The Role of IT Background for Metacognitive Accuracy, Confidence and Overestimation of Deep Fake Recognition Skills

Article *in* Lecture Notes in Computer Science · February 2022

9 authors, including:

Stefan Sütterlin
Hochschule Albstadt-Sigmaringen
128 PUBLICATIONS **1,850** CITATIONS
SEE PROFILE

Ric Lugo
Norwegian University of Science and Technology
74 PUBLICATIONS **396** CITATIONS
SEE PROFILE

Torvald Ask
Norwegian University of Science and Technology
18 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Basil Bärreiter
Hochschule Albstadt-Sigmaringen
3 PUBLICATIONS **3** CITATIONS
SEE PROFILE

The Role of IT Background for Metacognitive Accuracy, Confidence and Overestimation of Deep Fake Recognition Skills

Stefan Sütterlin[1,2], Ricardo G. Lugo[2,3], Torvald F. Ask[2,3], Karl Veng[1], Jonathan Eck[1], Jonas Fritschi[1], Muhammed-Talha Özmen[1], Basil Bärreiter[1], Benjamin J. Knox[2,3,4].

[1] Faculty of Computer Science, Albstadt-Sigmaringen University, Germany
[2] Faculty of Health, Welfare and Organisation, Østfold University College, Norway
[3] Centre for Cyber and Information Security, Norwegian University of Science and Technology, Norway
[4] Cyber Warfare Centre, Norwegian Armed Forces Cyber Defence, Norway
stefan.suetterlin@hs-albsig.de

Abstract. The emergence of synthetic media such as deep fakes is considered to be a disruptive technology shaping the fight against cybercrime as well as enabling political disinformation. Deep faked material exploits humans' interpersonal trust and is usually applied where technical solutions of deep fake authentication are not in place, unknown, or unaffordable. Improving the individual's ability to recognise deep fakes where they are not perfectly produced requires training and the incorporation of deep fake-based attacks into social engineering resilience training. Individualised or tailored approaches as part of cybersecurity awareness campaigns are superior to a one-size-fits-all approach, and need to identify persons in particular need for improvement. Research conducted in phishing simulations reported that persons with educational and/or professional background in information technology frequently underperform in social engineering simulations. In this study, we propose a method and metric to detect overconfident individuals in regards to deep fake recognition. The proposed overconfidence score flags individuals overestimating their performance and thus posing a previously unconsidered cybersecurity risk. In this study, and in line with comparable research from phishing simulations, individuals with IT background were particularly prone to overconfidence. We argue that this data-driven approach to identifying persons at risk enables educators to provide a more targeted education, evoke insight into own judgement deficiencies, and help to avoid the self-selection bias typical for voluntary participation.

Keywords. Metacognition; Deep Fake; Cybersecurity; Social Engineering

1 Introduction

In times of ever increasing Cyber Security (CS) threats and evolving CS defence technology, the resource-intensive arm's race between those intending to exploit technological weaknesses and those dedicated to protecting them, often leads to attack vectors circumventing technological defence structures by applying social engineering techniques. According to recent surveys, up to 98% of all cyberattacks in 2020 were social-engineering-enabled [1], with increasing tendencies [2, 3]. The term Social Engineering has been defined as "any act that influences a person to take an action that may or may not be in their best interest" [4, 5]. Cybercriminals consider the exploitation of stable human traits such as interpersonal trust, agreeableness and conscientiousness [6] as cheap, sufficiently reliable and sufficiently riskless to exploit. They do this by applying well established tactics of persuasion [7], such as social proof or reciprocity, depending on the target [8].

Persuasion tactics often aim to induce an emotional reaction (e.g. a sense of urgency, relatedness, compassion, sadness, fear, responsibility or duty) in the target. This will make them direct their attention towards internal processes such as the significance of emotional reactions or gut-feelings that lead to impulsive decision-making [9]. When attention is directed inwards, less attentional resources are directed externally to critically examine the details of the persuasion attempt, such as whether the information that is presented is plausible and warrants immediate action. One such example could be a cybercriminal pretending to be a relative or the superior of a target person while contacting them through a fake email or a fake profile on social media. The cybercriminal might be claiming that they are in trouble and that there is an urgent need for the target to transfer money from the company account or their own, to an account provided by the cybercriminal. If the cybercriminal is successful in instilling a sense of urgency in the target then the target may act without considering the scenario's actual probability. With the vast majority of cybercrime entailing some type of Social Engineering, and with political messaging relying on the spread of fake news that is often emotionally-laden, new and potentially disruptive technologies such as synthetic media (e.g. deep fakes (DF)) have raised attention both in the forensic as well as the political domain [10].

DF technology provides cybercriminals with a unique opportunity to impersonate other individuals with a degree of realism that is hard to falsify with human eyes [11, 12]. In a Social Engineering context, this is (broadly) achieved through 1) manipulating the facial expressions, lip movements, and voice of an individual that is known to the target by using DF technology on an existing video, 2) recording oneself while swapping faces with an individual that is known to the target, or 3) some combination of the two methods. In other words, DFs can be used in a Social Engineering attack through manipulation of facial expression or facial identity [13, 14]. The associated cost of DF scams was estimated to exceed 250 million USD in 2020 [15]. With the continuously increasing sophistication of AI-generated DF technology, the defence technology required to effectively detect and flag DFs may already be too advanced to be used by individuals from non-technical backgrounds [12, 16]. Social Engineering is more effective when aimed at unprepared and unaware individuals thus making them richer targets for Social Engineering attacks. Targeting unsuspecting victims with low levels of awareness of technological aids may effectively render technological solutions irrelevant in contexts where they are unknown, unavailable, or not likely to be applied by the user.

The lack of technological solutions in settings where DFs are most likely to occur highlights an urgent demand for education of users across societal settings and should therefore be integrated in future CS awareness training programs. It has been argued that, just as in other educational contexts, CS awareness training must follow individualised approaches to be most efficient [17, 18]. This tailored approach is required to avoid that participants are unmotivated due to inappropriate demands and difficulty levels. Comprehensive research efforts to map individual and cognitive-emotional factors affecting susceptibility to- and resilience against Social Engineering attacks have only recently started to emerge [19]. To successfully incorporate Social Engineering and DF detection skills in individualised CS awareness training, research must first address the current unanswered questions related to understanding how individual differences and related cognitive processes influence Social Engineering and particularly DF detection skills. The answers to these questions will be the basis for developing individualised approaches targeting cognitive processes on various levels. We argue that understanding trait-like precursors of underperformance in DF recognition can inform educational programs in CS awareness by identifying persons in need of feedback and particular attention. Consequently, these precursors may complement or to some extent, even substitute self-reported perceived competence and the self-selection to training schemes.

Metacognition is the cognitive ability to observe (or be aware of) one's own internal processes, such as thoughts and emotions, through two processes: (1) having knowledge about cognition and (2) how to regulate cognition [20]. Metacognitive awareness and accuracy vary between individuals, and may be a common denominator in both an individual's ability to be aware of- and think critically about their emotional and cognitive reactions during a Social Engineering attack. As well as possessing the requisite amount of self-knowledge about how accurately they are able to evaluate their ability to detect DFs. Previous research shows that having a technical or non-technical professional background does not differentiate susceptibility to phishing attacks. Even when having relevant education and coming from an Information Technology (IT) background, people are still overconfident in their abilities to detect Social Engineering attacks such as phishing emails [21, 22]. The non-differential ability of individuals with and without IT backgrounds to detect Social Engineering attacks may in part be explained by a lack of cognitive involvement, influenced by overconfidence rather than lacking detection abilities [22]. This matters because decisions in risky and uncertain situations are made based on a self-assessment of one's own perceived mastery, and not on actual performance which can only be known after the action has been performed and where there is the chance to receive an objective feedback on the outcomes.

There is a growing field of CS research on human-machine and human-human interactions in cyber-physical contexts (in the Hybrid Space) [23] aiming to improve decision-making that is based on how humans perceive and communicate their awareness of cyber threats. This research indicates that self-regulation and metacognitive skills are predictors of performance in such hybrid contexts [24-26]. While metacognitive awareness may allow individuals to detect when they are having emotional and cognitive reactions to DF-mediated Social Engineering attacks, self-regulation skills may allow individuals to divert attention away from the internal processes and towards critical examination of details such as DF artefacts. An individual with good metacognitive skills that knows their limitations may be more prone to seek out assistance or relevant information rather than making decisions based on an insufficient information basis. Conversely, people with lower levels of metacognition may not accurately judge their abilities to detect DF-mediated Social Engineering attacks, which may lead to overconfidence, thus being unprepared and at increased risk for victimisation. While metacognition has been researched extensively for performance in a cyber context in recent years, few studies have looked at metacognition (the confidence into one's performance and into the accuracy of one's self-assessment) in Social Engineering attacks. They find that participants can accurately judge legitimate emails, but the same participants displayed overconfidence in assessing phishing emails [27]. Accurate confidence judgments were only weakly associated with phishing identification [28].

This study investigated the role of metacognition in identifying DF social engineering attacks. We suggest the calculation of an overconfidence score may allow for the elimination of self-selection effects to cybersecurity awareness education, to ensure that training is delivered where it is needed most, and not only to those that are least confident. We hypothesise that overconfidence can have detrimental effects on DF recognition and that overconfidence is particularly pronounced in persons with IT-Backgrounds and self-reported IT-affinity.

- Hypothesis 1 (H_1): IT background and/or self-reported IT-affinity are not associated with higher DF recognition skills.
- Hypothesis 2 (H_2): IT background and/or self-reported IT-affinity are associated with higher *perceived* DF recognition skills.

2 Methods

2.1 Design

Data collection was done via the online platform Google Forms. All questions and instructions were presented in English. On the first page of the form, participants received instructions about the nature and purpose of the study. Upon study commencement, demographic information regarding participant characteristics were collected. Participants were asked to indicate whether they were using a PC, tablet, or smartphone, as well as their age, gender, country of residence, and level of educational attainment ranging from not having attended any education up to doctoral-level education. To collect information about IT and non-IT backgrounds, participants were asked to indicate whether they currently or previously worked in IT, or had received any formal IT education, and to indicate whether they previously or currently worked in CS/IT security. Participants also answered questions regarding their affinity to technology, prior DF knowledge and experience, and were asked to estimate their DF recognition skill level, the confidence they had in their skill level estimates, and how well they would perform on a DF recognition task. They were also asked to provide a number from 0 to 10 of how many clips they expected to correctly classify if shown 10 short clips that were either authentic or fake. Participants were also tested in the Group Embedded Figures Test which will be reported elsewhere. After all baseline information was collected, participants were taken to the task page in the online form. On the task page they were presented with 21 short videos where six of the videos were authentic and 15 of the videos were DFs. All clips were on average 15 seconds long and were retrieved from publicly available databases (github.com; kaggle.com). Participants were instructed to view each clip only once and DF videos were presented first. For each clip they rated if the video was authentic or faked. For each clip they rated as fake or authentic they were also asked to rate how certain they were about their individual judgements. After they finished rating the videos they were asked again to rate their DF recognition skills and how confident they were in their opinion about their skills. After finishing the study, participants were given the opportunity to express their opinions about how it was to participate, and also provide feedback for improvement. Participants were given a maximum of 55 minutes to complete the study including answering the questionnaires. A timer showing how much time they had left to complete the form was visible at the top of the page at all times. The average response time was approximately 20 minutes.

2.2 Questionnaires

Affinity for Technology Interaction (ATI) Scale. The ATI scale [29] was used to assess individual differences in affinity for interacting with technology. The ATI scale presents participants with nine statements about their technology habits such as “When I have a new technical system in front of me, I try it out intensively” and “It is enough for me to know the basic functions of a technical system”. The statements are judged on a 6-point scale, with responses ranging from ‘Completely disagree’ to ‘Completely agree’. Reliability for the ATI was acceptable (Cronbach’s α = .60).

Confidence in Abilities (CIA) Scale. To assess how confident participants were in their DF recognition abilities, the CIA questions were asked prior to and after the DF recognition task. The CIA scale asks participants to estimate confidence in their abilities on a 6-point scale ranging from ‘Very Good’ to ‘Very Poor’. Prior to the DF recognition task, participants were asked “How would you rate your skills to recognise deep fakes?”. After the DF recognition task, participants were asked “How would you rate your skills to recognize deep fakes now?”. Very high CIA scores and poor actual DF recognition performance indicate overconfidence, perhaps due to low metacognitive awareness and resulting metacognitive inaccuracy. Very low scores and good DF recognition performance indicate underconfidence due to low metacognitive awareness and accuracy. Judgement scores that match performance indicate good metacognitive awareness and accuracy

Judgement of Confidence (JOC) Scale. JOC questions were used prior to and after the DF recognition task. The JOC scale asks participants to judge how accurate they think their confidence in their abilities are on a 6-point scale ranging from ‘Very Good’ to ‘Very Poor’. Prior to the DF recognition task and following the CIA ratings, participants were asked “How confident are you that your rating above in which you describe your deep fake recognition skills, is accurate?”. After the DF recognition task and following the CIA ratings, participants were asked “How confident are you about this opinion about your recognition skill?”.

Certainty in Video Rating (CIVR) Score. To assess participants’ certainty about individual video ratings during the DF recognition task, participants were asked for each video they judged to also rate how certain they were about each ratings on a 4-point scale ranging from ‘Very unsure’ to ‘Very sure’. This requires processing of task difficulty, thus representing a task-oriented judgement of performance. Certainty in DF ratings were averaged as a CIVR DF score. Certainty in authentic ratings were averaged as a CIVR Real score. Total certainty ratings were averaged as a CIVR Overall score.

Overconfidence Scale (OCS). OCS was computed using the following formula:

$$OCS = \frac{\left(\frac{CIA\ Pre\ \times\ 100}{6}\right) + 1}{\%\ \text{of correct ratings} + 1}$$

To avoid dividing by zero, a constant (+1) was added on both sides of the fraction. Because the pre-task CIA score is divided by correct ratings, CIA estimates that were correct returned an OCS score of 1, while any CIA score that overestimated DF detection abilities returned an OCS score above 1. Thus, any OCS score above 1 represents an overconfidence in DF detection abilities. Likewise, values below 1 indicate an underestimation of detection skills.

2.3 Participants, recruitment, and ethical considerations

A total of 247 participants (92 females; 37.2%) were recruited via the online platform Amazon mTurk and financially compensated for their time. This study complies with the Declaration of Helsinki and is in line with the Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals. Informed consent was obtained from all participants prior to the study and they were all briefed about the purpose of the study prior to participation. Information was given to participants that they could withdraw from participation at any time. Participation was completely anonymous; neither IP address nor any personal information that could lead to identification of participants was registered. No methods of deception were applied in the study and all participants were informed that some of the videos would be faked.

2.4 Data reduction and analysis.

Participant characteristics were summarised as means (M) and standard deviations (SD) for continuous variables and number (count) and percentage (%) for ordinal variables and presented in tables. Visual inspection of variables showed that they were not normally distributed. Non-parametric tests were therefore used in all subsequent analyses.

Kruskal-Wallis H test was used to assess relationships between ordinal variables and DF detection skills. Results were reported as H statistics (degrees of freedom), p-values, and η2 (effect size). Dunn's Post-hoc test with Bonferroni adjustment was used to assess between-group differences. Results of post hoc tests were reported as Z statistics and Bonferroni adjusted p-values. Spearman correlations were performed to assess relationships between continuous variables and DF detection skills. Results were reported as Spearman's rho (ρ) and p-values and presented in a table. Separate linear regressions were performed for significant correlations.

In the first part of the analysis, Kruskal-Wallis tests were used to assess the relationships between IT and CS Background (grouping variables) and DF recognition performance (outcome variables). Spearman correlations were used to assess the relationship between ATI scale score and DF recognition performance. Additional Kruskal-Wallis tests were performed for IT and CS background with DF recognition performance (outcome variables) while weighting on ATI scores to see if technology affinity affected results. In the second part of the analysis, Kruskal-Wallis tests were used to assess differences in CIA, JOC, CIVR, and OCS scores (outcome variables) between IT/non-IT, and CS/non-CS groups (grouping variables). Spearman correlations were used to assess the relationship between ATI scale score and CIA, JOC, CIVR, and OCS scores. Additional Kruskal-Wallis tests were performed for IT and CS background with CIA and JOC scores (outcome variables) while weighting on ATI scores to see if technology affinity affected results. In the third part of the analysis, Kruskal-Wallis tests were used to assess differences in DF recognition performance, CIA, JOC, CIVR, and OCS scores between genders. Interval plots were generated for DF task performance, CIA scores, and OCS scores between IT and non-IT professionals and between genders and presented in tables. α level was set to .05 for all comparisons. All analyses were performed using JASP v0.16 [30].

3 Results

3.1 Participant characteristics

Sample background statistics can be found in Table 1.

Table 1. Sample characteristics (N = 247).

Demography	Count (%)	IT background and Device used	Count (%)
Male	155 (62.7)	IT professional	214 (86.6)
Higher education	216 (87.5)	CS professional	193 (78.1)
Country		Familiar with DF term	218 (88.2)
USA	120 (48.5)	Have seen a DF	214 (86.6)
India	104 (42.1)	**Device used**	
Other	23 (9.3)	PC	233 (94.3)
		Tablet	7 (2.8)
		Smartphone	6 (2.4)

Notes. IT = Information technology. CS = Cyber security. DF = Deep fake.

3.2 Descriptive statistics

Descriptive statistics for age, scale and test scores for IT and non-IT professionals, males and females, and for the total sample can be found in Table 2.

Table 2. Descriptive Statistics (N = 247).

	IT		**Non-IT**		**Male**		**Female**		**Total**	
Variables	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Age	34.2	10.3	40.8	13.4	35	11.1	34.9	10.6	35	10.9
ATI Scale	3.9	0.4	3.9	0.4	3.9	0.5	3.9	0.5	3.9	0.5
CIA Pre task	4.3	1.0	3.5	1.2	4.3	1.0	4.1	1.2	4.2	1.1
CIA Post task	4.4	1.0	3.8	1.2	4.5	0.9	4.1	1.2	4.4	1.0
JOC Pre task	4.5	1.0	4.5	1.2	4.6	1.0	4.4	1.2	4.5	1.0
JOC Post-task	4.6	1.0	4.4	0.9	4.7	0.9	4.4	1.1	4.6	1.0
CIVR DF	2.9	0.3	2.9	0.4	2.9	0.3	2.9	0.3	2.9	0.3
CIVR Real	2.9	0.3	3.0	0.5	2.9	0.3	2.9	0.4	2.9	0.4
CIVR Overall	2.9	0.3	3.0	04	2.9	0.3	2.9	0.3	2.9	0.3
OCS	1.7	5.7	1.2	0.8	1.8	6.7	1.2	0.6	1.6	5.3
DF rated real (%)	28.5	20.1	31.3	21.3	30	20.1	26.9	20.7	28.9	20.3
Real rated DF (%)	71.2	27.8	76.7	19.9	72.6	26.5	70.9	27.6	72	26.9
Correct ratings (%)	59.2	16.7	55.6	14.9	57.7	16.6	60.3	16.8	58.7	16.7

Notes. DF = Deep fake. ATI = Affinity for technology interaction. CIA = Confidence in abilities. JOC = Judgment of confidence. Pre task = Pre DF task. Post task = Post DF task. CIVR = Certainty in video rating. OCS = Overconfidence score.

Table 3. Descriptive Statistics and Spearman Correlations (ρ) (N = 247)

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Age	—												
2. Correct ratings	.095	—											
3. Real rated DF	-.181**	-.489***	—										
4. DF rated real	-.017	-.872***	.038	—									
5. ATI	.211***	.090	-.108	-.042	—								
6. CIA Pre	.027	.015	-.117	.058	.145*	—							
7. CIA Post	.040	-.038	-.166**	.124	.138*	.652***	—						
8. JOC Pre	-.150*	-.089	-.056	.152*	.233***	.466***	.461***	—					
9. JOC Post	-.078	-.094	-.092	.170**	.160*	.359***	.451***	.561***	—				
10. CIVR Overall	.126*	-.033	-.038	.057	.460***	.185**	.253***	.292***	.213***	—			
11. CIVR Real	.061	-.100	.064	.075	.350***	.204**	.205**	.229***	.182**	.840***	—		
12. CIVR DF	.132*	-.010	-.069	.047	.464***	.168**	.260***	.282***	.201**	.969***	.708***	—	
13. OCS	-.100	-.699***	.661***	.287***	.033	.632***	.292***	.381***	.292***	.076	.169**	.101	—

Notes. All correlations are 2-tailed. * p < .05. ** p < .01. *** p < .001.
DF = Deep fake. ATI = Affinity for technology interaction scale. CIA = Confidence in Abilities score. JOC = Judgement of confidence. CIVR = Certainty in video rating. OCS = Overconfidence score.

H_1: IT background and/or self-reported IT-affinity are not associated with higher DF recognition skills.

Spearman correlations for age, and scale and test scores can be found in Table 3. To test the hypothesis that IT background (H_1) does not influence DF recognition skills, Kruskal-Wallis tests were performed using IT and CS background as grouping variables and DF performance variables as outcomes. Spearman correlations were performed on ATI scores and DF recognition variables. All correlations can be found in Table 3. There were no significant differences in DF recognition (DF videos rated as authentic) for IT and non-IT (H = 0.08(2), p = .774, η^2 = -.003) or CS and non-CS professionals (H = .21(1), p = .649, η^2 = -.003).

There were no significant differences in rate of false positives (authentic videos rated as DFs) for IT and non-IT (H = 0.22(1), p = .638, η^2 = -.003) or CS and non-CS professionals (H = 0.42(1), p = .519, η^2 = -.002). There were no significant differences in overall DF task performance (correct ratings) for IT and non-IT (H = 1.41(1), p = .234, η^2 = .001) or CS and non-CS professionals (H = 0.09(1), p = .769, η^2 = -.003). ATI score was not associated with DF recognition (ρ = -.042, p = .509), false positive ratings (ρ = -.108, p = .092), or overall DF task performance (ρ = .090, p = .159). Figure 1 shows interval plots for DF detection performance between IT and non-IT- and CS and non-CS professionals.

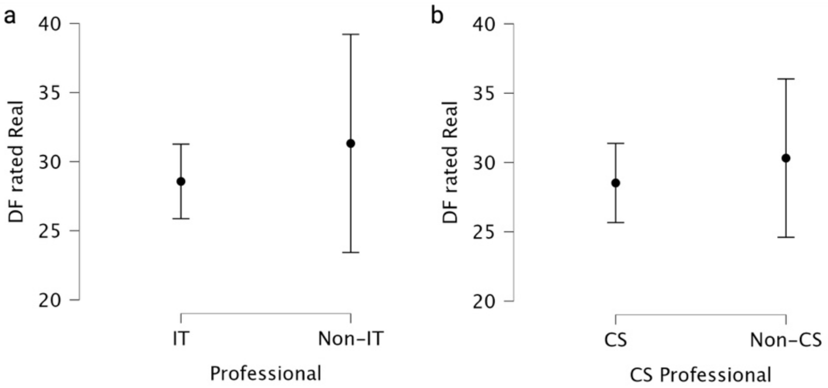

Figure 1: Interval plots for DF detection performance between IT and non-IT professionals. **a** IT background. **b** CS background.

The Kruskal-Wallis test was repeated using IT and CS background as grouping variables and DF performance variables as outcomes while weighting on ATI scores. This did not affect results.

H_2: IT background and/or self-reported IT-affinity are associated with higher perceived DF recognition skills.

CIA scores and IT background. To test the hypothesis that IT background influences belief in DF recognition abilities (H_2), Kruskal-Wallis tests were performed using IT and CS backgrounds as grouping variables and pre-task and post-task CIA scores as outcome variables. Spearman correlations were performed for ATI scores and pre- and post-task CIA scores (Table 3). There was a significant difference in pre-task CIA scores for IT ($H = 9.53(1), p = .002, \eta^2 = .034$) and CS backgrounds ($H = 14.84(1), p < .001, \eta^2 = .064$). Dunn's post hoc test showed that there was a significant difference between IT and non-IT professionals ($z = 3.09, p = .001$), and CS and non-CS professionals ($z = 3.85, p < .001$) with professionals having higher pre-task CIA scores than non-IT and non-CS professionals. There was a significant difference in post-task CIA scores for IT ($H = 8.08(1), p = .004, \eta^2 = .028$) and CS backgrounds ($H = 13.10(1), p < .001, \eta^2 = .049$). Dunn's post hoc test showed that there was a significant difference between IT and non-IT professionals ($z = 2.84, p = .002$), and CS and non-CS professionals ($z = 3.61, p < .001$) with professionals having higher post-task CIA scores than non-IT and non-CS professionals.

ATI scores were positively associated with pre-task CIA ($\rho = .145, p = .023$) and post-task CIA ($\rho = .138, p = .031$) scores ($\rho = .160, p = .012$). Post-task CIA was negatively associated with rating DFs as real ($\rho = -.666, p = .009$).

Linear regression analysis showed that ATI was a significant predictor of pre-task CIA score ($\beta = .136, p = .034, R^2_{Adj} = .014, F = 4.55$) but not post-task CIA scores.

Additional Kruskal-Wallis tests using IT and CS backgrounds as grouping variables and CIA scores as outcome variables while weighting on ATI scores did not affect results.

JOC scores and IT background. Kruskal-Wallis tests were performed using IT and CS backgrounds as grouping variables and pre-task and post-task JOC scores as outcome variables. Spearman correlations were performed for ATI scores and pre- and post-task JOC scores (Table 2). There was not a significant difference in pre-task JOC scores for IT and non-IT ($H = 0.02(1), p = .891, \eta^2 = -.004$) and CS and non-CS professionals ($H = 3.70(1), p = .055, \eta^2 = .011$). There was a significant difference in post-task JOC scores for CS and non-CS ($H = 5.50(1), p = .019, \eta^2 = .018$) backgrounds, but not for IT and non-IT backgrounds ($H = 1.59(1), p = .207, \eta^2 = .002$). Dunn's post hoc test showed that there was a significant difference between CS and non-CS professionals ($z = 2.34, p = .009$) with CS professionals having higher post-task JOC scores than non-professionals.

ATI scores were positively associated with pre-task JOC ($\rho = .233, p < .001$) and post-task JOC scores ($\rho = .160, p = .012$). Post-task JOC scores was positively associated with rating DFs as real ($\rho = .152, p = .017$). Linear regression analysis showed that ATI was a significant predictor of pre-task ($\beta = .185, p = .004, R^2_{Adj} = .030, F = 8.65$) and post-task JOC scores ($\beta = .160, p = .012, R^2_{Adj} = .021, F = 6.34$).

Additional Kruskal-Wallis tests using IT and CS backgrounds as grouping variables and JOC scores as outcome variables while weighting on ATI scores did not affect results.

CIVR scores and IT background. Kruskal-Wallis tests were performed using IT and CS backgrounds as grouping variables and CIVR scores as outcome variables.

Spearman correlations were performed for ATI scores and CIVR scores (Table 3). Significant differences in CIVR DF scores were found for CS and non-CS groups (H = 13.38(1), $p < .001$, $\eta^2 = .050$) but not for IT and non-IT groups (H = 0.68(1), p = .410, η^2 = -.001). Dunn's post hoc test showed that CS professionals scored significantly lower on CIVR DF ratings than non-professionals (z = -3.658, $p < .001$). There was not a significant difference in CIVR Real scores for IT and non-IT professionals (H = .45(1), p = .500, η2 = -.002) or CS and non-CS (H = 3.69(1), p = .055, η^2 = .010) professionals. There was not a significant difference in CIVR Overall scores for IT and non-IT (H = 1.53(1), p = .216, η^2 = .002) professionals. There was a significant difference in CIVR Overall scores for CS and non-CS (H = 13.17(1), p <.001, η^2 = .049) professionals. Dunn's post hoc test showed that CS professionals scored significantly lower on CIVR Overall scores than non-professionals (z = -3.63, $p < .001$).

ATI scores were positively associated with CIVR DF scores (ρ = .464, $p < .001$), CIVR Real scores (ρ = .350, $p < .001$), and CIVR Overall scores (ρ= .460, $p < .001$). CIVR scores were not associated with DF task performance variables.

Linear regression analysis showed that ATI was a significant predictor of CIVR DF scores (β = .392, p = < .001, R^2_{Adj} = .151, F = 44.41), CIVR Real scores (β = .287, p < .001, R^2_{Adj} = .079, F = 21.89), and CIVR Overall scores (β = .379, $p < .001$, R^2_{Adj} = .140, F = 40.85).

Additional Kruskal-Wallis tests using IT and CS backgrounds as grouping variables and CIVR scores as outcome variables while weighting on ATI scores did not affect results.

OCS scores and IT background. Kruskal-Wallis tests were performed using IT and CS backgrounds as grouping variables and OCS as outcome variable. Significant differences in OCS scores were found for CS and non-CS groups (H = 5.56(1), p = .018, η^2 = .018) but not for IT and non-IT groups (H = 2.06(1), p = .151, η^2 = .004). Dunn's post hoc test showed that CS professionals scored significantly higher on OCS than non-professionals (z = 2.36, p = .009). Figure 2 shows interval plots for DF task performance and OCS scores between IT and non-IT professionals. Figure 3 shows interval plots for DF task performance and OCS scores between CS and non-CS professionals.

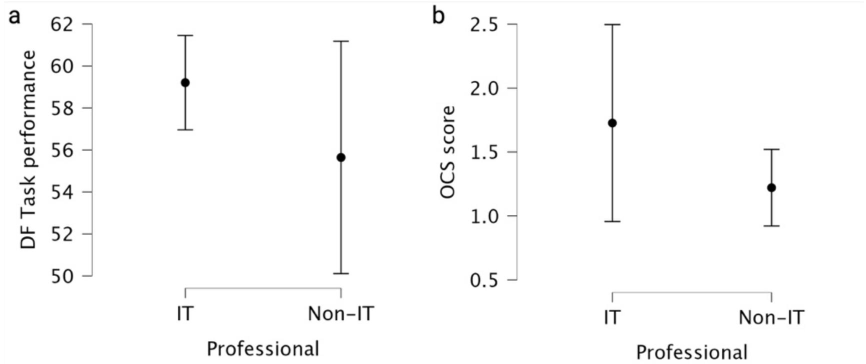

Figure 2: Interval plots for DF task performance and OCS scores between IT and non-IT professionals. **a** DF task performance. **b** OCS score. An OCS score > 1 means overestimation of DF detection skills; an OCS score < 1 means underestimation of DF detection skills.

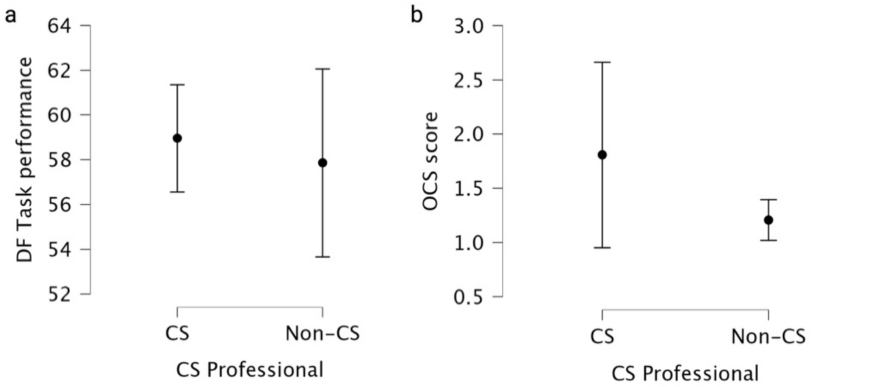

Figure 3: Interval plots for DF task performance and OCS scores between CS and non-CS professionals. **a** DF task performance. **b** OCS score. An OCS score > 1 means overestimation of DF detection skills; an OCS score < 1 means underestimation of DF detection skills.

Additional Kruskal-Wallis tests using IT and CS backgrounds as grouping variables and OCS scores as outcome variable while weighting on ATI scores did not affect results.

Gender differences. Kruskal-Wallis tests for differences in CIA scores, JOC scores, CIVR scores, DF task performance scores, and OCS scores for genders showed that there was only a significant difference for post-task CIA (H = 5.94(1), p = .015, η^2 = .020) and post-task JOC scores (H = 5.28(1), p = .022, η^2 = .017). Dunn's post hoc test showed that males had higher scores than females on post-task CIA (z = 2.43, p = .007) and JOC (z = 2.29, p = .011) scores.

Additional Kruskal-Wallis tests using gender as grouping variable and OCS scores as outcome variable while weighting on ATI scores did not affect results. Figure 4 shows interval plots for OCS and DF Task performance between genders.

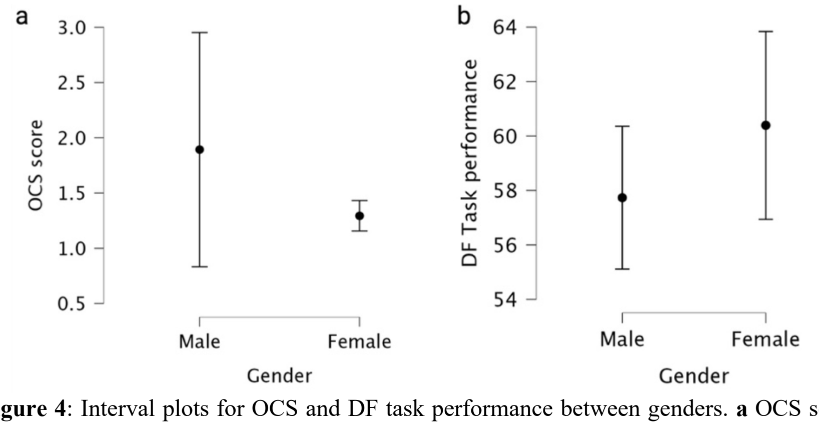

Figure 4: Interval plots for OCS and DF task performance between genders. **a** OCS score. An OCS score > 1 means overestimation of DF detection skills; an OCS score < 1 means underestimation of DF detection skills. **b** DF task performance.

4 Discussion

Developing skills that make individuals resilient against Social Engineering attacks requires individualised approaches to be effective [17]. However the field of research inquiring about the cognitive factors influencing individual Social Engineering susceptibility is currently in its infancy [19]. Recent research indicates that having relevant education or an IT background is not a protective factor against Social Engineering [21, 22]. It has been suggested that cognitive factors prevent targets of Social Engineering from engaging with details in a critical manner [22]. Advances in DF generation technology provide cybercriminals with a unique opportunity to impersonate individuals with high credibility, which could be weaponized in Social Engineering attacks against unsuspecting victims. As with the Social Engineering field in general, little is known about the cognitive factors influencing DF recognition skills. Previous research on human-machine interactions in CS contexts suggest that skills related to self-assessment accuracy such as metacognition is relevant for performance [23, 25, 26]. Overconfidence due to poor metacognitive abilities may serve as a possible explanation for why formal education or an IT professional employment does not guarantee resilience against Social Engineering attacks. Thus, in this study we aimed to assess the influence of overconfidence and IT backgrounds on DF recognition skills.

In line with previous research on phishing email susceptibility [22], and in support of our hypothesis that having an IT background does not influence DF detection skills, we found that individuals with an IT background were no better than non-professionals at judging DFs as authentic. This was true for both IT and CS professionals, and for individuals with an affinity for interacting with technology. Weighting results on ATI scores did not influence DF detection abilities between professionals and non-professionals.

We found support for our second hypothesis that having an IT background influences belief in own abilities to detect DFs. Individuals with an IT background, CS background, and individuals who had an affinity for interacting with technology all scored higher on confidence into their DF detection skills compared to participants without an IT background, or low affinity for interacting with technology. This was true both before and after the task. This suggests that persons describing themselves being close to the IT sector had higher belief in their DF detection abilities. There were no significant differences in their confidence regarding their self-assessment quality between IT professionals and non-IT professionals, suggesting that the confidence in their abilities to judge themselves was very similar. During the task, however, CS professionals were significantly less certain about their DF ratings compared to non-CS professionals, indicating that they were more doubtful of their actual ratings. This could suggest that CS professionals had a more analytical approach when judging performance on a case-by-case basis. These judgements may require processing task difficulty thus being a task-oriented judgement of performance as opposed to when judging their ability to perform which is arguably a more self-oriented judgment of performance. Despite this increased insecurity into their self-assessment, having an IT background, did not influence confidence into their perceived skills. Post-task confidence into their self-assessment were significantly higher for CS professionals compared to non-professionals. People with a higher affinity for interacting with technology had higher confidence in their abilities, were more certain about their task performance, and also had a higher belief that their confidence was accurate, suggesting that people with higher affinity are more confident in their abilities.

These findings could be explained by a lack of cognitive involvement, rather than inability, and are in line with elaboration likelihood models such as the Suspicion, Cognition, Automaticity Model (SCAM; Figure 5, a) [31]. Alternatively, it could be argued the effect may not be the result of a lack of cognitive involvement (i.e., motivation to elaborate the stimulus systematically), but that an IT Background does not provide superior skills in DF recognition. Following this thought, it can be argued that the presented tasks of visual perception and discrimination constituting a DF recognition paradigm are per se based on neuropsychological performance but are not technical tasks. The fact that DF are created and presented on devices with which persons working in the IT domain feel familiar with and show competencies in using, does add to a sense of familiarity and may thus result in overconfidence - even though the task (the systematic processing of a visual or audio-visual stimulus) is perceptual-psychological, rather than technical. The fact that CS professionals were less certain about their performance than non-professionals during task-oriented judgements but had higher belief in their performance for self-oriented judgements could indicate that this latter interpretation might be true at least for CS professionals. On the other hand, people with a higher affinity for interacting with technology were more certain of their case-by-case ratings despite not performing better, possibly indicating a lack of task-oriented judgement of performance which could indeed be reflective of a lack of cognitive engagement with the task.

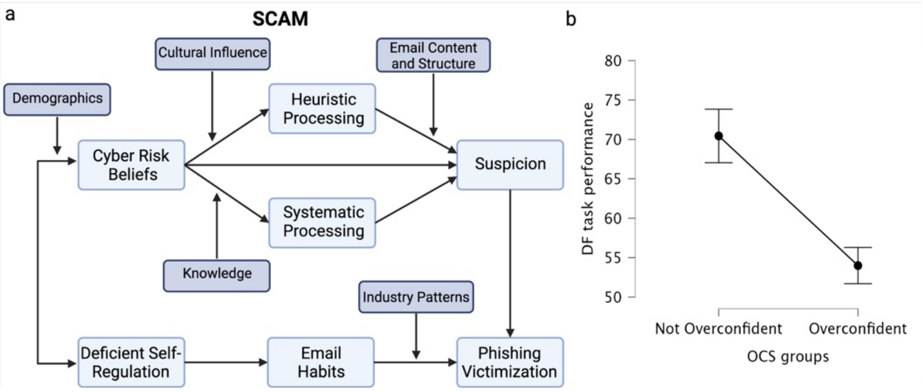

Figure 5: How overconfidence relates to Social Engineering susceptibility in the SCAM. **a** The SCAM. Adapted from [31]. **b** Performance differences between the Overconfident and Not overconfident groups. OCS = Overconfidence scale. SCAM = Suspicion, Cognition, Automaticity Model.

The present study does not determine whether the underlying mechanisms contributing to our findings are due to a lack of engagement or neuropsychological factors related to perceptual processing abilities. Previous research into the neuropsychological correlates of performance on metacognitive and perceptual tasks suggest that they rely on a common neural substrate [32-34]. This could suggest that measuring metacognitive accuracy is indicative of perceptual processing abilities; conversely it could mean that measuring metacognitive accuracy is only indicative of knowing how or when to apply the abilities but not the motivation to do so. Future research measuring confidence judgments and motivation to perform well while applying eye-tracking and EEG (e.g., to record event-related potentials related to perceptual processes and stimuli detection) will be needed to further assess how DF recognition performance relates to engagement and perceptual processes. This could also show whether these underlying mechanisms can be dissociated. Comparing how much time participants spend on individual task items to EEG data may be useful to indicate task engagement relative to perceptual abilities.

Based on these initial findings, we argue that the OCS is an easily obtainable indicator identifying individuals with particular need for systematic feedback and training for improved Social Engineering resilience. OCS assessment can easily be combined with a DF recognition task as demonstrated or be used in classical phishing simulation as they are common practice pre and post cybersecurity awareness interventions. While OCS may uncover persons at heightened risk for failures caused by overestimation of their performance in all sections of an organisation, persons with IT background and related job profiles may be of particular interest. This is due to their demonstrated vulnerability towards overestimation and potentially also due to increased likelihood of access to technical infrastructure and services allowing for a more efficient privilege escalation post-intrusion.

5 Conclusion

These study results suggest that understanding individual differences in DF recognition skills can improve teaching and training outcomes and strengthen Social Engineering resilience by providing a valuable parameter for individualised teaching and training methods. Obtaining a metacognition accuracy score allows us to flag employees with a particular need for improvement. Individual levels of overestimated skills as a major risk factor in safety- and security-critical socio-technical systems can be easily assessed. Individuals with a background in information technology are particularly prone to this vulnerability and do not perform superior in their performance. Further research should investigate the degree to which inaccurate self-assessment can be corrected via feedback mechanisms based on actual performance and consider effective treatment options optimised for this particular population. This way training could potentially contribute to better metacognitive awareness and thus better decision-making.

6 Acknowledgements

This study was supported by the Norwegian Research Council (project number 302941).

References

1. Purplesec, *Cyber Security Statistics*. 2021.
2. Security, I., *Cost of a Data Breach Report 2021*. 2021.
3. Verizon, *2021 Data Breach Investigations Report*. 2021.
4. Hadnagy, C., *Social engineering: The art of human hacking*. 2010: John Wiley & Sons.
5. Mouton, F., et al. *Towards an ontological model defining the social engineering domain*. in *IFIP International Conference on Human Choice and Computers*. 2014. Springer.
6. Uebelacker, S. and S. Quiel. *The social engineering personality framework*. in *2014 Workshop on Socio-Technical Aspects in Security and Trust*. 2014. IEEE.
7. Cialdini, R., *edition 3. Influence: Science and practice*. 1993, New York. Harper Collins College Publishers.
8. Parsons, K., et al., *Predicting susceptibility to social influence in phishing emails.* International Journal of Human-Computer Studies, 2019. **128**: p. 17-26.
9. Baek, E.C. and E.B. Falk, *Persuasion and influence: what makes a successful persuader?* Current opinion in psychology, 2018. **24**: p. 53-57.
10. Schick, N., *Deep Fakes and the Infocalypse: What You Urgently Need to Know*. 2020: Hachette UK.
11. Korshunov, P. and S. Marcel, *Deepfake detection: humans vs. machines.* arXiv preprint arXiv:2009.03155, 2020.
12. Rossler, A., et al. *Faceforensics++: Learning to detect manipulated facial images*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
13. Masood, M., et al., *Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward.* arXiv preprint arXiv:2103.00484, 2021.
14. Zollhöfer, M., et al. *State of the art on monocular 3D face reconstruction, tracking, and applications*. in *Computer Graphics Forum*. 2018. Wiley Online Library.
15. iProov, *The Threat of Deepfakes. The consumer view of deepfakes and the role of biometric authentication in protecting against their misuse*. 2020.
16. Hu, S., Y. Li, and S. Lyu. *Exposing GAN-generated faces using inconsistent corneal specular highlights*. in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. IEEE.

17. Drogkaris, P. and A. Bourka, *Cybersecurity culture guidelines: Behavioural aspects of cybersecurity.* European Union Agency for Network and Information Security (ENISA), 2019.
18. Egelman, S. and E. Peer. *The myth of the average user: Improving privacy and security systems through individualization*. in *Proceedings of the 2015 New Security Paradigms Workshop*. 2015.
19. Montañez, R., E. Golob, and S. Xu, *Human cognition through the lens of social engineering cyberattacks.* Frontiers in Psychology, 2020. **11**: p. 1755.
20. Schraw, G., *Promoting general metacognitive awareness.* Instructional science, 1998. **26**(1): p. 113-125.
21. Butavicius, M., et al., *Breaching the human firewall: Social engineering in phishing and spear-phishing emails.* arXiv preprint arXiv:1606.00887, 2016.
22. Jampen, D., et al., *Don't click: towards an effective anti-phishing training. A comparative literature review.* Human-centric Computing and Information Sciences, 2020. **10**(1): p. 1-41.
23. Jøsok, Ø., et al. *Exploring the hybrid space*. in *International Conference on Augmented Cognition*. 2016. Springer.
24. Jøsok, Ø., et al. *Macrocognition applied to the hybrid space: team environment, functions and processes in cyber operations*. in *International Conference on Augmented Cognition*. 2017. Springer.
25. Knox, B.J., et al., *Socio-technical communication: the hybrid space and the OLB model for science-based cyber education.* Military Psychology, 2018. **30**(4): p. 350-359.
26. Knox, B.J., et al. *Towards a cognitive agility index: the role of metacognition in human computer interaction*. in *International conference on human-computer interaction*. 2017. Springer.
27. Canfield, C.I., B. Fischhoff, and A. Davis, *Better beware: comparing metacognition for phishing and legitimate emails.* Metacognition and Learning, 2019. **14**(3): p. 343-362.
28. Kleitman, S., M.K. Law, and J. Kay, *It's the deceiver and the receiver: Individual differences in phishing susceptibility and false positives with item profiling.* PloS one, 2018. **13**(10): p. e0205089.
29. Franke, T., C. Attig, and D. Wessel, *A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale.* International Journal of Human–Computer Interaction, 2019. **35**(6): p. 456-467.
30. JASP, *JASP-Statistics*. 2021.
31. Vishwanath, A., B. Harrison, and Y.J. Ng, *Suspicion, cognition, and automaticity model of phishing susceptibility.* Communication Research, 2018. **45**(8): p. 1146-1166.
32. Chechlacz, M., et al., *Structural variability within frontoparietal networks and individual differences in attentional functions: an approach using the theory of visual attention.* Journal of Neuroscience, 2015. **35**(30): p. 10647-10658.

33. Shekhar, M. and D. Rahnev, *Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition.* Journal of Neuroscience, 2018. **38**(22): p. 5078-5087.

34. Zanto, T.P., et al., *Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory.* Nature neuroscience, 2011. **14**(5): p. 656-661.