Automated Deception Detection of Males and Females From Non-Verbal Facial Micro-Gestures

Keeley Crockett[1], James O'Shea[1], Wasiq Khan[2],
[1]Department of Computing and Mathematics,
Manchester Metropolitan University, Manchester, M1 5GD, UK, K.Crockett@mmu.ac.uk
[2]Department of Computer Science, Liverpool John Moores University, Byrom Street, L3 3AF, Liverpool, UK

***Abstract*—Gender bias within Artificial intelligence driven systems is currently a hot topic and is one of a number of areas where the data used to train, validate and test machine learning algorithms is under more scrutiny than ever before. In this paper we investigate if there is a difference between the non-verbal cues to deception generated by males and females through the use of an automated deception detection system. The system uses hierarchical neural networks to extract 36 channels of non-verbal head and facial behaviors whilst male and female participants are engaged in either a deceptive or truthful roleplaying task. An Image Vector dataset, comprising of 86584 vectors, is collated which uses a fixed sliding window slot of 1 second to record deceptive or truthful slots. Experiments were conducted on three variants of the dataset, all males, all females and mixed in order to examine if the differences in cues generated by males and females lead to differences in the accuracies of machine learning algorithms which classify their behavior. Results showed differences in non-verbal cues between males and females, with both genders at a disadvantage when treated by classifiers trained on both genders rather than classifiers specifically trained for each gender. However, there was no striking disadvantageous effect beyond the influence of their relative frequency of occurrence in the dataset.**

Keywords- micro-gestures, gender, deception detection, machine learning

I. INTRODUCTION

Data bias and the impact it has on users, is an important issue in machine learning when a system is trained, validated and tested on non-representative samples of the population. For example, Amazon had to stop using an Artificial Intelligence (AI) based recruiting tool after it had shown bias against women [1]. Initially the gender bias was traced back to the training data which comprised of resumes which were submitted to the company over 10-years – and consisted of a significantly higher proportion of men. This led to applications not being ranked in a gender-neutral way. In 2019, the Apple card faced a gender bias allegation when news broke that the AI algorithms gave men a higher credit limit than women resulting from biased historical data [2]. If the training data contains stereotypical concepts of gender, the resulting system will propagate this bias [3].

The research presented in this paper examines whether an automated deception detection system detects differences in the way that male and females deceive during an automated interview. The work utilizes a system known as Silent Talker (ST) [4, 5], which is capable of detecting deceptive behaviour by participants at levels significantly better than chance. Previous studies [6, 7] observed that when the groups were made demographically narrower, improved accuracies of deception detection were obtained. In addition to differences based on ethnicity, clear differences were found between males and females regarding the timing of non-verbal indicators within an answer to a question. These differences can be explained as products of human consciousness by the 'Background' theory proposed by the internationally regarded philosopher John Searle [8]. According to Searle the Background supports intentional states (beliefs etc.) consisting of capacities and presuppositions such as abilities, tendencies, habits, dispositions, taken-for-granted presuppositions and "know-how." Intentional states can only function against this Background that enables one to cope with the world. The Background is divided into Universal Background (e.g. we all walk upright, we all eat by putting food in our mouths) and local cultural practices.

This paper investigates if there is a difference in the deceptive / truthful nonverbal behavior between male and female groups when they are interviewed by an avatar which utilizes an automated deception detection system (ADDs). The ADDs system is trained, validated and tested on three datasets representing all females, all males and a mixed gender set of data. In this paper, dataset features are extracted from facial micro-gestures which are captured in real time during an automated interview which uses an avatar. We define a micro-gesture as a tiny gesture such a small head movement, pupil contraction, mouth corner twitch or one eyebrow raised momentarily [6]. Micro-gestures (combined by machine learning techniques) are fundamental components from which larger-scale expressions or gestures may be composed without relying on a particular psychological model for their definition or justification. A micro-gesture can hardly be noticed by the human naked eye without extreme focus on a specific location on the face. The main application area has been within investigations where a person is interviewed, and video recorded. In the past, human experts would then view the videos in slow motion to determine if any micro-gestures indicative of guilty behavior were exhibited by the interviewee. This task is arduous and

also very subjective based on the experience of the expert and their personal cognitive load level [9, 10]. Psychologists have found that the brain can process about 7 tasks simultaneously [9]. Liars use a significant amount of this capability to maintain consistency. In comparison, ADDs monitors 36 channels of micro-gestures and looks at patterns of behaviour across the channels. For definitions and descriptions of the 36 channels, see [7]. ADDs finds inconsistencies in the liar's non-verbal behavior (NVB), for example if people are coached to make eye-contact with the questioner they will make excessive contact showing a disruption in their NVB. The primary research question addressed in this paper is:

Is there a difference between the non-verbal cues to deception generated by males and females?

This produces the hypothesis pair:
H_0: There is no gender effect on the NVB cues to deception produced by males and females.
H_1: There is a statistically significant gender effect on the NVB cues to deception produced by males and females.

If such differences exist, this leads to a secondary question:

Do the differences in cues generated by males and females lead to differences in the accuracies of machine learning algorithms which classify their behavior?

This, in turn, produced the hypothesis pair:
H_0: There is no difference in the classification accuracies for deception of machine learning algorithms, between males and females.
H_1: There is a statistically significant difference in the classification accuracies for deception of machine learning algorithms, between males and females.

This paper is organized as follows; Section II presents related work on the possible influences of gender type in relation to deceptive behaviors and reviews the state of the art in automated deception detection. A data collection experiment is described in section III which uses an automated deception detection system to extract non-verbal behavior during an interview. Section IV describes the experimental methodology used to test the hypotheses with both male and female participants concerning truthful and deceptive conditions. Results and key findings are shown in section V and finally section VI presents the conclusions and future directions.

II. RELATED WORK

A) Influences of Gender on Deceptive Behaviour

Rezki [11] investigated physiological signals associated with liars using a traditional polygraph to determine if there were any differences between males and females. The results indicated that there was a divergence between the two categories in the sensitivity of each gender to specific questions. Work reported in [12] analyzed patterns extracted from thermal, linguistic, and visual responses from a sample of 104 truthful and deceptive participants engaged in three lab based scenarios. The experimental results indicated that that deception was easier to detect among females than males [12]. Lloyd et al. [13] reports that the literature is very inconsistent with regards to the gender effect and deception detection, and outlines the three identified gender effects: (1) Women are better at lie detection than are men; (2) Women are better liars than Men or Men are better liars than are women – different literature and empirical analysis supports different viewpoints; (3) Perceivers are better at detecting lies across gender lines. In a study outlined in [14], signal detection analysis was carried on the Miami University Deception Detection Database (MU3D), a free resource containing 320 videos of target individuals telling truths and lies. In MU3D, eighty participants (20 Black female, 20 Black male, 20 White female, and 20 White male) were recorded speaking honestly and dishonestly about their social relationships [14]. Perceivers where then randomly assigned to videos and asked to answer 4 questions per video, including "*Is this person telling a truth or a lie?*". Whilst this work presents a psychologist's view of trying to answer the question, *Does Gender matter in lie detection?* through human encoding of videos, it does present some interesting findings on understanding gender biases but is not conclusive [13]. Jung and Vranceanu [15] observed a gender effect in lying behavior whilst conducting a sender-receiver experiment that examined the gender interaction between the sender and the receiver and how it led to "dishonest communication strategies" [15]. The study found there was no gender differences in the frequency of lying but found that "*men tend to state bigger lies than women, and state the largest lies when paired with a woman."* [15].

In work presented in this paper, we argue that if there is a difference in observed behavior from a human perspective (as suggested by the literature), then this difference should be reflected, to a degree within an automated deception detection system, assuming that the human and the machine both make decisions on a person's non-verbal behavior.

B) Automated Deception Detection Systems

The need or desire to detect deception has been with us throughout human history. Humans have largely relied on subjective intuitions to judge others as truthful or deceptive. Nevertheless, from the earliest possible times, there have been attempts to use non-verbal behaviour such as rubbing the roots of the hair with the fingers (Vedas 900 BC) or use scientific measures such as measuring the pulse (Erasistratus 300-250 BC) to detect deception [16]. Medical developments in the late 19th century produced instruments capable of making objective measurements of pulse and blood pressure. Experiments were performed with these instruments by Lombroso to question suspects in robbery and murder cases at the turn of the 20th century [16]. The experiments paved the way for the invention of the Polygraph by Larson [17]

which remains the best-known lie detector. Even objective measurements from instrumentation may be interpreted subjectively by a human interviewer. Various methods for formalizing result interpretation by polygraph examiners were proposed around 2009 and there has been some interest in automated analysis. As reported in 2018, US security agencies still relied on human polygraph examiners [18]; the FBI requires 5 years of investigative experience as a special agent and training in approved Polygraph examiners course for its interviewers [19].

Two prominent, AI driven, automated deception detection systems are Silent Talker [5] and AVATAR [20]. Silent Talker, designed to be used in a natural interviewing situation, classifies multiple visible signals from the upper body to create a comprehensive time-profile of a subject's psychological state [5]. AVATAR, a kiosk based system has been trialed at US-Mexico, US-Canada and selected EU borders, with reported deception detection accuracies of between 60-80% [21]. Both systems are reported to be consistently above human accuracy and not subjective. Apart from research on the use of instrumentation, the recent focus has been improving interviewing techniques [5] or on developing automatic (AI) systems to analyse the results of instrumentation [22]. Early protagonists of AI deception detection were keen to point that machines are not subject to fatigue and are free of human bias [6, 7]. In reality there are serious concerns about bias in machine learning AI systems due to lack of diversity in the developers (the "white guy problem") [23] or poor representation of the general population in the developers or datasets [3]. This has been widely publicized in controversial arguments about the COMPASS prison release system [25]. Consequently, there is need for controlled experiments to determine the susceptibility of AI algorithms to learning bias towards minorities.

III. Data Collection through an Automatic Deception detection system

A) Data Collection Methodology

Raw video data was collected from 32 participants (22 male and 10 female) who consented to take part in role playing activity which was either a truthful or deceptive task. The task involved first packing the contents of a suitcase that the participant had packed for a holiday where they would be travelling from an airport. During the task, participants were interviewed through an automated interviewing system with each video interview lasting between 3 and 6 minutes depending on the detail given in the answer. Data was captured using a web-cam using the default video resolution of 640*480 and 30 frames per second (fps). Using the Silent Talker system, image vectors were extracted from the participant videos. The image vectors were comprised of a collection of 36 non-verbal channels from the Object Locator ANNs' outputs, the Pattern Detector ANNs' outputs, facial geometrical calculations and logical expressions. The three categories of channels were related to facial movement (15 channels), eye position and movement (16 channels) and the angle of the face (5 channels). To extract the channel data, a fixed sliding window slot of 1 second (30 frames per second) was used to collate information on channel states. This information was used to formulate the Image Vector dataset which was used in this study.

B) The Image Vector dataset

The full Image Vector data set comprised of 86584 rows of data, split between make and female vectors as shown in Table I. The imbalance in the dataset reflects the imbalance in the gender of the participants i.e. 22 male and 10 female. For a vector to be included in the Image Vector dataset, the slot where the vector was extracted must be valid. A valid slot is one where all channel information is present from all 36-non-verbal channels. If for example, a participant turns their head during the role playing activity and one eye is not visible to the camera then these image vectors fall below the threshold and are not included in the dataset. Other factors that can reduce the number of vectors included are 1) poor lighting, incorrect positioning in relation to the webcam and failure to follow instructions of the role-play activity. Each vector in the dataset is labelled either (-1) Truthful or (1) Deceptive based on the whether the person was undertaking a truthful or deceptive role-playing task.

Table I: Image Vector Dataset Description

Gender	Truthful (-1)	Deceptive (1)	Total Vectors
Male Vectors (1)	34618	25581	60499
Female Vectors (-1)	8432	17653	26085

IV. Experimental Methodology

This section describes the methodology to conduct a quantitative empirical study of non-verbal behaviour with samples of volunteer participants concerning truthful and deceptive conditions. The hypothesis pair tested was:

H_0: There is no gender effect on the NVB cues to deception produced by males and females.
H_1: There is a statistically significant gender effect on the NVB cues to deception produced by males and females.

In order to maximize the sensitivity of the test, all of the vectors were used. For each condition, the vectors were randomly split, 50/50, into mutually exclusive training and testing sets. Consequently, the results shown in table II are derived from 26,085 female vectors, table III results are derived from 60,499 male vectors and table IV results are derived from 86,584 combined male and female vectors. The implications of these choices will be included in the discussion section. A number of well-known machine learning (ML) algorithms were trained on each dataset (the "J48 Best" entries are for the optimally pruned J48 trees). These are presented in tables II – IV). ZeroR is the baseline model used. The ZeroR rule simply guesses that every vector

in a dataset belongs to the majority class. For comparative purposes a number of representative and common ML algorithms are used including the decision tree J48 (based on Quinlan's C4.5 algorithm), a simple multi-layer perceptron (MLP) and the Naïve Bayes probabilistic classifier. Decision trees were used to offer some degree of explainability on the interactions between non-verbal channels providing transparency in the decision making process to expert stakeholders [27]. Additionally, the Weka attribute ranker was used to determine the relative importance of the channels (non-verbal behavior cues) for each gender and a comparison made. The top 10 ranked non-verbal channels for males and females are shown in table V.

V. RESULTS AND FINDINGS

A) Results

The results show that ML algorithms operate in a similar manner to an instrument based deception detection system where the system is calibrated with baseline information from the participant (e.g. polygraph and other biometric methods). Consequently, the results should not be taken as a claim for the performance of ADDS deployed in the real world (which would not have seen sample of the interviewee's non-verbal behavior in advance). Therefore, it is stressed that the focus of this paper is not on the accuracy of ADDs, but to examine the *gender effect on the NVB cues to deception produced by males and females.*

Table II shows the % classification accuracy overall and for both deceptive and truthful classes for a number of ML algorithms using only females within the Image Vector dataset.

Table II: Results for Female Gender

Model	%Accuracy	%Deceptive Correct	%Truthful Correct
ZeroR	67.3%	100%	0%
J48 Default	97.5%	98.0%	96.2%
J48 Best *	94.7%	92.6%	95.8%
Naïve Bayes	77.3%	80.8%	70.1%
Random Forest	99.8%	100%	99.5%
MLP	99.6%	99.7%	99.4%

* We define a best pruned tree as one in which MNO (minimum number of objects) is just below the number that would cause a significant reduction in classification accuracy from the default MNO (=2).

The best pruned tree was 94.7% with MNO=11 (Table II). The J48 default tree contained 449 leaves (897 nodes in total), with the channel *lright* (movement of the left eye to the right), being the most significant node. The J48 best pruned tree on females comprised of 350 leaves and 699 nodes with again the channel *lright* being the most significant node.

Table III shows the results of experiments conducted using only the male image vectors.

Table III: Results for Male Gender

Model	%Accuracy	%Deceptive Correct	%Truthful Correct
ZeroR	57.2%	0%	100%
J48 Default	97.5%	97.0%	97.9%
J48 Best *	96.9%	97.2%	95.8%
Naïve Bayes	75.6%	68.1%	81.1%
Random Forest	99.8%	99.9%	99.7%
MLP	96.4%	96.1%	96.7%

In Table III, the J48 default tree contained 746 leaves (1491 nodes in total), with *lhleft (*movement of the left eye half to the left position*)* being the most significant node. This was also reported for the J48 best pruned tree which gave 96.9% with MNO=9. In order to assess the effect on the classification accuracy using an unbalanced and un-representative sample (as a stress-test using the conditions under which bias would be expected to occur), the same ML algorithms were run on the full Image Vector dataset (Table IV).

Table IV: Results for Male and Female (Full dataset)

Model	%Accuracy	%Deceptive Correct	%Truthful Correct
ZeroR	50.4%	100%	0%
J48 Default	96.5%	96.5%	96.5%
J48 Best *	95.4%	94.1%	96.7%
Naïve Bayes	70.1%	74.6%	65.1%
Random Forest	99.8%	99.8%	99.8%
MLP	91.7%	92.9%	90.8%

To further investigate the differences in classification performance between the two genders, significance tests were performed, these were the 2-sample t-test and the N-1 Chi Square test, The results are summarised in table V. Model is the type of classifier, *%CA difference* is the difference in classification accuracy (truthful and deceptive cases) between the male and female specialized classifiers, p-value t-test is the p-value for the two-sample t-test (independent groups) and p-value χ-square is the p-value for the N-1 Chi Square test.

Table V: Variation of overall CA between genders

Model	%CA difference	p-value t-test	p-value χ-square
J48 Default	0.00 %	1.0	1.0
J48 Best *	2.20 %	< 0.01	< 0.01
Naïve Bayes	1.70 %	< 0.01	< 0.01
Random Forest	0.00 %	1.0	1.0
MLP	3.20 %	< 0.01	< 0.01

The experiment detected no difference between the treatment of males and females by the J48 default and random Forrest classifiers. There were small (but significant) differences between them when classified by the best (pruned) J48, Naïve Bayes and MLP classifiers. In order to assess the importance of the non-verbal channels in the decision-making process, Weka's Information Gain attribute ranker [26] was applied and the top 10 influential channels are shown in Table V.

Table V: Non-verbal Channel Ranking

Attribute Rank	Females	Males
1	ffm	fbm
2	fbm	ffm
3	fmc	fmac
4	fmac	fmc
5	lright	lhleft
6	lclosed	rhright
7	rleft	lright
8	fs	fs
9	rclosed	lhright
10	lhclosed	fmuor

On analysis of the channels in Table V, it can be seen that the top 4 channels are important cues to deception but appear in a different order for males and females. Furthermore, different cues appear in the top 10 between males (lhleft, rhright, lhright and fmour) and females (lclosed, rleft, rclosed and lhclosed) A more formal indicator of the relationship between the two genders can be shown by the Spearman rank correlation coefficient (ρ) which gives 0.75 with $p < 0.01$, a strong and significant correlation.

C) Discussion

It should be noted that in tables II and III the 100% / 0% distributions for the ZeroR rule are reversed. There is no significant gender effect shown by this, it is a simple outcome of the difference in distributions of the classes between the two genders. ZeroR is not an AI classifier, it is a baseline measure to help understand the performance of classifiers.
Using the 2-sample t-test, initial analysis of tables II-IV shows that the lowest classification accuracy for males is 1R, significantly different from the ZeroR score ($t = 48.148$, DoF = 120996, $p < 0.01$). The same is true for the females ($t = 38.264$, DoF = 52168, $p < 0.01$). Therefore all of the classifiers are performing better than chance for both genders.

If decision trees are allowed to grow, without any constraints during training, they may become over-trained – effectively memorizing the data set rather than extracting principles from it (Overfitting). Consequently, the size of a decision tree may be constrained by pruning parameters. Pruning was performed in these experiments using Minimum Number of Objects (MNO) pruning, which does not permit leaves of the tree to exist which would contain fewer than the set number of cases from the training set. Further analysis of tables II and III was performed by comparing the relative advantages of being classified separately (for each gender) as opposed to being classified collectively. The average improvement in classification accuracy for males was 3.3% and the average for females was 6%.

The channel rankings provide an insight into the relative importance of the different non-verbal channels between females and males. The Spearman ρ between the two genders was calculated as 0.75 with $p < 0.01$. A rule of thumb for interpreting correlation coefficients in [28] describes a value of ρ between 0.70 to 0.89 as "A strong correlation." Additionally, the p-value for this correlation is <0.01. As ρ is clearly < +1.0 we can conclude that the relative importance of the set of NVB cues to deception is not identical. However, we can state that we have found strong evidence to support the view that there is high similarity, between these cues for females vs. males in this dataset. This finding is in agreement with Searle's concept of the Universal Background [8]. It may also be interesting to examine the nature of the channels which were most different in ranks and which were identical in ranks for the two genders.

The most different were lhleft and lleft (16 rank positions difference) followed by fma (14 rank positions difference). lhleft is "left eye half left", lleft is "left eye left" and fma is "face movement angle-change." These are two channels from the group "eyes" and one from the group "face angle." The identical channels were fbla, fs, fvs, lblink (all with 0 rank positions difference). fbla is "face blanch", fs is "face scale change", fvs is "face vertical shift" and lblink is "left eye blink." These are 3 channels from the group "face" and one from the group "eyes." Although there is some difference between the kinds of channels that are most similar and most different between males and females, one should be wary of reading too much into this. The next three channels in terms of rank difference come from the "face" and "eye" categories.

The differences in the classification accuracy shown in Table *V* were small. Despite the small size of these differences, they were detected as significant due to the relatively large sample sizes, in terms of vectors. It is noted that the two statistical tests agreed on the statistical significance, from their different perspectives.

An overview of the performance of ML across a range of classifiers was obtained by averaging their performances, in particular by averaging the improvement in classification accuracy for each gender by classifying it separately, compared with classifying that gender using a classifier trained on a mix of both genders. This suggested an advantage for males (6%) compared with females (3.3%). However, it would be hasty to take this as evidence that the females have been subject to an inherent disadvantage as the combined set it skewed towards the males. One attempt to normalize this would be to multiply the female CA by the ratio of males to females in the dataset, giving 7.7%, close to the figure for males.

VI. Conclusions and Further Work

This paper has investigated the influence of gender in the classification of deception from nonverbal behavior. In addressing the research question "*Is there a difference between the non-verbal cues to deception generated by males and females?"*

The evidence from this investigation suggests that the NVB cues (channels) in this dataset are highly similar in their importance but are, nevertheless, different. Both genders appear to be disadvantaged when treated with a combined classifier than when they are treated with classifiers tailored for their gender. There is a gender effect between classifiers

trained specifically for each gender. In the 3 cases there is a significant difference (but very small) between versions of the same classifier trained for different genders. The more interesting finding is that examining the top 10 NVB cues (individually, with OneR) shows differences in the relative power of the cues, even though there is a high degree of similarity.

The second question, "*Do the differences in cues generated by males and females lead to differences in the accuracies of machine learning algorithms which classify their behavior?"* may be addressed by examining the differences in accuracy of classification of truthful and deceptive cases between male and female for the various classifiers. However, due to the relatively large size of the dataset, the T-test for differences between percentages shows even the smallest of differences to be significant. Nevertheless we can reflect on the figures. For the specialised female classifier, four out of five classify the deceptive cases more accurately. For the specialised male classifiers, three out of the five classify the truthful cases more accurately. It would be unwise to interpret this as an inherent bias of AI against females. If these classifiers were used in real-world applications, the practical outcome would be that more female deceivers would be classified as truthful and more truthful males as deceivers. In fact, the aggregates of the differences between deceptive and truthful for each gender (D-T), are almost mirror images of each other, +13.3 / -3.2 for females, +1.6 / -14.5 for males. In reality, classifiers developed with the experimental methodology described here will never be deployed in a practical application. NVB deception detection falls under the domain of Signal Detection Theory (SDT), which provides a theoretical basis for setting a valid discrimination threshold for purposes of classification. In a developing a real system, as well as using larger volumes of data for improved modelling, principles of SDT will be used to set appropriate boundaries for risk scores coming from ST to determine the classification of truthful vs. deceptive.

Further work should include larger scale experiments for three reasons. To balance the dataset, to achieve greater statistical power in determining the significance of gender differences and to investigate whether larger training sets will lead to better ML models which server both genders effectively. It should also be noted that gender and ethnicity attributes were excluded (as channels) from the dataset. Furthermore, the nature of the other channels does not support the identification of the gender that the vectors in the dataset belong to. This makes the classifiers developed in this study particularly robust to developing gender- based bias.

Further work will also involve a more detailed analysis of the explanatory power of individual cues to deception (channels) and combinations thereof, and how such combinations should be selected.

Finally, it may be argued that the hard binary gender divide in the dataset is inappropriate for the more fluid view of gender in the modern world. For example, Facebook introduced a set of 58 gender categories in 2014 [29]. Nonetheless, Bivens [29] also reported that beneath the 58 user-declared options, Facebook reconfigured them into a binary system. Future studies should include more gender options, but the use of binary gender in this study fits its purpose of investigating differences based on gender.

REFERENCES

[1] Dastin, J. (2018), Amazon scraps secret AI recruiting tool that showed bias against women, [online], Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, [Accessed 22/11/2018]

[2] Barkho, G. (2019), How Goldman Sachs Can Regain User Trust After Apple Card Discrimination, [online], Available: https://observer.com/2019/11/goldman-sachs-bias-detection-apple-card/ [Accessed 04/01/2020]

[3] Leavy,.S (2018), Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning, 2018 ACM/IEEE 1st International Workshop on Gender Equality in Software Engineering,pp14-16. DOI 10.1145/3195570.3195580

[4] Al Bandar, Z.G., McLean, D.A., O'Shea, J.D. and Rothwell, J.A. (2015), Manchester Metropolitan University. Methods and apparatus for analysing the behaviour of a subject. U.S. Patent 8,992,227.

[5] Silent Talker Ltd, (2020), [online], Available at: https://www.silent-talker.com/ [Accessed 5 Jan. 2020]

[6] Rothwell J (2002) artificial neural networks for psychological profiling using multichannels of nonverbal behaviour. PhD Thesis, Manchester Metropolitan University

[7] Rothwell, J., Bandar, Z., O'Shea, J. and McLean, D., (2006). Silent talker: a new computer-based system for the analysis of facial cues to deception. Applied cognitive psychology, 20(6), pp. 757-777.

[8] Searle, J.R., (1980). The background of meaning. In Speech act theory and pragmatics (pp. 221-232). Springer, Dordrecht.

[9] Miller, G.A., 1994. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological review, 101(2), p.343.

[10] Slavin, R.E., (2019). Educational psychology: Theory and practice.

[11] Rezki, M., (2019). Detecting Lie-A Practical Approach. Revue d'Intelligence Artificielle, 33(2), pp.127-132

[12] Abouelenien, M., Burzo, M., Pérez-Rosas, V., Mihalcea, R., Sun, H. and Zhao, B., (2018). Gender Differences in Multimodal Contact-Free Deception Detection. IEEE MultiMedia

[13] Lloyd, E.P., Summers, K.M., Hugenberg, K. and McConnell, A.R., (2018), Revisiting Perceiver and Target Gender Effects in Deception Detection. Journal of Nonverbal Behavior, 42(4), pp.427-440.

[14] Lloyd, E.P., Deska, J.C., Hugenberg, K., McConnell, A.R., Humphrey, B.T. and Kunstman, J.W., (2019). Miami University deception detection database. Behavior research methods, 51(1), pp.429-439.

[15] Jung, S. and Vranceanu, R., (2017). Experimental Evidence on Gender Differences in Lying Behaviour. Revue économique, 68(5), pp.859-873.

[16] Trovillo, P.V., A History of Lie Detection (1939). J. CRIM. L., 29, p.848.

[17] Grubin, D. Madsen, L. (2005) Lie detection and the polygraph: A historical review, The Journal of Forensic Psychiatry & Psychology, 16:2, 357-369, DOI: 10.1080/14789940412331337353

[18] Burst, A. (2018), To tell the truth: What current and hopeful federal employees should know about polygraphs, [online], Available: https://federalnewsnetwork.com/explainers/2018/08/to-tell-the-truth-how-federal-agencies-use-polygraphs-in-hiring-and-screening, [Accessed 04/01/2020]

[19] Agencies using polygraphs, [online], Available: https://oig.justice.gov/reports/plus/e0608/results1.htm, [Accessed 04/01/2020]

[20] Marsh, A. A Brief History of the Lie Detector (2019), [online], IEEE Spectrum, Available: https://spectrum.ieee.org/tech-history/heroic-failures/a-brief-history-of-the-lie-detector, [Accessed 04/01/2020]

[21] Daniels, J. (2018), Lie-detecting computer kiosks equipped with artificial intelligence look like the future of border security, [online], Available: https://www.cnbc.com/2018/05/15/lie-detectors-with-

artificial-intelligence-are-future-of-border-security.html, [Accessed 04/01/2020]
[22] Vrij, A., Granhag, P.A., Mann, S. and Leal, S., (2011). Outsmarting the liars: Toward a cognitive lie detection approach. Current Directions in Psychological Science, 20(1), pp.28-32.
[23] O'Shea, J., Crockett, K., Khan, W., Kindynis, P., Antoniades, A. and Boultadakis, G., (2018), July. Intelligent Deception Detection through Machine Based Interviewing. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
[24] Crawford, K. (2016), Artificial intelligence's white guy problem, The New York Times, [online], Available: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html , [Accessed 04/01/2020]
[25] Flores, A.W., Bechtel, K. and Lowenkamp, C.T., (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. Fed. Probation, 80, p.38.
[26] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., (2018). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
[27] Crockett, K. Stoklas, J. O'Shea, J. Krügel, T. Khan, W. Reconciling Adapted Psychological Profiling with the New European Data Protection Legislation (2020), Computational Intelligence, Eds: Sabourin, C. Mereio, J. Barranco, N. Madani, K. Warwick, K. Springer, *in-press*
[28] Fowler, R.L., (1987). Power and robustness in product-moment correlation. Applied Psychological Measurement, 11(4), pp.419-428.
[29] Bivens, R., (2017). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. New Media & Society, 19(6), pp.880-898.