Deepfake Detection in Super-Recognizers and Police Officers

Meike Ramon | University of Lausanne and AIR – Association for Independent Research
Matthew Vowels | University of Lausanne, Lausanne University Hospital and University of Lausanne, and The Sense Innovation and Research Center
Matthew Groh | Northwestern University

We examined human deepfake detection performance (DDP) in relation to face identity processing ability among Berlin Police officers, including Super-Recognizers (SRs). While we find no relationship, further research into human DDP using state-of-the-art static deepfakes is needed to establish the potential value of SR-deployment.

The present study is the first empirical investigation of the relationship between human deepfake detection performance (DDP) and individuals' face identity processing (FIP) ability. Using videos from the Deepfake Detection Challenge, we investigated DDP in two unique observer groups: Super-Recognizers (SRs) and "normal" officers from within the 18,000 members of the Berlin Police. SRs were identified either via previously proposed lab-based procedures or the only existing tool for SR identification involving increasingly challenging authentic forensic material: the Berlin Test For Super-Recognizer Identification (beSure). Participants judged either pairs of videos or single videos in a two-alternative forced-choice (2AFC) decision setting (that is, which of the pair or whether a single video was a deepfake or not). We explored speed–accuracy tradeoffs and compared DDP between lab-identified SRs and non-SRs and police officers as a function of their independently measured FIP ability. Interestingly, we found no relationship between DDP and FIP ability. Further work using static deepfakes created with current state-of-the-art generative models is needed to determine the value of SR deployment for deepfake detection in law enforcement.

Digital Object Identifier 10.1109/MSEC.2024.3371030

Introduction

Perception Versus Reality

Our perception of the world around us is highly subjective; presented with the same information, we interpret it in vastly different ways. Our perception is influenced by several factors, most of which operate without our knowledge or control.[1] Perceptual illusions provide compelling examples of the highly subjective nature of human perception and of how it is influenced by both external stimuli and internal cognitive processes. Simply put, the world as we perceive it reflects a unique interaction between incoming information and the way it is processed depending on our abilities, prior experiences, and expectations. Adding to this *human* complexity, technological advances provide means to alter or even create entirely novel information (for a review, see Farid[2]). Whether consciously or not, all of us are likely to have already experienced some form of synthetic media—or *deepfake*.

Deepfakes have been used in the realm of art, for example, to (re)create characters or scenarios, including interactive installations to create immersive experiences.[3] In wider society, the word "deepfake" is typically associated with misinformation, that is, the

Copublished by the IEEE Computer and Reliability Societies

intentional manipulation of audio content or facial information. Facial deepfakes can take various forms, for example, swapping the entire face or individual features or manipulating them—either in static 2D images or dynamic video sequences. Such manipulations can include masking or enhancing information or changing characteristics that are stable (for example, gender and ethnicity) or those that vary across different time scales (for example, age and expressions of emotion). These manipulations aside, facial deepfakes also encompass the creation of entirely new artificially generated *synthetic facial identities*.

Deepfake Detection in Real Life and Laboratory Conditions

In 2022, several European mayors, including those from Berlin, Madrid, and Vienna, were deceived into holding video calls with a deepfake impersonating Vitali Klitschko, the mayor of Kyiv.[4] In the case of Berlin, Mayor Franziska Giffey became suspicious ~15 min into the call when the fake Klitschko started discussing controversial topics regarding Ukrainian refugees. The deception, which was confirmed through diplomatic channels, emphasizes the usage and impact of misinformation in the political realm.

Notwithstanding their importance, the known number of instances of deepfake deployment to influence politics is modest compared to the much more frequent targeting of celebrities, public figures, and everyday people.[5] Deepfakes are considered to pose crucial "risks to our democracy and to national security" as well as "individuals and businesses fac[ing] novel forms of exploitation, intimidation, and personal sabotage."[6] Given the challenges associated with facial deepfakes, the increasing number of studies emerging in this domain is not surprising. However, these studies typically report algorithmic approaches to tackle deepfake detection. Importantly, the speed of deep learning-based deepfake detection theoretically makes them suitable for large-scale implementation. However, human and machine-based processing operates based on different features, which remain to be clearly defined (Wichmann and Geirhos[7]). Differences between machines and humans aside, our understanding of humans' ability for deepfake detection actually remains largely unexplored. Considering that humans will always be required to make final decisions, it is critical to understand the limits of our ability to detect deepfakes.

A few studies have investigated humans' perception of deepfakes. For example, Groh et al.[8] reported that "ordinary humans perform in the range of the leading machine learning model on a large set of minimal context videos" (p. 1). Although the highest DDP could be achieved by combining human and model predictions, humans often incorrectly updated their responses when exposed to inaccurate model predictions (that is, machine-based responses that were not correct). Thus, integrating human with model predictions can result in an *increase or decrease* in DDP. Moreover, the authors reported that manipulations that are known to disrupt human FIP, notably stimulus inversion, were associated with decreased human—but not model—performance. These findings were interpreted as supporting "a role for specialized cognitive capacities in explaining human deepfake detection performance" (Groh et al.[8]).

Knowledge Gap

An important question that remains unanswered is whether—and to what degree—deepfake detection performance varies across observers. For instance, highly motivated trained law enforcement professionals might outperform neurotypical observers, who are not professionally tasked with face or deepfake processing. On the other hand, it is possible that stable individual differences in FIP ability, notably *superior* abilities (for example, Ramon[9] and Ramon and Vowels[10]), may be a better predictor of DDP. Conceivably, compared to neurotypical observers, individuals with substantially inferior or exceptionally superior FIP ability may exhibit markedly different sensitivity to information manipulation.

Over the past decade, there has been a surging interest in so-called Super-Recognizers, individuals with an apparently innate superiority in face identity processing (Ramon,[9] Ramon and Vowels,[10] and Mayer and Ramon[11]). These individuals are of interest not only to cognitive (neuro)scientists but also to law enforcement and policing (Ramon[9] and Ramon et al.[12]). The most consistent strategy for identifying these unique individuals has been proposed by Ramon[9], whose diagnostic framework for lab-based SR identification comprises challenging behavioral tests assessing perception and recognition memory for facial identities. The only existing tool designed to identify law enforcement professionals using authentic police images was proposed by Ramon and Rjosk[13]. While empirical evidence into the mechanisms underlying SRs' ability is mounting (for example, Nador et al.[14], Nador et al.[15], and Linka et al.[16]), to date, no study has investigated DDP in SRs or law enforcement professionals.

Understanding factors that influence our perception and detection of deepfakes is critical considering their potentially wide-ranging societal implications. Such knowledge is particularly pertinent for organizations that are expected to monitor and mitigate threats by deepfakes: law enforcement professionals. Therefore, in this study, we investigated the impact of human factors—professional occupation and individual differences in face identity processing ability—on

deepfake detection performance. We did so by testing two unique cohorts of observers: previously reported SRs (Ramon[9]) and law enforcement professionals from within the 18,000 officers of the Berlin Police (Ramon and Vowels[10]). Their performance was measured across using identical stimulus material, experimental settings, and neurotypical control observers' data as reported previously (Groh et al.[8]).

Methods

This research complies with all relevant ethical regulations, and the Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects approved the deepfake detection portion of this study as Exempt Category 3 – Benign Behavioral Intervention. This study's exemption identification number is E-3354. All procedures and protocols were approved by the University of Fribourg's Ethics Committee (approval number 473) and conducted in accordance with both their guidelines as well as those set forth in the Declaration of Helsinki. All participants were healthy volunteers, were provided with informed written consent, and were not financially compensated for their participation.

Experiments

Participants were invited to participate in two deepfake detection tests reported previously by Groh et al.[8] and exemplified in Figure 1(a). The first experiment involves presenting two stimuli in a 2AFC design; the second presents a single stimulus. Observers are required to decide which of the two stimuli in the 2AFC setting represents a deepfake and report their confidence in the single-video stimulus being a deepfake. Participants could complete as many trials as they wished. The full 2AFC and single-video experiments comprised a total of 56 and 56 trials, respectively (for full details, see Groh et al.[8]).

Participants

The data reported in this study originated from different sources. First, data published previously by Groh et al.[8] included nonrecruited observers (who arrived at the website via organic links on the Internet) and observers recruited from Prolific[17]. These data were considered as representing neurotypical controls (as no independent measure of their FIP ability was available). Second, data from lab-identified SRs reported previously (Ramon[9]) and thereafter using the same lab criteria were invited to participate. Finally,

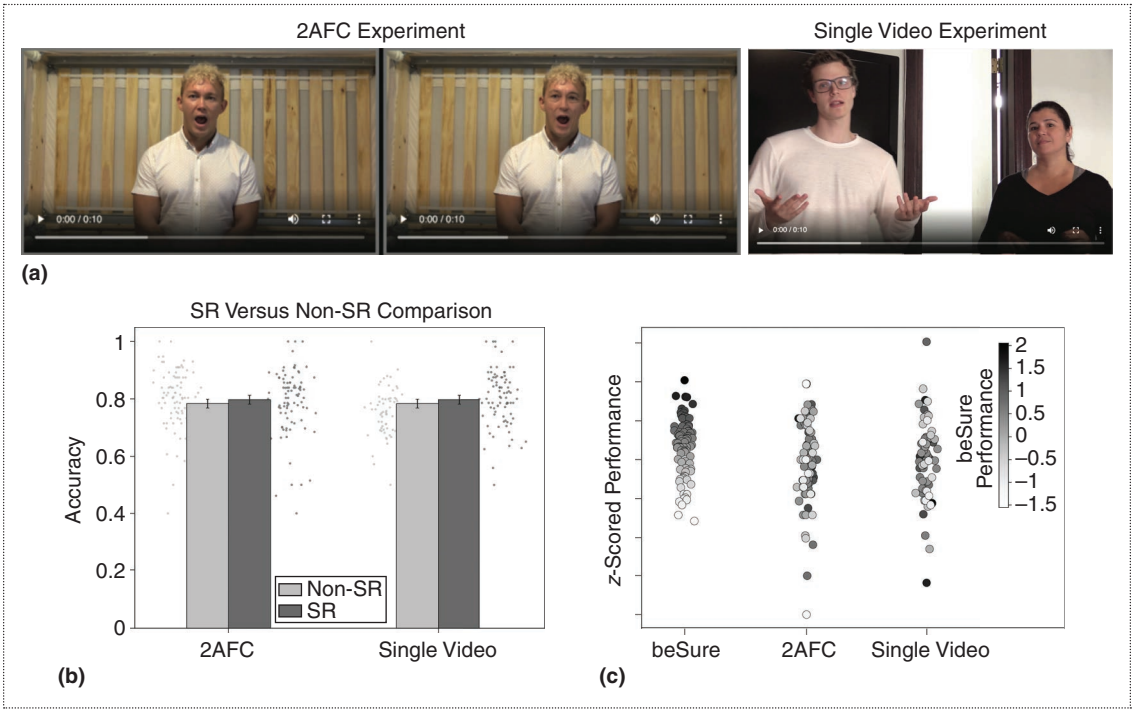

Figure 1. Stimuli and results for DDP. (a) Example stimuli presented in the 2AFC (left) and single-video (right) experiments. (Source: Adapted from Groh et al.[8]) (b) DDP for each of the two experiments for SRs and control observers (dark and light grey). (c) Relationship between different performance measures along the *x*-axis: performance across beSure (left) and both deepfake experiments (middle and right). Colors indicate beSure performance rank to visualize the (in)dependence between FIP ability measured by beSure, and observers' performance for the deepfake experiments.

Berlin Police officers who had previously participated in beSure (Ramon and Vowels[10] and Ramon and Rjosk[13]), the only existing tool for SR identification with authentic police material (for details, see Ramon and Vowels[10] and Ramon and Rjosk[13]), were invited to participate in the deepfake detection experiments.

In total number, 193 individuals contributed data to the first experiment (decision: which video in a pair was a deepfake), and 132 contributed to the second experiment (decision: whether individually presented videos were real or fake). Of these, 106 and 68 met the SR lab criteria (Ramon[9]). Note that the majority of SRs were thus not Berlin Police officers but from the ~90 individuals tested in the AFC Lab. Eighteen lab-identified SRs were from the sample of participating police officers. Note: Demographic information can be provided only for participating Berlin Police officers (Ramon and Vowels[10]) and are summarized in Table 1. [No information is available for the nonrecruited/recruited observers reported originally by Groh et al.[8] or the SRs reported previously (Ramon[9]), and after this publication, using the same lab criteria as participation did not require the provision of personal information.]

Analyses

To investigate the relationship between FIP ability and DDP, we considered performance on lab- and police-based procedures (beSure; Ramon and Rjosk[13]) and deepfake stimuli across both the single-video and 2AFC experiments. All such analyses were performed using the software/language R (R Core Team[18]) by using a zero-and-one inflated beta regression model (BEINF), implemented using the gamlss package (Rigby and Stasinopoulos[19]), for the single-video experiment (each trial within which has a fractional [0,1] outcome) and a multilevel logistic model for the 2AFC experiment (each trial within which has a binary correct/not correct outcome). The BEINF model is generally employed in statistical analyses when the outcome variable of interest is continuous but bounded within a specific interval and where the data also exhibit a nonstandard distribution within the interval, such as a pronounced skewness or the presence of peaks at the boundaries, which are common in proportion or percentage data. For this model, we create an average of the individual trial results and regress it onto the group variable.

The multilevel binomial logistic regression model, implemented using the lme4 package (Bates et al.[20]), was employed to examine the effect of group membership on a binary outcome (that is, correct versus incorrect responses) while accounting for the nonindependence of repeated measures within individuals. Specifically, the model incorporates a fixed effect for the group variable to assess its influence on the likelihood of a correct response and a random intercept for participants to model the variability in baseline log odds of success across participants, thereby accommodating the repeated measures design.

A key characteristic of the BEINF model is its ability to accommodate data with excess zeros or ones, a phenomenon often referred to as *inflation* at the boundaries. Traditional models, such as the beta regression, are well suited for continuous outcomes constrained within (0, 1); however, they struggle with boundary inflation because they assume that the distribution of the outcome variable is smooth across the entire interval. The BEINF model extends the Beta regression by incorporating parameters that explicitly model the probability of observing these boundary values, thus providing a more nuanced understanding of the data distribution.

A full set of comparisons can be found in the supplementary materials. Given that we undertook multiple comparisons, we used an adjusted alpha level of 0.001. Reaction time (RT) data (measured in seconds) were winsorized (between the fifth and 95th percentiles to deal with outliers where participants left the response survey software open without participating) and then z-scored to improve model convergence. Altogether, our analyses aimed at answering three distinct questions.

First, we investigated whether there is a relationship between DDP and processing duration, that is, a correlation between accuracy and RTs in the newly acquired data (lab-identified SRs and Berlin Police officers). This served to determine the potential presence of speed–accuracy tradeoffs that could account for the obtained findings.

Second, labeling observers categorically, we asked whether individuals identified as lab-identified SRs (Ramon[9]) excel at deepfake detection relative to neurotypical observers reported previously (Groh et al.[8] and Ramon and Vowels[10]). We combine all available data and consider those observers as SRs who met the proposed lab-based criteria for SR identification (Ramon[9]), including those among

Table 1. Demographic information for participating Berlin Police officers.

	n (Sample Size)	Mean Age	Standard Deviation	Handedness (Right/Left/ Ambidextrous)	Gender (Female/ Male/ Diverse)
2AFC	89	42	9	75/12/2	27/62/0
Single video	65	42	9	56/7/2	21/41/0

Berlin Police officers. Note that additional analyses performed for all subgroup-wise comparisons are provided in the supplementary materials in the accompanying OSF project.

Third, and finally, we sought to determine the potential relationship between FIP ability and DDP by considering police officers' FIP ability in a continuous manner through their previously measured performance across all five subtests of the bespoke police tool beSure (Ramon and Vowels[10] and Ramon and Rjosk[13]). To this end, we first performed linear regressions for performance in deepfake experiments and beSure performance. Additionally, given the possibility that the relationships may be nonlinear, we also explored whether a data-driven approach would indicate predictive potential. To this end, we undertook the same regressions for the single-video and 2AFC experiments—but this time with a random forest (Breiman[21]). Random forests are a type of data-adaptive, nonparametric, and tree-based machine learning algorithm that learn a function that maps from the predictors to the dependent variable. The *forest* element refers to the fact that multiple trees are used, each of which is trained on a bootstrapped subsample of the input data and input variables. This bootstrapping process helps to prevent overfitting, a phenomenon whereby data-adaptive approaches tend to learn ungeneralizable functions that exhibit only good performance on the data on which they are trained.

For the random forest, we use the sklearn implementation (Pedregosa et al.[22]) with its default values, which have been shown to yield consistently good performance across a range of tasks without needing hyperparameter tuning (Probst et al.[23]). Specifically, the core hyperparameters were as follows: number of estimators, 100; maximum features, all; maximum depth, unlimited; minimum sample split, two; and criterion, squared error. No experiments were undertaken to evaluate whether better hyperparameters could be identified (we assume that the algorithm is already substantially more flexible than the alternative linear regressors under comparison). We follow a leave-one-out cross-validation process to evaluate the out-of-sample mean-squared-error performance of the random forest and compare it to a "dummy" regressor, which simply predicts the average value of the outcome.

Results

Relationship Between Performance and RTs

First, we explored the extent to which RTs in both experiments would be predictive of DDP. Specifically, we aimed to determine whether higher performance accuracy could be accounted for by prolonged RTs, that is, a speed–accuracy tradeoff.

To this end, for the 2AFC experiment, we fit the multilevel logistic model to the data to assess the relationship between DDP (correct/incorrect) and a standardized RT while accounting for random intercepts associated with individual users. In terms of the fixed effects, the (standardized) RT was negatively associated with the log odds of correct deepfake detection (B = −0.373, SE = 0.029, z = −12.97, $p < 0.001$). Here, "B" represents the fixed effect regression coefficient for the standardized RT, indicating its effect on the log odds of correctly detecting a deepfake. "SE" is the standard error of the estimate for "B," quantifying the uncertainty/variability. The "z" value serves as the test statistic for assessing the significance of the effect, and the associated "p" value indicates the probability of observing such an effect (or stronger) under the assumption that there is, in fact, no association.

Taking the exponent of the fixed effect "B," we get an odds ratio of approximately 0.69. In other words, for every one-standard-deviation increase in the RT, the odds of correctly detecting a deepfake are decreased by about 31% relative to the odds of someone reacting in an average amount of time. It is important to note that this association between an increased RT and the decreased detection accuracy does not imply causality. The observed relationship might suggest that longer RTs are linked to greater uncertainty in distinguishing deepfakes, potentially because more challenging decisions require longer deliberation. However, this interpretation is speculative, and further research would be necessary to explore the underlying mechanisms.

On the other hand, for the single-video tasks, which have a fractional performance measure $[0,1]$, we use a zero-and-one inflated beta generalized additive regression model (Stasinopoulos et al.[24]), which we fit to, again, assess the association between the standardized RT and performance. The main coefficient to be evaluated is μ (estimate = −0.013, SE = 0.048, t = −0.272, p = 0.786). An interpretation of these results follows in a similar manner to those for the multilevel model. Here, "t" is the test statistic rather than "z." These results indicate that there is no significant relationship between the RT and the expected score—the threshold for significance is taken to be $\alpha = 0.05$, and the value of "p" is above this.

Taken together, analyses for both the single-video and the 2AFC experiment have ruled out speed–accuracy tradeoffs. If anything, we observed the opposite pattern—lower performance associated with prolonged RTs. Therefore, only performance accuracy was considered in further analyses.

Group Differences: SRs Versus Controls

The relationship between independently measured FIP ability and DDP was first investigated by categorizing observers according to their SR status. Recall that observers originated from different groups: 1) previously reported SRs (Ramon[9]) Berlin Police officers who met the lab criteria and those who did not and 2) recruited and nonrecruited observers reported previously (Groh et al.[8]). We combined all non-SR and SR data, respectively, to investigate potential group differences in DDP. A bar and scatter plot for the comparison can be seen in Figure 1.

In addition, for the 2AFC experiment, we fit a generalized linear mixed model (with binomial/logistic link function) to the data to assess the relationship between DDP (correct/incorrect) and SR status while accounting for random effects associated with individual users. The likelihood of a correct response for SRs was not significantly different from the reference group ($B = 0.082$, SE = 0.082, $z = 1.001$, $p = 0.317$).

Similarly, for the single-video tasks, which have a fractional performance measure $[0,1]$, we use a zero-and-one inflated beta generalized additive regression model (Stasinopoulos et al.[24]). For the μ component, which represents the mean of the average score, and recalling our adjusted alpha level of 0.001 in light of the complete set of group comparisons undertaken and presented in the supplementary materials, the likelihood of a correct response for SRs was not significantly different from the reference group ($B = 0.156$, SE = 0.077, $t = 2.019$, $p = 0.046$).

Individual Differences: Continuous Measure of FIP via beSure

Finally, focusing on Berlin Police officers, we investigated the relationship between individual differences in DDP and their FIP ability as measured by the five subtests of beSure (Ramon and Vowels[10]).

First, we investigated potential linear relationships via Spearman correlations between the z-standardized averages across the beSure subtest rank performances and observers' ranking in the single-video and 2AFC experiments. For this analysis, results from the single-video and 2AFC trials were averaged to generate a singular summary score for each participant's performance, which then served as the dependent variable in the respective regression. For the single-video experiment, the rank-order correlation with the beSure performance ranking was −0.03 ($p = 0.84$). For the 2AFC experiment, the rank-order correlation with the beSure performance ranking was 0.12 ($p = 0.27$). As such, no significant relationship was identified. Comprehensive multiple regression outcomes for both experiments are presented in Tables 2 and 3, respectively. The sole significant predictor was the beSure Subtest 4 accuracy performance for the single-video experiment, $B = -0.032$, $t(5) = -2.790$, $p = 0.007$—however, with an effect in the *opposite* direction as one might expect. Nevertheless, due to the modest R^2 values for both single-video and 2AFC experiments (0.123 and 0.053, respectively), we abstain from interpreting this specific finding.

Second, we addressed potential nonlinear relationships via regressions performed with a random forest (Breiman[21]) for both experiments. For the single-video experiment regression, we find that the random forest has an out-of-sample mean-squared error of 0.0072, while the mean-squared error for the dummy is 0.0067. Similarly, the random forest out-of-sample mean-squared error for the 2AFC experiment was 0.0121, while that of the dummy regression was 0.0104. As such, in both cases, the dummy regression was better than the random forest (lower mean-squared error). These results suggest that even a relatively powerful data-adaptive algorithm is able to predict neither single-video nor 2AFC experiment performance via an independent and continuous measure of FIP ability, derived from all five beSure subtests (Ramon and Vowels[10] and Ramon and Rjosk[13]).

Discussion

Society is confronted with increasing amounts of digital misinformation and a lack of solutions for their detection. Compared to the number of studies reporting automatic solutions developed toward this end, empirical studies of human ability for deepfake detection remain severely limited. Moreover, existing studies have not considered two potential determinants of deepfake detection performance: stable individual differences in face identity processing ability and

Table 2. Linear regression results for the single-video experiment performance as the dependent variable with beSure subtest performance as predictors.

	Coefficient	SE	*t*-Value	*p*-Value
Constant	0.755	0.01	77.09	<0.001
Subtest 1 accuracy	0.004	0.013	0.265	0.792
Subtest 2 accuracy	0.01	0.015	0.712	0.48
Subtest 3 accuracy	0.004	0.013	0.298	0.767
Subtest 4 accuracy	−0.032	0.012	−2.790	0.007
Subtest 5 accuracy	0.006	0.012	0.516	0.608
R^2				0.123
Adjusted R^2				0.049

professional occupation. To address this knowledge gap, we leveraged access to two unique groups of human observers: previously reported SRs and motivated officers from within the entire group of ~18,000 employed by the Berlin Police (Ramon,[9] Ramon and Vowels,[10] and Mayer and Ramon[11]). The latter had participated in beSure (Ramon and Vowels[10] and Ramon and Rjosk[13])—the only existing police FIP assessment tool using *authentic police material*. In this manner, we could relate DDP to two independent, challenging, and complementary means of FIP assessment. In light of the challenges that synthetic misinformation represents, we sought to expand our understanding of the human limits for facial deepfake detection.

No Evidence of Speed–Accuracy Tradeoffs

Independently of FIP ability, we sought to determine whether DDP is characterized by speed–accuracy tradeoffs. It is conceivable that high performance could be attributed to the depth with which individuals opt to process information. In this case, high performance would come at the expense of prolonged RTs. On the other hand, an absence of such speed–accuracy tradeoffs would suggest that other factors may be more meaningful determinants of observers' deepfake detection. Overall, across diverse cohorts, we did not find a speed–accuracy tradeoff, that is, *improved* performance due to associated with prolonged processing (that is, response) time. If anything, for the 2AFC experiment, performance deteriorated with processing time, while no relationship was found for the single-video experiment.

SRs Versus Controls

Next, we sought to determine whether stable differences in FIP ability might affect DDP. To this end, we examined if individuals categorized as SRs according to previously proposed lab-based diagnostic procedures (Ramon[9]) would outperform those who did not. Indeed, recent evidence has demonstrated that SRs excel at forensic perpetrator identification (Mayer and Ramon[11]). Moreover, they outperform non-SRs in challenging identity-matching scenarios measured via beSure, the only FIP assessment tool that involves authentic police material (Ramon and Vowels[10] and Ramon and Rjosk[13]). It is thus conceivable that SRs' superiority extends to the detection of synthetic disinformation.

We analyzed an extensive dataset of single-trial responses solicited in a single-video and a 2AFC experiment, respectively. Observers belonged to two groups: 1) civilians or Berlin Police officers identified as SRs via lab tests (Ramon[9] and Ramon and Vowels[10]) who represent the core of a deep-data neuroscientific research agenda pursued in the Applied Face Cognition Lab (https://afclab.org/) and 2) non-SRs who were previously reported neurotypical observers (Groh et al.[8]) and officers of the Berlin Police who did not meet the SR criteria (Ramon[9]). The results indicate that DDP was not related to group membership.

These findings may be accounted for by the stimulus material presented and used. SRs outperform controls when the processing of static images of faces is required (Ramon,[9] Ramon and Vowels,[10] and Ramon and Rjosk[13]). Here, however, observers judged *dynamic* stimuli. The availability of motion information may have leveled the field across observers.

Table 3. Linear regression results of 2AFC experiment performance as the dependent variable with beSure subtest performance as predictors.

	Coefficient	SE	*t*-Value	*p*-Value
Constant	0.804	0.011	74.61	<0.001
Subtest 1 accuracy	0.004	0.015	0.25	0.803
Subtest 2 accuracy	0.012	0.015	0.824	0.412
Subtest 3 accuracy	0.002	0.014	0.169	0.866
Subtest 4 accuracy	−0.013	0.013	−0.969	0.335
Subtest 5 accuracy	0.017	0.013	1.301	0.197
R^2				0.053
Adjusted R^2				−0.004

Individual Differences in FIP in Police Officers

To complement the categorical approach comparing SRs to non-SRs, our final analysis concentrated on police officers, who had undergone testing of FIP ability via a novel bespoke police tool: beSure (Ramon and Vowels[10] and Ramon and Rjosk[13]). This was done to address whether a potential association between FIP ability and DDP would require a more sensitive individual differences approach. This continuous analytical approach again provided a null finding; officers' DDP was unrelated to their FIP ability rank determined via the challenging five subtests of beSure (Ramon and Vowels[10] and Ramon and Rjosk[13]).

Limitations and Future Outlook

Collectively, our results suggest that neither increased processing time, which can be considered a proxy for motivation, nor FIP ability measured via two independent approaches are associated with DDP. These findings emerge within a large, diverse, and unique group of observers, which we believe represent society at large as well as motivated law enforcement professionals.

An important consideration concerns the different number of trials completed across participants' subgroups. For the first two analyses, we combined the previously reported dataset (Groh et al.[8]) with our newly acquired one. According to Groh et al.,[8] "[r]ecruited participants [were] asked to view 20 videos while nonrecruited participants [could] view up to 45 videos." Provided uninterrupted participation, observers of the present cohort were exposed to the complete set of deepfake stimuli. As such, we cannot rule out a greater learning effect for these observers. However, these considerations do not hold for the third analysis, which was performed exclusively on Berlin Police officers' data. Here, we also did not find any significant association between ability and DDP. [However, since we did not know a priori whether any differences would emerge, further work testing for null effects (for example, via equivalence tests; Lakens[25]) is required.]

One obvious caveat is that our findings are linked to the stimulus material used, which represents a subsample of instances submitted to the Deepfake Detection Challenge (https://www.kaggle.com/c/deepfake-detection-challenge) (see Groh et al.[8]). Since this challenge, the number of solutions available for deepfake creation has increased substantially. This means that today's deepfakes will vary much more in terms of their quality and likely detection difficulty. Indeed, it is possible that facial deepfake stimuli created using state-of-the-art approaches might be processed more proficiently by individuals with high(er) FIP ability. However, the stimuli used here are arguably the most extensively studied among humans (Groh et al.[8]).

Another open question concerns within-observer reliability—or consistency in judging the same deepfake stimulus. Given repeated exposure to the same stimuli, it is possible that the consistency of observers' judgments is related to their FIP ability (Ramon[9]). Relatedly, previous work has shown that severely impaired individuals especially benefit from instructions that reveal task-relevant diagnostic information (for example, Ramon and Rossion[26]). Thus, future work should address the extent to which judgments are influenced by prior information on or familiarity with deepfakes could affect observers' performance—and potentially interact with individual differences in FIP ability. Finally, facial deepfakes may be combined with audio content, which in isolation can facilitate or hamper DDP (Groh et al.[27]). Potentially, the detection of deepfakes involving both audio and visual information could relate to stable individual differences in multisensory integration. ■

Acknowledgement

Meike Ramon thanks Simon Rjosk, the Berlin Police, in particular, the Center for Innovation and Science Management for the fruitful long-standing collaboration and shared dedication to scientific quality and transparency and all participating police officers for their support and service. Meike Ramon is supported by a Swiss National Science Foundation PRIMA (Promoting Women in Academia) Grant (PR00P1 179872). M. R. and M. G. conceived the experiments; M. R. conducted face identity processing assessments; M. G. conducted the deepfake experiments; M. R. and M. V. analyzed face identity processing assessment data; M. V. and M. G. analyzed the deepfake experimental data; M. R. wrote the initial manuscript, and all authors edited the final version. Anonymized research data reported (that is, behavioral responses provided by data acquired from previously identified SRs and Berlin Police officers) subject to analysis and analysis code can be found on the accompanying OSF project (https://osf.io/zw7vm/). For previously published data used here, please refer to the original report (Groh et al.[8]). This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects (performed in line with exemption identification number E-3354) and the University of Fribourg's Ethics Committee (approval number 473).

References

1. D. J. Kersten, P. Mamassian, and A. L. Yuille, "Object perception as Bayesian inference," *Annu. Rev. Psychol.*, vol. 55, no. 1, pp. 271–304, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:2230247
2. H. Farid, "Creating, using, misusing, and detecting deep fakes," *J. Online Trust Saf.*, vol. 1, no. 4, pp. 1–33, 2022, doi: 10.54501/jots.v1i4.56.
3. P. Pataranutaporn et al., "Ai-generated characters for supporting personalized learning and well-being," *Nature Mach. Intell.*, vol. 3, no. 12, pp. 1013–1022, 2021, doi: 10.1038/s42256-021-00417-9.
4. "European leaders targeted by deepfake video calls imitating mayor of kyiv," *The Guardian*, Jun. 2022. [Online]. Available: https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko
5. S. Dunn, "Women, not politicians, are targeted most often by deepfake videos," Centre for International Governance Innovation, Waterloo, ON, Canada, 2021. [Online]. Available: https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/

6. R. M. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *California Law Rev.*, vol. 107, p. 1753, Jul. 2018. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954
7. F. Wichmann and R. Geirhos, "Are deep neural networks adequate behavioural models of human visual perception?" *Annu. Rev. Vis. Sci.*, vol. 9, no. 1, pp. 501–524, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257880120
8. M. Groh, Z. Epstein, C. Firestone, and R. W. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 1, 2021, Art. no. e2110013119, doi: 10.1073/pnas.2110013119.
9. M. Ramon, "Super-recognizers – A novel diagnostic framework, 70 cases, and guidelines for future work," *Neuropsychologia*, vol. 158, Jul. 2021, Art. no. 107809, doi: 10.1016/j.neuropsychologia.2021.107809.
10. M. Ramon and M. J. Vowels, 2023, "Large-scale super-recognizer identification in the berlin police," OSF, doi: 10.31234/osf.io/x6ryw.
11. M. Mayer and M. Ramon, "Improving forensic perpetrator identification with super-recognizers," *Proc. Nat. Acad. Sci. USA*, vol. 120, no. 20, 2023, Art. no. e2220580120, doi: 10.1073/pnas.2220580120.
12. M. Ramon, A. Bobak, and D. White, "Super-recognizers: From the lab to the world and back again," *Br. J. Psychol.*, vol. 110, no. 3, pp. 461–479, 2019, doi: 10.1111/bjop.12368.
13. M. Ramon and S. Rjosk, *beSure—Berlin Test for Super-Recognizer Identification: Part I: Development*. Frankfurt am Main, Germany: Verlag fur Polizeiwissenschaft, 2022. [Online]. Available: https://www.polizeiwissenschaft.de/suche?query=978-3-86676-762-1
14. J. Nador, T. Alsheimer, A. Gay, and A. Ramon, "Image or identity? Only super-recognizers'(memor) ability is consistently viewpoint-invariant," *Swiss Psychol. Open*, vol. 1, no. 1, pp. 1–15, 2021, doi: 10.5334/spo.28.
15. J. Nador, M. Zoia, M. Pachai, and M. Ramon, "Psychophysical profiles in super-recognizers," *Sci. Rep.*, vol. 11, no. 1, 2021, Art. no. 13184, doi: 10.1038/s41598-021-92549-6.
16. M. Linka, M. D. Broda, T. A. Alsheimer, B. de Haas, and M. Ramon, "Characteristic fixation biases in super-recognizers," *J. Vis.*, vol. 22, no. 8, p. 17, 2022, doi: 10.1167/jov.22.8.17.
17. Prolific, 2021. [Online]. Available: https://www.prolific.com/
18. R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, Version 4.1.0, 2021. [Online]. Available: https://www.R-project.org/
19. R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape," *J. Roy. Statistical Soc. C (Applied Statistics)*, vol. 54, no. 3, pp. 507–554, 2005, doi: 10.1111/j.1467-9876.2005.00510.x.
20. D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Statistical Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.
21. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
22. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, and B. Thirion, "Scikit-learn: Machine learning in Python," *JMLR*, vol. 12, no. 85, pp. 2825–2830, 2011.
23. P. Probst, M. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wires Data Mining Knowl. Discovery*, vol. 9, no. 3, 2018, Art. no. e1301, doi: 10.1002/widm.1301.
24. M. D. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, and F. D. Bastiani, *Flexible Regression and Smoothing: Using GAMLSS in R*. Boca Raton, FL, USA: CRC Press, 2017.
25. D. Lakens, "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," *Social Psychol. Personality Sci.*, vol. 8, no. 4, pp. 355–362, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:39946329
26. M. Ramon and B. Rossion, "Impaired processing of relative distances between features and of the eye region in acquired prosopagnosia—Two sides of the same holistic coin?" *Cortex*, vol. 46, no. 3, pp. 374–389, 2010, doi: 10.1016/j.cortex.2009.06.001.
27. M. Groh, A. Sankaranarayanan, N. Singh, D. Y. Kim, A. Lippman, and R. Picard. "Human detection of political speech deepfakes across transcripts, audio, and video." Papers With Code. [Online]. Available: https://paperswithcode.com/paper/human-detection-of-political-deepfakes-across

Meike Ramon is a Swiss National Science Foundation Promoting Women in Academia group leader and an assistant professor. She leads the Applied Face Cognition Lab and directs the Cognitive and Affective Regulation Laboratory at the University of Lausanne, 1015 Lausanne, Switzerland. Her research interests include face processing and recognition, cognitive neuroscience, and its applications in government and industry. Ramon received a Ph.D. focused on personally familiar face processing in the healthy and damaged brain from UCLouvain. She is a board member of the Association for Independent Research. Contact her at meike.ramon@gmail.com.

Matthew Vowels is a junior lecturer at the Institute of Psychology at the University of Lausanne, 1015 Lausanne Switzerland, a visiting research fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey, and a senior researcher for The

Sense Innovation and Research Center in the department of radiology for the Lausanne University Hospital. His research interests include machine learning, computer vision, causality, and statistics. Contact him at matthew.vowels@unil.ch.

Matthew Groh is a Donald P. Jacobs Scholar and assistant professor in the Department of Management and Organizations at Kellogg School of Management and by courtesy the Department of Computer Science at McCormick School of Engineering, Northwestern University, Evanston, IL 60208 USA. His research interests include human-AI collaboration, computational social science, affective computing, deepfakes, and generative AI. Groh received a Ph.D. in media arts and sciences from MIT. Contact him at matthew.groh@kellogg.northwestern.edu.