictures, of which half are considered attractive and the other half are considered unattractive. In the inductive learning condition, there were a total of 96 images shown with a description of whether this is AI or a real image. In addition to the pictures, participant's confidence in their ability to classify images was assessed three times in the experimental condition and twice in the control condition, but will not be further discussed in this research as this was investigated by one of the collaborators.

Visual Processing Skills

The Leuven Embedded Figures Test (L-EFT) measures the ability to disembed information from context (de-Wit et al., 2017). A higher score on this test indicates better visual-processing skills in the disembedding of visual information. This test originally consists of 64 questions, but this would have caused concerns for the total length of the questionnaire. The L-EFT is a newer version of the original embedded figures test by Witkins, which has been validated in short form with 12 and even 6 questions (Jackson, 1956; Mumma, 1993). While this did not exist for the L-EFT, the data set including accuracy results was made publicly available by de-Wit et al. (2017). Based on this data of 255 participants, the 10 most difficult items were selected, as Schlooz and Hulstijn (2014) found that increased difficulty showed better increased sensitivity. In the research by de-Wit et al. (2017), items had an accuracy range of 0.49-0.70, with an average of 0.59. While the validation of the short version of the L-EFT has yet to be determined, the L-EFT is validated to test visual-perception without major

influences from broader cognitive abilities (Huygelier et al., 2018). We used the unvalidated, shorter version of the L-EFT to assess visual processing factors. The questionnaire first showed a figure at the top of the screen. The figure was presented together with three response options, among which only one contained the embedded figure. Participants were asked to choose the image in which they believed the figure appeared (M_{score} = 7.97; *SD* = 1.81).

Theory of Mind

The RMET (Baron-Cohen et al., 2001) was used to assess participants' ToM. The RMET is a visual test suited for an adult population. The RMET is the only test that does allow for differentiation between adults, without running into a ceiling effect (Baron-Cohen et al., 2001; Yeung et al., 2024). According to Chander et al. (2020), the original 36-item RMET can be condensed into a relatively accurate shorter version. The validity of the RMET has been debated, suggesting results should be interpreted more broadly than a direct link to ToM (Higgins et al., 2024). On the contrary, Olderbak et al. (2015) describes the short form as more reliable than the longer version. During the RMET, participants were presented with eight cropped facial images showing only the eye region. For each image, participants were asked to indicate the emotion they believed best matched the person's expression. Each photo included four possible emotion options from which they had to choose the correct one (M_{score} = 5.82; *SD* = 1.36).

Demographical characteristics

Participants were asked to indicate their gender, age group, and preferred language. The questionnaire was available in both Dutch and English. Gender was reported using the options *male, female, non-binary, other* or *prefer not to say.* Age was reported in the categories *18-24, 25-34, 35-44,* and so on, up to *65+*. Native language could be indicated through the options *English, Dutch, German, other* or *prefer not to say.* For all demographic questions, the opportunity to not disclose this type of information was included for ethical reasons.

Image Classification

For the classification of GAI or human made images we asked with every shown portrait: "This image is:" with two response options. "AI-generated" or "A real photo" with an allowed response time of 15 seconds. In the inductive learning condition, participants completed a training

session where they viewed a series of faces correctly labelled as GAI or human made. These were shown in an alternating pattern to help the participants notice the visual distinctions between both categories. Participants in the control condition did not receive this training and instead went straight to the test. After the survey participants had the opportunity to write any remark or feedback about the experiment if they wished to. Finally, they could see a message thanking their participation, which marked the end of the procedure.

Results

Table 1

Frequency Table of the Demographic within the Sample

Categorical Variables	Level	Frequency	Percentage
Age	18-24	228	85.4
	25-34	15	5.6
	35-44	4	1.5
	45-54	11	4.1
	55-64	3	1.1
	65+	6	2.2
Gender	Female	200	74.9
	Male	60	22.5
	Non-binary	5	1.9
	Prefer not to say	2	0.7
Language	English	110	41.2
	Dutch	157	58.8
Sona	Sona	223	83.5
	Other	44	16.5
Screen resolution	Above 512x512	202	75.7
	Below 512x512	65	24.3

Preliminary results

The analyses examined whether GAI image detection scores were improved in the inductive learning condition, in comparison to the control condition and whether ToM and visual processing contributed to detection performance. The number GAI images correctly identified were scored and converted to a detection accuracy percentage based on the total number of questions answered. This was done to account for people missing a question, as it cannot be determined whether they ran out of time or did not know the answer. The RMET and L-EFT scores were calculated by adding up the scores for each of the tests. The screen resolution of the device used varies between the participants, which is divided to below 512x512, where at least one side is below 512, and above 512x512, in which case both sides are higher than 512. The variation of the demographics in the sample can be seen in the frequency table (Table 1).

Table 2

Correlation Table

	Variables		1	2	3	4	5	6
1.	L-EFT	Pearson's r						
		p-values						
2.	RMET	Pearson's r	.107					
		p-values	.082					
3.	Accuracy (percentage)	Pearson's r	.033	.119				
		p-values	.596	.051				
4.	Accuracy (score)	Pearson's r	.046	.108	.982			
		p-values	.459	.104	<.001**			
5.	Accuracy Score AI	Pearson's r	.074	.089	0.824	.834		
		p-values	.225	.104	<.001**	<.001**		
6.	Accuracy Score Real	Pearson's r	.008	.083	.857	.877	.466	
		p-values	.895	.178	<.001**	<.001**	<.001**	
7.	Age	Pearson's r	-.037	-.034	-0.396	-.384	-0.420	-0.250
		p-values	.552	.585	<.001**	<.001**	<.001**	<.001**

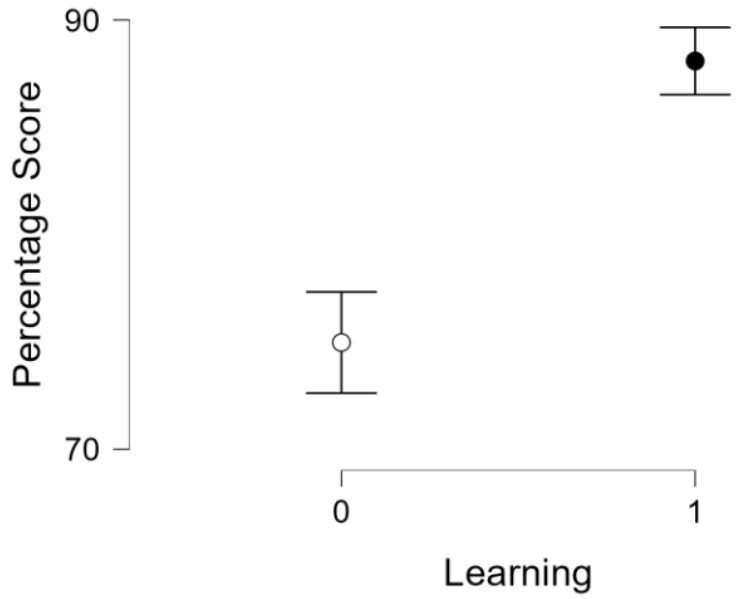
*Note: ** correlation is significant at .01 level (2-tailed)*

In table 2 the correlations for the L-EFT, RMET, the accuracy of GAI images and age are shown. The detection accuracy percentage and detection accuracy score are significantly correlated (r = .997, p <.001). No significant difference is found between detection accuracy as percentage and as score, the latter projecting accuracy in combination with time management. Percentage scores are therefore used in future analysis, as table 2 shows no significant difference between them. Age was explored as well, finding a significant negative correlation with percentage score (r = -.396, p <.001), but the unequal group sizes must be noted.

Hypothesis tests

Figure 2

Difference in detection accuracy percentage scores for the control (0) and learning condition (1)



Hypothesis 1: Effect of Learning

The first hypothesis explored whether learning improved GAI detection accuracy. An independent sample T-test was used to test whether the learning condition (M_{score} = 88.09; SD = 13.78) had higher detection accuracy percentage than the control group (M_{score} = 74.97; SD = 14.50), see figure 2. Shapiro-Wilk test indicated a violation of the normality, which was moderately violated (skewness = -1.28, kurtosis = 1.78), however the sample size was sufficiently large. The visual inspection of the QQ plot showed no extreme outliers. The Brown-Forsythe test indicated that the homogeneity of variance was violated. Therefore, Welch's t-test was used, which showed significant

differences in percentage of scores (t (245.6) = -9.172, p <.001) The effect size was large (d = -1.10), indicating that training increased detection accuracy substantially, supporting hypothesis 1.

Table 3

Effect Table Multiple Linear Regression with Learning and RMET

	b	SE b	t	p
Learning	13.63	1.49	9.14	<.001
RMET	1.66	1.66	3.04	.003

Hypothesis 2: Effect of Theory of Mind

The second hypothesis tested whether ToM, measured through RMET, predicted detection accuracy. The Pearson correlation between percentage score and RMET approached significance (r = .119; p = .051), suggesting a weak positive trend. Because learning conditions had a significant effect on detection accuracy as seen in hypothesis 1, a hierarchical multiple linear regression was conducted to examine whether RMET predicted the performance after accounting for learning conditions. Detection accuracy percentage was entered as the dependent variable, and learning condition and RMET as independent variables. Table 3 shows RMET as a significant predictor (b = 1.66, t(264) = 3.04, p = .003), suggesting that higher ToM scores were associated with better detection accuracy after adjusting for the learning condition. Overall, hypothesis 2 is partially supported, as the bivariate correlation between RMET and detection accuracy approached significance, while the relationship adjusting for learning condition was significant.

Table 4

Effect Table Multiple Linear Regression with Learning, RMET and L-EFT

	b	SE b	t	p
Learning	13.65	1.49	9.14	<.001
RMET	1.62	0.549	2.95	.003
L-EFT	0.26	0.411	0.62	.534

Table 5

Partial and Semi-partial Correlations

	Semi-partial *Model 2*	Semi-partial *Model 3*
Learning	0.487	0.488
RMET	0.162	0.158
L-EFT	-	0.033

Note: Model 2 includes Learning and RMET; Model 3 includes Learning, RMET and L-EFT

Hypothesis 3: Theory of Mind controlled for L-EFT

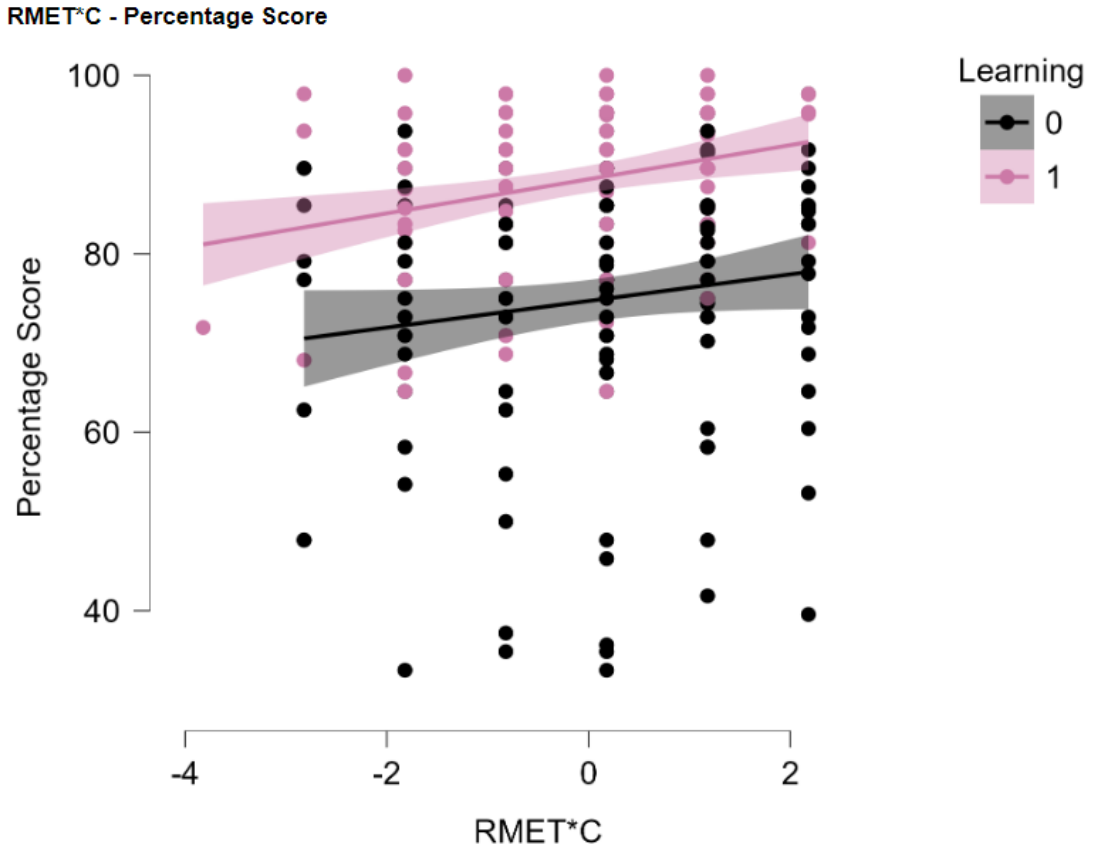
The third hypothesis tested the shared variance between the L-EFT and the RMET in GAI detection accuracy. The bivariate correlation of L-EFT and detection accuracy percentage is not significantly correlated (r = .033, p = .596). Contrary to expectations, visual disembedding did not contribute to GAI detection accuracy. To see whether visual processing played a role in the RMET scores, the L-EFT was added to the hierarchical multiple linear regression. As seen in table 4, L-EFT is not a significant predictor (b= .026, t(264) = 0.62, p = .534). The semi-partial correlations, shown in table 5, indicate that the variance in detection accuracy explained by L-EFT (sr = .033) is negligible. Adding L-EFT does not significantly decrease RMET semi-partial correlation (sr = .158). RMET remained a significant predictor for detection accuracy, whereas L-EFT did not explain additional variance and did not alter the relationship of ToM and GAI detection accuracy. Thus, no support for hypothesis 3 could be found.

Table 6

Coefficients Table Linear Regression

	b	SE	t	p
RMET centred	1.535	0.703	2.183	.030
Learning	13.556	1.493	9.089	<.001
Interaction	0.378	1.116	0.338	.735

Figure 3



Note: Control (0) and Learning (1)

Hypothesis 4: Interaction effect Theory of Mind

The fourth hypothesis tested whether ToM had an effect on the improvement of learning, a linear regression with the interaction between RMET and learning was done. While the ANOVA was significant (F (3, 263) = 29.21, p <.001), the interaction itself was not significant (b = 0.378, p = . 735), as can be seen in table 6. As shown in figure 3, the control and learning are parallel lines, indicating that ToM does not influence learning ability in the identification of GAI images. Thus, the fourth hypothesis for the interaction of learning and RMET as moderator was not supported.

Table 7

Multivariate Test of RMET on Emotion-Specific Detection Accuracy

	Value	F	df	p
Pillai's Trace	.01	0.89	3,263	.447

Exploratory analysis

In addition to the hypotheses stated before, an additional analysis is explored to see whether ToM influenced the detection accuracy for the three emotions differently. This potential difference is based on the innate concept of ToM being based on inferring cognitive or emotional states, creating the possibility that some emotions are better recognized than others. The scores for the three emotions happy, scared and angry were added up and the percentage was used. The RMET was not significantly correlated with the emotions separately (Happy $r = .083$, $p = .172$; Scared $r = .099$, $p = .104$; Angry $r = .079$; $p = .197$). A multivariate linear model was performed in SPSS (version 31.0.0), using the three dependent variables detection accuracy percentage for happy, scared and angry and the RMET score as the independent variable. The multivariate effect of RMET was not significant (Pillai's Trace = .01, $F(3, 263) = 0.89$, $p = .447$). This indicates that ToM was not differently associated with GAI image detection accuracy across the emotions happy, scared and angry.

Discussion

This research investigated whether humans can be trained to be better at GAI image detection through inductive learning and whether individual differences in socio-cognitive or visual processing skills contribute to this. The results show that learning was related to higher accuracy in detecting GAI images. The socio-cognitive construct ToM explained a smaller, but unique portion of variance after accounting for inductive learning. Visual disembedding did not play a role in the RMET and did not contribute to the detection accuracy. Lastly, no interaction was found between inductive learning and ToM. These findings highlight inductive learning as a main contributor, with ToM as a smaller, independent contributor. The following sections interpret these findings and consider the implications to counter the danger of the inability to detect GAI.

Summary of Results and Implications

Inductive Learning

Hypothesis 1 was supported as GAI image detection was improved based on inductive learning, demonstrating that humans can learn and improve to recognize GAI images. This is in line with Kornell and Bjork (2008) stating that inductive learning allows people to make distinctions more easily, which extends to the distinction between GAI and human-made images. This effect cannot be

attributed to mere exposure, as the images need to be labelled in order to provide feedback and allow the participant to make connections (Prince & Felder, 2006). This allows for extraction of features from the GAI image, rather than memorization of the image. As contemporary GAI has progressed far beyond the obvious visual flaws, these findings are particularly informative in showing that humans can still identify subtle differences if exposed to multiple labelled examples. In this sample, this low intensity training was effective in improving performance, suggesting practical implications for scalable interventions against the dangers of GAI.

Theory of Mind

Although the bivariate association of RMET and detection accuracy only partially supported hypothesis 2, support was found after accounting for the strong inductive learning effect. Specifically, ToM explained a unique part of the variation in detection accuracy, indicating a contribution of the socio-cognitive factors to GAI detection beyond the effects of training. Importantly, the distribution in the RMET scores in the sample is reliable with an average of 72.25% correct, which is in line with the known average for the general population of 72.78% correct (Baron-Cohen et al., 2001). This indicates that observed results are unlikely to be driven by an atypical sample distribution and reflect the intended socio-cognitive factors. Overall, these findings suggest that while inductive learning was the prominent predictor of GAI detection, ToM contributes independently to baseline detection sensitivity.

One possible explanation for this effect is that people scoring higher on RMET are more sensitive to a holistic social coherence that is lacking in GAI images. This interpretation aligns with the socially meaningful cues described by Carlson et al. (2013), where ToM allows people to rely on the context of a situation, which could translate to whether the image feels socially logical. Because ToM reflects a relatively stable socio-cognitive factor, this sensitivity becomes more apparent when accounting for the short-term improvement from inductive learning. This is consistent with the evidence of relatively stable individual differences in ToM over longer periods of time (Fernández-Abascal et al., 2013). Taken together, this account suggests ToM as a contributor to GAI detection through a sensitivity to social coherence of human faces, although alternative explanations should also be considered.

An alternative explanation for ToM being influential for GAI image detection, is that ToM facilitates social reasoning and understanding complex intentional decisions, such as deception (Frith & Frith, 2005). From this perspective, individuals with higher ToM scores may consider the underlying intention and manipulation of the GAI images. The GAI images were carefully selected by the researchers with the intent to simulate authentic human made images. The communicative purpose of the images might be recognized as an overarching theme, rather than exclusively considering the perceptive content of the image. Some participants describe the intention of mimicking facial imperfection such as skin texture and pimples, suggesting an awareness of the intent of the researchers. However, in this scenario one would suspect inductive learning as being influential by presenting more images to confirm the type of manipulation in the images, which was not the case. A possible direction to explore is epistemological vigilance (Sperber et al., 2010), related to individual differences in inferring hidden mental states to avoid deception. Future research could explore the connection between ToM and epistemological vigilance in relation to GAI detection. Currently, this remains a speculative interpretation, as it likely involves broader metacognitive processes that were not measured in this study. Therefore, the idea that ToM influences the interpretation of deceptive cues in GAI images, should only be considered as an addition to the socio-cognitive sensitivity described before.

These findings suggest that sensitivity to social coherence allows some individuals to be better at GAI detection. This highlights a potential weakness of GAI systems, which produces realistic individual features, but fails to present the social coherent facial expression present in human portraits. Additionally, socio-cognitive factors influence individuals ability to detect GAI images, which can be translated into risk factors. From a ToM perspective, the vulnerable groups to the dangers of GAI are young children and people with ASD or other neurodevelopmental disorders who are known to score lower on ToM tests (Hutchins et al., 2012; Korkmaz, 2011). Thus, ToM contributing to the recognition of GAI images seems to be based on a socio-cognitive sensitivity, which should be considered in determining vulnerabilities for the dangers of GAI.

Visual disembedding

Hypothesis 3 was not supported, as the visual disembedding measured by the L-EFT did not predict GAI detection accuracy, nor did it share variance with ToM as a predictor of GAI detection. This suggests that visual disembedding does not play a role in GAI detection. Because the RMET is a visual test, it was initially expected that visual processing factors would partially account for the observed association of ToM and GAI detection. The absence of this effect further supports that the socio-cognitive factors are the crucial contributor to GAI detection, rather than lower level visual processing skills. Based on these findings, ToM does not rely on visual disembedding, contributing to the idea that socio-cognitive factors are more relevant in GAI detection than visual processing factors.

Another possible explanation is that the L-EFT did not measure the intended visual processing relevant for GAI detection. Although the L-EFT should be a precise measure of visual disembedding, research has shown that it can be influenced by fluid intelligence and cognitive flexibility (Huygelier et al., 2018). Van der Hallen et al. (2015) found that the L-EFT depends most on line continuity and figure complexity, but lacks 3D depth cues and meaningful embedding contexts. In contrast, the GAI images included in this research are portraits including complexities such as lights and shadows, anatomical accuracy and a holistic view of the features in the face. As participants had the option to leave a comment, perceptual cues were often noted as suggestive of GAI content. The most common mentions were unrealistic light in the eyes, overly saturated skin colours and skin texture or wrinkles that seemed out of place. These detailed cues classify as perceptually complex, which is not measured with the L-EFT, suggesting visual disembedding in itself should not be discounted as an effect.

The absence of an effect of visual disembedding has implications for the interpretation of the RMET, suggesting that it measures something that is not related to simple visual processing. If visual processes influence the RMET, these are likely related to more complex processing related to a more holistic perceptual style, allowing to assess social coherence. For future research, a visual test more focused on the global-local perceptual styles has potential to further investigate the role of visual processing on both RMET and GAI image detection. Navon's paradigm, as described by Gerlach and Poirel (2018), could be used to explore whether perceptual styles focusing on the details or on the overall picture are beneficial to GAI image detection. However, based on the present findings of

visual disembedding measured by the L-EFT, there was no effect on GAI image detection or ToM measures.

Interaction Effect

Hypothesis 4 was not supported, as no interaction between inductive learning and ToM was found in predicting GAI detection accuracy. This indicates that the training overall was equally beneficial across participants, independent of the individual differences in ToM. Contrary to what was expected, ToM does not improve inductive learning. One possible interpretation is that the socio-cognitive processes are not primarily engaged in inductive learning. Instead, the training relied on labelled examples, allowing for detection through similarity-based discrimination (Kornell & Bjork, 2008). In the context of GAI, this likely relied on perceptual irregularities between the GAI and human made image, which appear to be separate from socio-cognitive processes. Importantly, individuals scoring higher on ToM still benefited equally from the inductive training, suggesting that the benefits from inductive learning are additive to the baseline socio-cognitive benefits. From a practical perspective, this indicates that inductive learning improves GAI detection broadly and does not require adaptation to individual differences in socio-cognitive factors. Additionally, this suggests that the vulnerable groups described earlier could benefit from inductive learning to an equal extent. Overall, absence of the interaction suggests that ToM is an independent contributor to GAI detection, adding to the larger improvements in GAI detection made by inductive learning.

Exploratory Findings

Lastly, exploratory analysis revealed that there is no significant difference in the association of ToM and GAI detection with different emotions. A possible reason could be that the socio-cognitive process aiding in recognition does not depend on the emotions. Not finding a difference is not surprising, as the GAI images in this study are strong expressions without ambiguity. ToM focuses more on subtle differences between emotions or socially complex cues (Lee et al., 2014). Future research could expand on this research by using emotionally ambiguous stimuli or contextualized scenes to explore this connection. Based on the current results, no emotion-specific differences in the relationship between ToM and GAI detection was found.

Strengths, limitations and suggestions for future research

A key strength of this study is the highly standardized design, allowing for exploration of underlying factors in GAI detection, while necessarily introducing limitations regarding generalizability. The GAI portraits were selected to have similar backgrounds, clothing and pose. This increased internal validity, ensuring that the emotional face is the distinguishing factor. However, the external validity is reduced, as in real life the images will contain more context, which might influence the GAI detection rate. The high level of standardization may have contributed to the relatively strong learning effect, aiding in the detection of commonalities between the human made images, which originated from the same database. Additionally, some potential noise was introduced due to using two different AI models. The latter was done to ensure the faces, especially in the fear condition, would vary enough and increase realism. The GAI images were generated and selected by the researchers, introducing potential bias, but ensuring qualitative images that allowed for a challenging test. Overall, this design allowed for a precise examination of GAI detection, while also highlighting the need to interpret the findings within the context of this study.

Given the global relevance of accurately detecting GAI images, an important limitation is the cultural generalizability. While the images were generated keeping diversity in mind, our sample was predominantly White university students. Cultural differences in socio-cognitive factors may influence how facial cues are interpreted, as suggested by the social orientation hypothesis describing a more holistic view in Asian countries (Varnum et al., 2010). This hypothesis suggests differences in socio-cognitive factors, making it likely that GAI detection could be influenced by cultural differences. This highlights that, while relevant to understanding GAI detection, the findings should be interpreted with caution when extending to culturally diverse populations.

The decision to have the survey be filled out online was another trade-off between internal and external validity, in which the latter was prioritized. The online environment can potentially reduce optimization in attention and motivation. However, this allowed for a large sample and more participants. By letting all participants use the device they are accustomed to, external validity was increased, which was further collaborated by the fact that the quality of the device was checked in the

analysis and did not seem to have an effect on detection accuracy. Overall, the online design allowed for better external validity while maintaining sufficient robustness.

Another trade-off that had to be considered was the length of the individual measures in relation to the length of the survey. In order to avoid compromising the attention span of the GAI training and test, short versions of the RMET and L-EFT were used. While both short-versions of the test were not validated, other research has included shortened versions, such as Chander et al. (2020) and Mumma (1993). Unfortunately, tests for ToM are limited, as most tests are only designed for children or adults with a neurodevelopmental disorder (Byom & Mutlu, 2013; Hutchins et al., 2012). There is a lack of ToM tests without a ceiling effect that do not rely on visual cues. Further research in a clinical population with ASD could be insightful, due to the expectancy to have relatively high scores on the L-EFT, with low scores on the RMET or other ToM tests (Brewer et al., 2017; Cribb et al., 2016; Hutchins et al., 2012; Van der Hallen et al., 2015). Using a verbal test for ToM, such as the Strange Stories Test by Happé (1994), would give insight into the core workings of how ToM contributes to GAI image detection. Looking into a clinical population, especially with a non-visual ToM test, would allow further exploration into whether the differences between socio-cognitive and visual processing skills influence GAI image detection.

While we found a clear effect that the learning was beneficial, it is unclear how long this result may last. While it is impossible to make an estimation over the duration of the effect of learning as this is influenced by multiple factors, previous findings about inductive learning and inductive reasoning describe longer lasting results (Klauer & Phye, 2008; Tomic & Klauer, 1996). Future research could look into a longitudinal study to see whether this inductive learning effect remains over a longer period of time. It would also be interesting to see the development of GAI over this period of time, potentially incorporating multiple trials from images generated at the start of the study and at the end of the study. While there are many directions on where to take GAI research, the main focus should be on seeing whether inductive learning can create a reliable long-term training to help against the dangers of GAI.

In conclusion, this research showed that inductive learning improves the detection of GAI images and that socio-cognitive skills used in ToM were associated with higher detection accuracy

independent of training. The visual disembedding was not linked to improved detection accuracy, indicating that this form of visual processing skills may be less relevant for GAI detection. While the dangers of the inability to recognize GAI images persist, these findings offer a promising potential in training to prepare society to better recognize reality.

References

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The 'Reading the Mind in the Eyes' Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *42*(2), 241–251.

Brewer, N., Young, R. L., & Barnett, E. (2017). Measuring Theory of Mind in Adults with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *47*(7), 1927–1941. https://doi.org/10.1007/s10803-017-3080-x

Byom, L. J., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience*, *7*, 413. https://doi.org/10.3389/fnhum.2013.00413

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT* (No. arXiv:2303.04226). arXiv. https://doi.org/10.48550/arXiv.2303.04226

Carlson, S. M., Koenig, M. A., & Harms, M. B. (2013). Theory of mind. *WIREs Cognitive Science*, *4*(4), 391–402. https://doi.org/10.1002/wcs.1232

Chai, L., Bau, D., Lim, S.-N., & Isola, P. (2020). *What makes fake images detectable? Understanding properties that generalize* (No. arXiv:2008.10588). arXiv. https://doi.org/10.48550/arXiv.2008.10588

Chander, R. J., Grainger, S. A., Crawford, J. D., Mather, K. A., Numbers, K., Cleary, R., Kochan, N. A., Brodaty, H., Henry, J. D., & Sachdev, P. S. (2020). Development of a short-form version of the Reading the Mind in the Eyes Test for assessing theory of mind in older adults. *International Journal of Geriatric Psychiatry*, *35*(11), 1322–1330. https://doi.org/10.1002/gps.5369

Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence: Individual differences in the discernment of human from AI texts. *Scientific Reports*, *14*(1), 25989. https://doi.org/10.1038/s41598-024-76218-y

Cribb, S. J., Olaithe, M., Di Lorenzo, R., Dunlop, P. D., & Maybery, M. T. (2016). Embedded Figures Test Performance in the Broader Autism Phenotype: A Meta-Analysis. *Journal of Autism and*

Developmental Disorders, *46*(9), 2924–2939. ERIC. https://doi.org/10.1007/s10803-016-2832-3

de-Wit, L., Huygelier, H., Van der Hallen, R., Chamberlain, R., & Wagemans, J. (2017). Developing the Leuven Embedded Figures Test (L-EFT): Testing the stimulus features that influence embedding. *PeerJ*, *5*, e2862. https://doi.org/10.7717/peerj.2862

Fernández-Abascal, E. G., Cabello, R., Fernández-Berrocal, P., & Baron-Cohen, S. (2013). Test-retest reliability of the 'Reading the Mind in the Eyes' test: A one-year follow-up study. *Molecular Autism*, *4*(1), 33. https://doi.org/10.1186/2040-2392-4-33

Ferretti, V., & Papaleo, F. (2019). Understanding others: Emotion recognition in humans and other animals. *Genes, Brain and Behavior*, *18*(1), e12544. https://doi.org/10.1111/gbb.12544

Frith, C., & Frith, U. (2005). Theory of mind. *Current Biology*, *15*(17), R644–R645. https://doi.org/10.1016/j.cub.2005.08.041

Gerlach, C., & Poirel, N. (2018). Navon's classical paradigm concerning local and global processing relates systematically to visual object classification performance. *Scientific Reports*, *8*(1), 324. https://doi.org/10.1038/s41598-017-18664-5

Grassini, S., & Koivisto, M. (2024). Understanding how personality traits, experiences, and attitudes shape negative bias toward AI-generated artworks. *Scientific Reports*, *14*(1), 4113. https://doi.org/10.1038/s41598-024-54294-4

Haan, W. (2023, August 25). *Hoe een AI-gegenereerde foto in het NOS Journaal terechtkwam* [News]. NOS. https://over.nos.nl/nieuws/hoe-een-ai-gegenereerde-foto-in-het-nos-journaal-terecht-kwam/

Happé, F. (2013). Embedded Figures Test (EFT). In *Encyclopedia of Autism Spectrum Disorders* (pp. 1077–1078). Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1698-3_1726

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154. https://doi.org/10.1007/BF02172093

Higgins, W. C., Kaplan, D. M., Deschrijver, E., & Ross, R. M. (2024). Construct validity evidence reporting practices for the Reading the Mind in the Eyes Test: A systematic scoping review. *Clinical Psychology Review*, *108*, 102378. https://doi.org/10.1016/j.cpr.2023.102378

Hutchins, T. L., Prelock, P. A., & Bonazinga, L. (2012). Psychometric Evaluation of the Theory of Mind Inventory (ToMI): A Study of Typically Developing Children and Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *42*(3), 327–341. https://doi.org/10.1007/s10803-011-1244-7

Huygelier, H., Hallen, R. V. der, Wagemans, J., de-Wit, L., & Chamberlain, R. (2018). The Leuven Embedded Figures Test (L-EFT): Measuring perception, intelligence or executive function? *PeerJ*, *6*, e4524. https://doi.org/10.7717/peerj.4524

Islam, M. R. (2024). *Generative AI, Cybersecurity, and Ethics*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/rug/detail.action?docID=31805206

Jackson, D. N. (1956). A short form of Witkin's embedded-figures test. *The Journal of Abnormal and Social Psychology*, *53*(2), 254–255. https://doi.org/10.1037/h0043845

Klauer, K. J., & Phye, G. D. (2008). Inductive Reasoning: A Training Approach. *Review of Educational Research*, *78*(1), 85–123. https://doi.org/10.3102/0034654307313402

Korkmaz, B. (2011). Theory of Mind and Neurodevelopmental Disorders of Childhood. *Pediatric Research*, *69*(8), 101–108. https://doi.org/10.1203/PDR.0b013e318212c177

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the 'enemy of induction'? *Psychological Science*, *19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Lee, C. H. (2025, September 25). Can people still tell real photos from AI images in 2025? - Conjointly. *Conjointly*. https://conjointly.com/blog/real-vs-ai-images-2025/

Lee, S. B., Koo, S. J., Song, Y. Y., Lee, M. K., Jeong, Y.-J., Kwon, C., Park, K. R., Park, J. Y., Kang, J. I., Lee, E., & An, S. K. (2014). Theory of Mind as a Mediator of Reasoning and Facial Emotion Recognition: Findings from 200 Healthy People. *Psychiatry Investigation*, *11*(2), 105–111. https://doi.org/10.4306/pi.2014.11.2.105

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., & Ouyang, W. (2023). *Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images* (No. arXiv:2304.13023). arXiv. https://doi.org/10.48550/arXiv.2304.13023

Ma, D. S., Kantner, J., & Wittenbrink, B. (2021). Chicago Face Database: Multiracial expansion. *Behavior Research Methods*, *53*(3), 1289–1300. https://doi.org/10.3758/s13428-020-01482-5

Meletti, S. (2016). Emotion Recognition. In M. Mula (Ed.), *Neuropsychiatric Symptoms of Epilepsy* (pp. 177–193). Springer International Publishing. https://doi.org/10.1007/978-3-319-22159-5_11

Mumma, G. H. (1993). The embedded figures test: Internal structure and development of a short form. *Personality and Individual Differences*, *15*(2), 221–224. https://doi.org/10.1016/0191-8869(93)90029-3

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01503

Prince, M. J., & Felder, R. M. (2006). Inductive Teaching and Learning Methods: Definitions, Comparisons, and Research Bases. *Journal of Engineering Education*, *95*(2), 123–138. https://doi.org/10.1002/j.2168-9830.2006.tb00884.x

Rapp, A., Di Lodovico, C., Torrielli, F., & Di Caro, L. (2025). How do people experience the images created by generative artificial intelligence? An exploration of people's perceptions, appraisals, and emotions related to a Gen-AI text-to-image model and its creations. *International Journal of Human-Computer Studies*, *193*, 103375. https://doi.org/10.1016/j.ijhcs.2024.103375

Roca, T., Roman, A., Vega, J., Duarte, M., Wang, P., White, K., Misra, A., & Lavista Ferres, J. (2025). *How good are humans at detecting AI-generated images? Learnings from an experiment*. https://doi.org/10.48550/arXiv.2507.18640

Romero-Moreno, F. (2025). Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Computer Law & Security Review*, *58*, 106162. https://doi.org/10.1016/j.clsr.2025.106162

Schlooz, W. A. J. M., & Hulstijn, W. (2014). Boys with autism spectrum disorders show superior performance on the adult Embedded Figures Test. *Research in Autism Spectrum Disorders*, *8*(1), 1–7. https://doi.org/10.1016/j.rasd.2013.10.004

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Stephens, R., Dunn, J., Hayes, B., & Kalish, M. (2020). A test of two processes: The effect of training on deductive and inductive reasoning. *Cognition*, *199*, 104223. https://doi.org/10.1016/j.cognition.2020.104223

Stewart, S. L. K., Wright, C., & Atherton, C. (2019). Deception Detection and Truth Detection Are Dependent on Different Cognitive and Emotional Traits: An Investigation of Emotional Intelligence, Theory of Mind, and Attention. *Personality and Social Psychology Bulletin*, *45*(5), 794–807. https://doi.org/10.1177/0146167218796795

Sun, Y., Yang, C.-H., Lyu, Y., & Lin, R. (2022). From Pigments to Pixels: A Comparison of Human and AI Painting. *Applied Sciences*, *12*(8), 3724. https://doi.org/10.3390/app12083724

Ting, T. T., Ling, L. Y., Azam, A. I. B. A., & Palaniappan, R. (2023). Artificial intelligence art: Attitudes and perceptions toward human versus artificial intelligence artworks. *AIP Conference Proceedings*, *2823*(1), 020003. https://doi.org/10.1063/5.0162434

Tomic, W., & Klauer, K. J. (1996). On the effects of training inductive reasoning: How far does it transfer and how long do the effects persist? *European Journal of Psychology of Education*, *11*(3), 283–299. https://doi.org/10.1007/BF03172941

Tredinnick, L., & Laybats, C. (2023). The dangers of generative artificial intelligence. *Business Information Review*, *40*(2), 46–48. https://doi.org/10.1177/02663821231183756

Van der Hallen, R., Chamberlain, R., de-Wit, L., & Wagemans, J. (2015). The Leuven Embedded Figures Test (L-EFT): Re-embedding the EFT into vision sciences. *Journal of Vision*, *15*, 336. https://doi.org/10.1167/15.12.336

Varnum, M. E. W., Grossmann, I., Kitayama, S., & Nisbett, R. E. (2010). The Origin of Cultural Differences in Cognition: Evidence for the Social Orientation Hypothesis. *Current Directions in Psychological Science*, *19*(1), 9–13. https://doi.org/10.1177/0963721409359301

Velásquez-Salamanca, D., Martín-Pascual, M. Á., & Andreu-Sánchez, C. (2025). Interpretation of AI-Generated vs. Human-Made Images. *Journal of Imaging*, *11*(7), 227. https://doi.org/10.3390/jimaging11070227

Yeung, E. K. L., Apperly, I. A., & Devine, R. T. (2024). Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews*, *157*, 105481. https://doi.org/10.1016/j.neubiorev.2023.105481

Zimmer, L., Richardson, H., Pletti, C., Paulus, M., & Schuwerk, T. (2025). Predictive responses in the Theory of Mind network: A comparison of autistic and non-autistic adults. *Cortex, 187*, 159–171. https://doi.org/10.1016/j.cortex.2025.04.006

Appendix A

Examples of Used Prompts

Females

Angry (Attractive)

Make a realistic photo of a 40 year old white woman. She has an angry expression, wearing a grey t-shirt (round neck) against a pure white background. Natural skin texture, minimal makeup, unstyled hair, and realistic lighting.

Angry (Unattractive)

Make a frontal photograph of an ugly-looking black woman (45 years old) with an irritated expression, wearing a grey t-shirt (round neck) against a white background. Uneven skin texture, tired eyes, unstyled long hair, no makeup, realistic lighting and colours, round, chubby face.

Happy (Attractive)

Make a front-facing photograph of a middle aged attractive black woman with a happy expression (closed mouth), wearing a grey t-shirt (round neck) against a white background. Minimal make-up, no jewellery and realistic lighting.

Happy (Unattractive)

Create a frontal photograph of an ugly-looking young black woman with a happy expression (open mouth), wearing a grey t-shirt with a round neck against a completely white background. Her face must be round and full, she should not be wearing make-up and her teeth should be a little crooked.

Fear (Attractive)

Create a picture of an attractive middle-aged white woman which could be used for a database for emotions. The picture should have a white background and she should be

wearing a grey shirt with a round neck. She should look scared, but realistically. She should have her hair up and have no jewellery. Her shirt and shoulders should be visible.

Fear (Unattractive)

Make a picture of an unattractive adolescent white woman with acne (23 years old) who is considered ugly exhibiting the emotion fear, she has to be terrified. Completely white background. Gray shirt (round neck). Should not be too close up (shirt and shoulders visible). Realistic human with facial asymmetry, and add some blemishes in the skin. Not too pretty looking. Not too much light and reduce shininess on the face.

Males

Angry (Attractive)

Make a realistic photo of a 40 year old white man. He has an angry expression, wearing a grey t-shirt with a round neck against a pure white background. Natural skin texture and realistic lighting.

Angry (Unattractive)

Create a frontal photograph of an unattractive adolescent black man with an angry/irritated expression, wearing a grey t-shirt with a round neck against a completely white background. He is overweight and has skin blemishes. Shirt and shoulders should be visible.

Happy (Attractive)

Make a front-facing picture of an attractive white man (25 years old) who is smiling (mouth closed) exhibiting the emotion of being happy. Emotion recognition task theory of mind database. Grey shirt with a white neck and pure white background. Not too zoomed in, his shirt and shoulders should be visible. No additional shine on the face.

Happy (Unattractive)

Produce an image of an ugly middle-aged white man who is happy (smiling with mouth open). He should have facial asymmetry, eye bags, skin blemishes, a grey shirt with a round neck and a fringe. The background needs to be white. Should be indistinguishable from reality.

Fear (Attractive)

Image of an attractive black man (30 years old) looking scared against a white background. The picture should be usable in a database for emotion recognition. Gray shirt with a rounded neck. Shoulders should be visible. Realistic lightning and reduce shininess on the face.

Fear (Unattractive)

Make a front facing image of a black man (20 years old) who is considered ugly and unattractive. He should be looking frightened. Give him facial asymmetry and acne. He should be wearing a grey shirt with a round neck and his shoulders should be in the picture. Do not add any jewellery or additional facial shininess.

Example GAI Images Used in the Study

Appendix B

AI Use Summary

Introduction

- AI system: ChatGPT
- Final prompts used: “How to test whether the visual abilities influence ToM measures, avoiding the ceiling effect” & “Existing tests with 2 modalities that could be used to test either ToM or visual processes”
- Use case: To brainstorm about topics to include and enhance flow and reduce overcomplication by introducing too many topics, this led to the inclusion of EFT as a measure and the removal of the idea of testing 2 modalities.
- Modifications: I did not use any AI text, everything is written by me.