# Appraising the AVATAR for Automated Border Control

## Results of a European Union Field Test of the AVATAR System for Interviewing and Passport Control

**Conducted at the Henri Coandă International Airport, Bucharest, Romania, December 2013**

by Aaron Elkins, Elyse Golob, Jay Nunamaker, Judee Burgoon, and Douglas Derrick

October 2014

*a report to*

FRONTEX

*prepared by the*

National Center for Border Security and Immigration

BORDERS

THE UNIVERSITY OF ARIZONA

blank page

# Appraising the AVATAR for Automated Border Control

## Results of a European Union Field Test of the AVATAR System for Interviewing and Passport Control

Conducted at the Henri Coandă International Airport,
Bucharest, Romania, December 2013

by Aaron Elkins, Elyse Golob, Jay Nunamaker, Judee Burgoon,
and Douglas Derrick

October 2014

*a report to*

FRONTEX

*prepared by the*

National Center for Border Security and Immigration

BORDERS THE UNIVERSITY OF ARIZONA.

## About the Authors

**Aaron Elkins** is a postdoctoral researcher with BORDERS at the University of Arizona. He investigates how the voice, face, body, and language, reveal emotion, deception, and cognition for human-computer interaction (HCI) and artificial intelligence applications. **Elyse Golob**, Ph.D., is the executive director of BORDERS at the University of Arizona. Her expertise includes cross-border trade, economic development, and border management policy. **Jay Nunamaker** is a Regents' Professor and the Soldwedel Chair of Management Information Systems at the University of Arizona. He is director of BORDERS and of the Center for the Management of Information. Professor **Judee Burgoon** is director of research for the Center for the Management of Information at the University of Arizona. Her research focuses on nonverbal communication, deception, trust, interpersonal interaction, and new technologies. **Douglas Derrick** is an assistant professor of IT innovation at the University of Nebraska at Omaha. He received a Ph.D. in management information systems from the University of Arizona. His research interests include human-agent interactions, intelligent agents, data fusion, decision support systems, and persuasive technology.

Elkins, Aaron, Elyse Golob, Jay Nunamaker, Judee Burgoon, and Douglas Derrick. 2014. *Appraising the AVATAR for Automated Border Control: Results of a European Union Field Test of the AVATAR System for Interviewing and Passport Control.* Tucson: National Center for Border Security and Immigration (BORDERS), University of Arizona.

## National Center for Border Security and Immigration (BORDERS)

BORDERS is a consortium of 18 premier institutions that is dedicated to the development of innovative technologies, proficient processes, and effective policies that will help secure international borders, foster international trade, and enhance long-term understanding of immigration determinants and dynamics.

BORDERS | The University of Arizona
McClelland Hall, Room 427
P.O. Box 210108
Tucson, AZ 85721-0108
520.621.4475
www.borders.arizona.edu



This report was edited by Robert Merideth.

The authors appreciate the support provided by Frontex to the AVATAR research program. Special thanks go to Edgar Beugels, head of research and development, and to Monica Gariup, senior research officer, and to her dedicated staff. The authors also appreciate the cooperation of the Romanian Border Police for its assistance in the Bucharest field test, and thank the many European border guards who participated in the AVATAR workshops described in this report. Their contributions were invaluable.

# Table of Contents

blank page

# Executive Summary

The Automated Virtual Agent for Truth Assessment in Real-time (AVATAR) is a system designed to automate screening, interviewing, and credibility assessment of persons crossing international borders (traveling by air, land, or sea transport). The AVATAR was developed by researchers at the National Center for Border Security and Immigration (BORDERS), a United States Department of Homeland Security Center of Excellence headquartered at the University of Arizona.

The AVATAR conducts primary and secondary screenings of entrants using a virtual human agent and an array of non-invasive sensors to automate the analysis of a person's credibility, documents, and identity. In addition to automating screenings, the AVATAR is designed to support border guards and operational staff by providing a real-time risk score of an individual to indicate if additional scrutiny or investigation of the person is warranted. By having one officer manage multiple AVATAR kiosks, the system is designed to improve border control efficiency and effectiveness.

## BORDERS–FRONTEX COLLABORATION

Since 2009, BORDERS has collaborated on the development of the AVATAR with Frontex through annual workshops, experiments, and field tests. The goals of these activities were: (1) to introduce border guards to the AVATAR system, (2) to learn the requirements of border guards and other end-users and to collect their feedback in using the system, and (3) to explore the technology's potential utility for the European Union.

Each workshop built upon the outcomes and lessons of the previous one, culminating in the 2013 field test at the Henri Coandă International Airport in Bucharest, Romania. The results of this trial support the maturity of the current AVATAR for further field testing and applications. The automated interviewing, biometric identification, document authentication, and credibility assessment features performed reliably and were consistent with previous lab testing.

## WORKSHOPS AND EXPERIMENTS

Major outcomes from the workshops (and their respective experiments) conducted in cooperation with Frontex include:

- During the **fake bomb experiment** (2010 workshop), the AVATAR identified 100% of the "guilty" participants with an overall accuracy of 87% using fused (aggregated) data from vocalic and eye-behavior sensors.
- In the **document fraud experiment** (2011 workshop), the AVATAR achieved an overall accuracy rate of 95% with fused vocalic and eye-behavior sensor data.

- The **hooligan experiment** (2012 workshop) demonstrated that the AVATAR's accuracy rate surpassed that of humans in identifying imposters and passing innocent participants. The AVATAR identified 85% of imposters, while humans detected only 67%. Humans had up to an 80% false positive rate of "innocents," while the AVATAR's false positives were 15%.

## AIRPORT FIELD TEST

A major objective of the 2013 Bucharest, Romania, field test was to assess the newest generation AVATAR in an operational setting. Major findings included:

- **Processing time**. The AVATAR interview took an average of 30 seconds for the shortest interview type (Romanian citizen) and an average of 90 seconds for the longest interview type (third country national from a non-visa-waiver country).
- **Reducing false positives**. Across all interviews, only 2% of passengers were flagged as high risk and 15% as low/medium risk. Because ground truth was unknown, specific accuracy measures could not be calculated.
- **Passenger acceptance**. Passengers overwhelmingly liked using the AVATAR (90%), indicated they would use it in the future (75%), and found it easy to use (96%).
- **Operational performance**. The AVATAR performed well in terms of improved sensing, auto-height adjustment, installation, maintainability, and ruggedness for daily use.

## FUTURE RESEARCH

The AVATAR has the potential to improve and extend existing automated border control (ABC) technology in the EU. While existing systems provide automated biometric-based identification and document authentication, they lack behavioral and physiological screening to assess credibility. Future areas for AVATAR research in the EU may include:

### Smart Borders initiatives

- Integrating the AVATAR into existing screening procedures, including e-Gate technologies
- Testing the AVATAR as a next generation ABC technology to facilitate an Entry-Exit System (EES) and Registered Traveler Program (RTP) for EU citizens and third-country nationals (TCNs)

### Advanced behavioral analysis

- Incorporating computer vision and automated facial analysis to detect imposters
- Using biometric identification augmented by behavioral analysis to detect imposters

### Countermeasures

- Testing methods to prevent attempts to defeat the AVATAR's screening
- Identifying situations and passenger types that affect the AVATAR's detection accuracy

### Border guard decision support

- Optimizing display of the AVATAR's information to officers to improve decision making
- Determining the optimal number of AVATAR kiosks that individual border officers can reasonably monitor and maintain vigilance

### Passenger acceptance

- Testing more realistic 2D or 3D AVATAR interviewing agents to enhance passenger interaction and acceptance

This report describes the AVATAR system and its underlying research. It summarizes the BORDERS–Frontex workshops and experiments, and provides full details of the Bucharest, Romania, airport field test. The report also identifies further research areas for the AVATAR in the EU context.

# 1. Introduction

Every day, millions of people cross international borders at ports of entry for air, land and sea transportation, with a steady growth in such arrivals projected. This constant stream of passengers brings never-ending challenges to border agencies that manage national security. Although officers must screen all travelers while ensuring the flow of legitimate travel and trade, there is no surefire method to detect imposters, false documents, or contraband. No single device or indicator ("Pinocchio's nose") can accurately identify threats or deception, while minimizing false alarms and false negatives. If not based on relevant facts, an officer's assessment of an individual may be inaccurate, arbitrary, ineffective, or discriminatory ("gut feeling").

A "one-size-fits-all" screening approach may create long queues, customer inconvenience, and high personnel costs. Alternatively, scaled-down checks may permit entry to dangerous persons and contraband. A risk-segmentation approach, such as a trusted traveler program, affords minimal screening to low-risk individuals. Although these programs are becoming more popular, many individuals who may qualify do not enroll due to high costs, lack of knowledge, or infrequent travel schedules.

The Automated Virtual Agent for Truth Assessment in Real-time (AVATAR) is designed to address these limitations. The AVATAR can conduct primary and secondary screenings using an array of non-invasive sensors to detect suspicious persons and to verify travelers' identities and documents while ensuring individual privacy. These data are fused and analyzed to provide officers with a real-time risk score to determine if further investigation of a person is warranted. The AVATAR can act as a "force multiplier," reducing the workload of border officers while identifying potential threats and providing risk assessments to expedite the flow of legitimate travel and trade.

This report describes the AVATAR system, the system's underlying research, the series of BORDERS–Frontex workshops and experiments held 2010–2012, and a culminating field test in 2013 at the Henri Coandă International Airport in Bucharest, Romania.

## BACKGROUND

The National Center for Border Security and Immigration (BORDERS) is a United States Department of Homeland Security (DHS) Center of Excellence headquartered at the University of Arizona. Its mission is to provide scientific knowledge, develop and transition technologies, and evaluate policies to meet the challenges of border security and immigration.

Since its inception in 2008, BORDERS has established joint ventures with international institutions engaged in border management and research. In 2009, BORDERS

initiated a multi-year collaboration with Frontex to conduct a series of workshops, experiments, and field tests. The purpose was to introduce border guards to the AVATAR, collect end-user requirements and feedback, and explore the technology's potential utility for the European Union (Table 1). Details on these events are provided later in this report.

**Table 1. BORDERS-Frontex collaborative activities**

| YEAR | WORKSHOPS, EXPERIMENTS, AND FIELD TEST | LOCATION |
|---|---|---|
| 2010 | WORKSHOP: Artificial Intelligence for Screening and Decision Support at Border Crossings (including the "fake bomb" experiment) | Warsaw, Poland |
| 2011 | WORKSHOP: Improving Border Checks with Next-generation Artificial Intelligence and Advanced Sensor Technology: Decision Support for Assessment, Screening and Interviewing (including the "document fraud" experiment) | Warsaw, Poland |
| 2012 | WORKSHOP: Simulation on Passenger Risk Assessment, Joint Operation Champions League (including the "hooligan" experiment) | Apeldoorn, Netherlands |
| 2013 | FIELD TEST: AVATAR for Automated Interviewing and Passport Control at the Henri Coandă International Airport | Bucharest, Romania |

## AVATAR SYSTEM


**Figure 1. AVATAR kiosk**

The AVATAR (Figure 1) is a kiosk-based automated screening system for credibility assessment developed by BORDERS. The AVATAR conducts natural, brief interviews in a number of screening contexts, such as trusted traveler programs, pre-employment applications, personnel reinvestigations, visa adjudications, and other scenarios that entail credibility assessments. During the interaction, the kiosk's non-contact sensors measure and identify suspicious or irregular behavior that warrants further investigation.

Unlike traditional screening technologies, the AVATAR incorporates a virtual human agent who conducts a fully automated interview (Figure 2). During this process, the AVATAR scans and authenticates travel documents, such as passports and visas, identifies passengers through biometrics, and conducts a

**Figure 2. AVATAR interviewing agent**

credibility assessment tailored to each passenger. As indicated, the interaction is natural and brief to facilitate rapid screening and to conduct risk segmentation. The interviewer's demeanor, visage, and questions can be customized. Based on a real-time analysis of the passenger's answers and behavior, the AVATAR dynamically adjusts its questioning to elicit the most diagnostic responses and cues. The interview results are transmitted wirelessly to an officer who uses this information to make a final decision on the passenger's admittance.

When a passenger approaches the kiosk, the AVATAR adjusts its height and calibrates its interface and sensors in response to the individual's stature, culture, and language. The kiosk then collects fingerprints and facial biometrics, and scans and reads the passport for identification. Next, the AVATAR conducts a screening interview based on the traveler's identity (taking into account immigration status and travel history). During this voice-based interaction, the AVATAR's multiple behavioral and physiological sensors measure the person's nonverbal and verbal behavior to dynamically adjust the interview and assess credibility. The sensor results are then fused and analyzed to rate the passenger's credibility and risk. All results are transmitted in real-time to an attending officer via a mobile tablet (Figure 3).

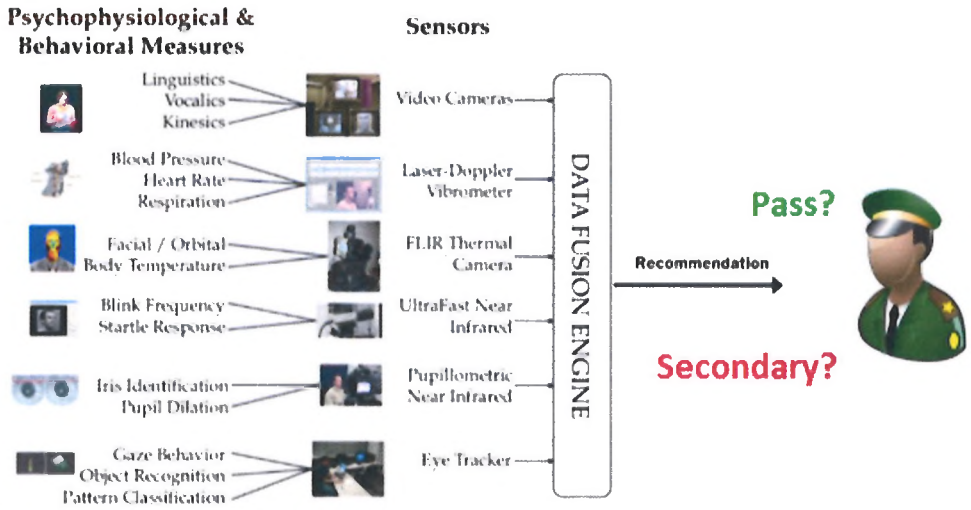


**Figure 3. Behavioral and physiological sensor fusion and credibility assessment**

The AVATAR has been field-tested for border security applications in the United States and in Europe. For example, in conjunction with the DHS Customs and Border Protection (CBP), the AVATAR was tested in 2012 at the Nogales, Arizona, Port of Entry Trusted Traveler Application Center (where candidates apply for expedited processing as pre-approved, low-risk travelers). In cooperation with Frontex, the AVATAR

was operationally tested in 2013 at the passport control facility at the Henri Coandă International Airport in Bucharest, Romania.

## KEY FUNCTIONS

The AVATAR contains several sensors, scanners, readers, and other tools to facilitate operational processes such as conducting interviews, collecting biometric identification and payments, scanning documents, and sending wireless alerts related to the identified risk (Figure 4). This section describes the key elements of the AVATAR.

### Intelligent agent interview

A critical capability of the AVATAR is the automation of interviews. The AVATAR incorporates video monitors to display the **embodied virtual agent** (EVA) to conduct interviews while the sensors assess other cues.

Border officers expend much time screening low-risk individuals and diverting suspicious individuals to secondary screening. These time-intensive interactions incur high personnel costs. By deploying multiple AVATARs at a checkpoint, one officer can monitor several interviews.

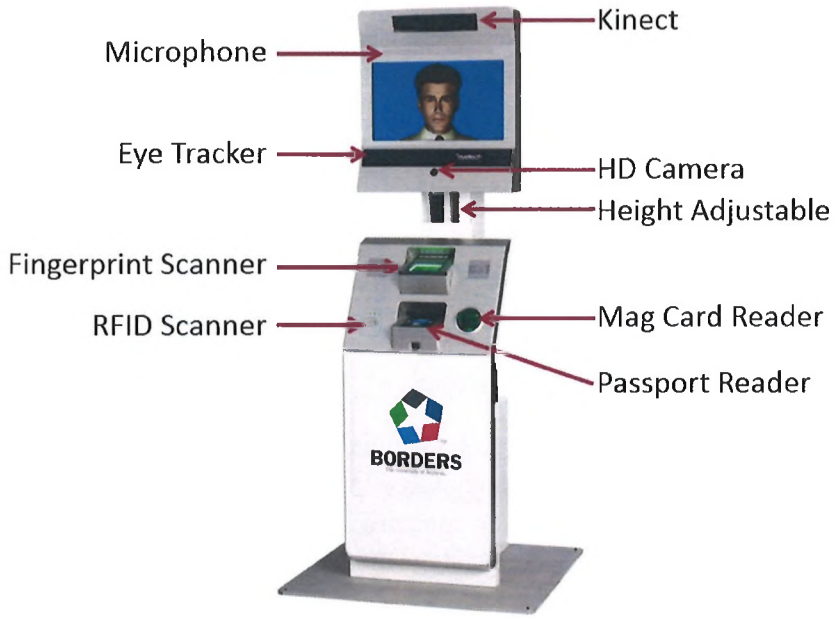


**Figure 4. AVATAR kiosk**

### Identification and authentication

The system includes a biometric **fingerprint scanner** to confirm passenger identity. It also includes a **passport reader** to authenticate via eye tracking algorithms (see below) that the document presented belongs to the person.

### Sensors

Because there is no one cue that indicates deception, the AVATAR uses non-contact sensors to gather multiple behavioral and physiological cues correlated with deception. These non-invasive sensors facilitate rapid and unobtrusive passenger processing. These include:

- **Eye tracker.** This oculometric sensor measures gaze fixations and movements (saccades), and pupil dilation.
- **Two-dimensional video cameras.** These sensors capture kinesic (facial, head, postural, and gestural movement) and proxemics (spatial distance) cues from video streams.
- **Three-dimensional video cameras.** The Microsoft Kinect and other devices track body movements for cues of deception, such as "freezing up" or showing less movement.
- **Microphones.** These vocalic sensors measure the voice, including pitch, quality, and hesitation. Measures are taken to ensure that these data are captured as cleanly as possible, since noisy screening environments may decrease the signal-to-noise ratio. Techniques such as using array microphones and software filtering can help to clean up the vocal signal.

### Sensor fusion

The AVATAR system fuses multiple sensors and cues simultaneously to provide an accurate credibility assessment and transmits the information to an officer.

There are three major methods for fusing multiple sensors: anomaly detection, decision algorithms, and collation and fusion analytics.

- **Anomaly detection.** This measures individual deviations from an established baseline to indicate high levels of arousal that may suggest deception. A given response can be compared to an individual's baseline or to an aggregate baseline. An advantage of using the individual baseline is that people who show a high degree of arousal on innocuous questions will not be flagged when showing the same degree of arousal on sensitive questions.
- **Decision algorithms.** This method uses machine-learning techniques applied to datasets to teach the algorithms to identify deception. Using a

subset of data with ground truth, a decision tree or other algorithm is developed to identify deception in another set. A testing set is used to evaluate the consistency of those cues in a new population. These types of decision algorithms have often performed in the range of 70% to 90% accuracy (Fuller, Biros, & Delen, 2011; Meservy et al., 2005). These algorithms can process massive amounts of data very quickly without incorporating human biases in the decision process.

- **Collation and fusion analytics.** In order to process streams of data in real-time, the AVATAR processes the response to each question separately. The AVATAR then combines the information from each question through multimodal data fusion. Researchers have used low-level classifiers on voice and eye-tracker data and integrated the predictions to achieve much higher accuracy in deception detection (Alipour, Zeng, & Derrick, 2012; Derrick, Elkins, Burgoon, Nunamaker, & Zeng, 2010).

For an overview of the AVATAR interviewing process, see Figure 5.


**Figure 5. AVATAR interviewing process**

# 2. Research Approach

The science and validity of the AVATAR has been investigated by BORDERS and its research partners for more than two decades.[1] A design science approach was used that addresses real-world problems through multidisciplinary and multi-methodological research (Figure 6). Theoretical insights are generated and useful knowledge is gained through an iterative process of prototyping, laboratory experiments, and field tests (Nunamaker, Twyman, & Giboney, 2013). During the AVATAR's development, the underlying technologies were tested in a series of experiments involving a total of some 7,000 participants (see Appendices A and B). This design science process thus produces a final product for end-users that has been rigorously and repeatedly tested in multiple scenarios.

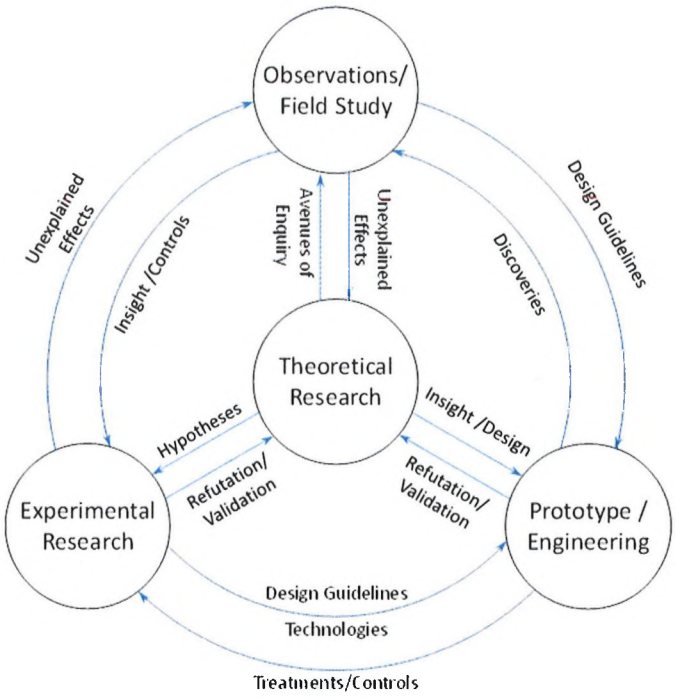


**Figure 6. The Nunamaker model of design science research**

[1] For a complete list of publications presenting the results of investigations about the science of the AVATAR, see: http://borders.arizona.edu/cms/content/journal_papers.

# 3. Research Questions

Our research questions have centered on three primary areas that have the greatest impact on an automated screening system. First the research on intelligent agents will be described, as that is the form the AVATAR interviewer takes. Then the literature pertaining to the potential cultural impacts as they relate to behavior and deception will be reviewed. Finally, and most importantly, the cues of deception used in our research will be described.

## INTELLIGENT AGENT

The AVATAR interviewer is an intelligent agent designed as the "face" of the system. Research on deception shows that the interviewer can have a strong effect on the interviewee, influencing his or her level of comfort and ability to deceive. BORDERS researchers have studied extensively the influence of various attributes of our intelligent agent, including appearance and demeanor.

The appearance of the intelligent agent is important in many ways (Pickard, 2012). A male interviewer was selected after first conducting experiments comparing different aspects of both the face and voice. When comparing male and female faces, people perceived the female face as friendlier, and the male face as more powerful (Nunamaker, Derrick, Elkins, Burgoon, & Patton, 2011). To imbue the system with more authority, the more powerfully perceived male face was selected. The additional effects of various features of the AVATAR, including facial attractiveness, facial similarity to the interviewee, voice quality, and assumed authority (in the form of clothing related to authority figures) were also tested. The AVATAR interviewer in its current form was designed through this research on appearance and embodiment.

Furthermore, nonverbal communication and perceptions as they relate to the demeanor of the agent has been investigated. Since users of computer systems tend to apply human-like social rules and expectations to the system (Nass & Moon, 2000; Reeves & Nass, 1996), users respond to a computer system's synthetic emotions in ways similar to reactions in human-human interactions (Brave, Nass, & Hutchinson, 2005). The AVATAR research has demonstrated that embodied-agent demeanor significantly influences users' perceptions of the agent (Nunamaker, Derrick, Elkins, Burgoon, & Patton, 2011). Embodied agents with a neutral face were perceived as more powerful than smiling agents, and, as mentioned earlier, male agents were judged as more powerful than female agents. Clearly, understanding how the embodiment of the agent can influence users' response is an important part of designing an embodied agent system.

## IMPACTS OF CULTURE

Whether deception is expressed the same way everywhere or varies by culture has been the topic of significant debate. Deception research related to culture has primarily focused on perceived universal indicators of deception, such as emotional displays and the ability to detect deception between and across cultures. The primary cultural theories about deception suggest that the outward cues of deception are caused by phenomena that are universal. Neuro-Cultural Theory (Ekman & Friesen, 1969; Ekman, Sorenson, & Friesen, 1969; Ekman, 1971; Ekman et al., 1987; Ekman, 1994) suggests that all humans are equipped with certain neurological programs that produce universal emotional displays, such as fear and anger. However, other displays are culturally specific depending on social norms. Cultures also have implicit rules for when expressions should be displayed or not and with what consequences. In general, cultural norms and rules have been found to shape physiological and social responses during deception, and to influence what is regarded as deception, motivations to deceive, and the acceptability of deception.

Cultures vary widely on how much they control their verbal and nonverbal expressions (Burgoon, Nunamaker, Metaxas, Levine, & Park, 2009). For instance, African-Americans, Europeans, and Hispanics score higher on expressive control compared to Indians and Central Asians. In a cross cultural study of Middle Easterners and Asians navigating a virtual airport security environment, Wiederhold (2008) found that Asians experienced higher levels of anxiety than Middle Easterners. Additionally, the study found that Iraqis were more relaxed than Iranians when navigating the virtual environment. These findings reveal that responses to stress are not universal.

In summary, the effect of culture on deception is still a topic of study and debate. By incorporating current theories into the design of the AVATAR, the system is more resilient to cultural differences. One primary benefit of such a system is that it is not subject to cultural biases. Also, since the AVATAR is designed to evaluate each person against his or her respective baseline (see "Vocalic measures" in section 5), the confounding factors of culture and other individual differences can be mitigated.

## INDICATORS OF DECEPTION

The ability to identify deception is based on the premise that the nonverbal and verbal behaviors exhibited by truth tellers and deceivers differ. Many of these cues to deception can be categorized as **vocalic** (Sporer & Schwandt, 2007), **kinesic** (Burgoon, Jensen, et al., 2010), and **oculometric** (DePaulo et al., 2003; Hartwig & Bond, 2011) features. Beyond these three primary cues, the AVATAR research has included deception detection using **linguistic**, **cardiorespiratory**, and **thermal** measures. However, at this time the remote sensors for these are not mature enough for operational use.

Some research indicates that humans have limited cognitive capacity, and thus can only process a limited number of tasks simultaneously. What this means, and sup-

ported by prior work (Nunamaker et al., 2011), is that monitoring a large number of heterogeneous cues might exceed the threshold of cognitive ability necessary to monitor and control those cues under most circumstances.

A description of indicators of deception and methods used for their detection are presented below. Deception detection sensors currently being used in the AVATAR system are marked with an asterisk (*).

### Kinesic*

The kinesic indicators of deception are many and varied. Only a few will be mentioned here. Evidence suggests that humans exhibit different degrees of face and body movement when deceiving compared with telling the truth. For instance, people experience a decrease in subtle movements during deception (Vrij, Semin, & Bull, 1996). BORDERS' research has discovered a core rigidity effect of deception (Twyman, Elkins, & Burgoon, 2011; Twyman, Schuetzler, Proudfoot, & Elkins, 2013), possibly a result of a defensive response. A meta-analysis by DePaulo and others (2003) found some support for using lip presses, chin raises, fidgeting, illustrators, facial pleasantness, and overall tension ratings for differentiating truth from deception. Some of these are likely due to arousal, whereas others may have alternative causes.

### Oculometric*

Additional work has been done using novel methods to tie eye movement patterns to deception and concealed knowledge. The protocols employed in this initial work have included a variety of approaches. Twyman and colleagues (2010) tested the use of eye tracking to conduct a concealed information test. Subjects committed a mock crime and later viewed a number of objects located in the room where the crime was committed. The results indicate that how persons look at objects present in a crime scene can be a significant predictor to identify which objects have been used to commit the crime.

Pupil dilation has been shown to be associated with deception in many different contexts (Goldwater, 1972; Nieuwenhuis, De Geus, & Aston-Jones, 2011; Nunnally, Knott, Duchnowski, & Parker, 1967). The dilation of the pupil varies not only with changes in light, but also with cognitive processing (Beatty & Wagoner, 1978), memory load (Kahneman & Beatty, 1966), orienting (Goldwater, 1972; Nieuwenhuis et al., 2011; Nunnally et al., 1967), and attention and effort (Kahneman, 1973). Significant differences in pupil dilation have been found between innocent interviewees and those concealing information (Lubow & Fein, 1996; Proudfoot, Twyman, & Burgoon, 2013). Furthermore, pupil dilation may be a robust measure for rapid screening as it is an autonomic response, thus not easily subject to countermeasures.

Osher (2007) leveraged the use of pupil diameter, eye movement, and response times to test the validity, reliability, and objectivity of using these measures to identify deception. Two experiments were conducted to test the classification accuracies using these oculometric measures. In these experiments participants were instructed to

lie and tell the truth during a computerized personnel-screening interview. In this interview, participants were asked questions that were simultaneously displayed on the screen. By analyzing how the guilty participants looked and reacted to the text of questions they lied to, the researchers were able to achieve up to 85% detection accuracy. Specifically, the researchers found indications that liars were more vigilant (than the truth tellers) when looking at the text of questions they lied too (i.e., liars spent more time fixating and re-reading the questions).

## Vocalic*

The *pitch* of the voice, or the fundamental frequency at which someone is speaking, is controlled by contractions of the larynx in the throat that vibrate the vocal folds. The muscles about the larynx are affected by tension. Thus, stress or nervousness affects the vocal pitch of the speaker's voice (Ekman, O'Sullivan, Friesen, & Scherer, 1991; Liu, 2005). Attributes of a person's pitch that can be used in understanding emotion include the average pitch during an utterance, as well as the variability in pitch over the length of the utterance (Scherer, 1986). The average pitch of a speaker tends to increase when the person is under stress, such as when deceiving, or when highly motivated (Streeter, Krauss, Geller, Olson, & Apple, 1977). Human deception detection based on pitch alone is also no less accurate than detection based on all attributes of speech (Streeter et al., 1977). Interestingly, persons who believe pitch increases when people lie exhibit even larger changes in pitch than those who do not hold such a belief (Villar, Arciuli, & Paterson, 2012).

Other vocal research employing standard acoustic instrumentation has found that liars speak with greater and more varied vocal pitch (Apple, Streeter, & Krauss, 1979; DePaulo et al., 2003; Zuckerman, Rosenthal, & DePaulo, 1982), shorter durations (Rockwell, Buller, & Burgoon, 1997; Vrij et al., 2008), less fluency, and longer response latencies (silences before a speaking turn) (DePaulo, Stone, & Lassiter, 1985; deTurck & Miller, 2006; Rockwell et al., 1997; Sporer & Schwandt, 2006).

One tool that measures multiple vocal features is layered voice analysis (LVA). Initial tests by BORDERS using custom LVA software produced a classification accuracy rate of 63% (Burgoon, Nunamaker, & Metaxas, 2010). In follow-up studies, when vocalic features were combined with cardiorespiratory features (see section below), overall detection accuracy increased to 72% (Burgoon, Nunamaker, et al., 2010).

## Linguistic

When people are deceptive, they use different words and grammar than when they tell the truth. Response detail, plausibility, logical structure, discrepancies, involvement, immediacy, and repetitions, as well as spontaneous corrections, admissions of lack of memory, and related external associations may be useful indicators of deception (DePaulo et al., 2003). The reliability of these cues is still being debated and the effects of some cues may be context-dependent. Using linguistic cues—such as non-immediacy terms, suppressing answers, feelings, and cognitive load--BORDERS re-

searchers were able to correctly classify truthful emergency telephone callers 88% of the time, and deceptive callers 92% of the time (Burns, Moffitt, Jenkins, & Nunamaker Jr, 2011). BORDERS researchers are currently investigating ways to help the AVATAR understand the context of the interviews and to change its queries based on interviewee answers.

### Cardiorespiratory

Laser Doppler vibrometers (LDV) are devices that capture pulse, blood pressure, and respiration through a medically safe laser aimed at the interviewee's neck. In this way, some of the traditional measures captured by the polygraph can be measured remotely and non-invasively. Initial investigations of the LDV showed promise, but limitations in the hardware make it currently infeasible to include in the AVATAR (Derrick et al., 2010). For example, the laser must be manually aimed at the interviewee's neck; some collars would obscure the interviewee's neck to laser measurement; and for safety reasons, the laser would need to shut off if the interviewee moved too much. There continue to be improvements in remote cardiorespiratory measurement, including recent methods using video cameras (e.g., measuring subtle color changes in the face) that are currently being researched for inclusion in in future versions of the AVATAR.

### Thermal

Thermal sensors capture infrared signals to monitor facial blood flows. In a deception context, most researchers focus on the orbital areas of the face associated with fight-or-flight responses (around the nose and eyes) (Pavlidis, Eberhardt, & Levine, 2002). Throughout the interview, changes in blood flow are monitored to detect differences in arousal to questions. More research is needed to determine the reliability of thermal signals (Burgoon et al., 2008).

## AVATAR'S INTEGRATED SCREENING PROCESS

Figure 7 depicts the processes involved within the AVATAR to screen passengers and provide decision support to border guards. In this graphic, a passenger or interviewee receives messages (border screening questions) from the intelligent agent (AVATAR face and voice). While the interviewee responds, nonverbal and verbal behavior and physiological signals are measured using the sensors integrated in the kiosk. These measurements are merged, processed, and fused before being analyzed to determine interviewee risk and credibility. The risk information is then transmitted as decision support to the border guard and to the intelligent agent to guide its credibility assessment and screening strategies. This process continues until either the passenger is cleared or an intervention is requested by a monitoring border guard.

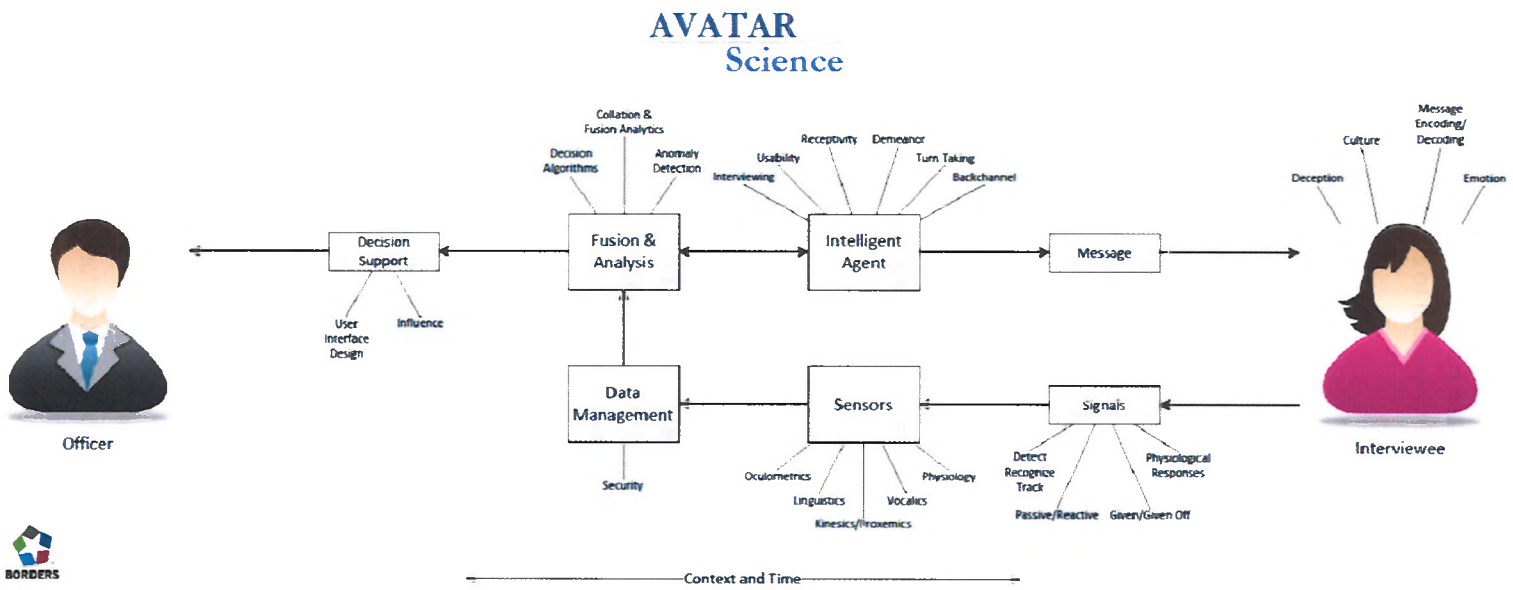


**Figure 7. Depiction of the processes involved in automated credibility assessment**

# 4. BORDERS–Frontex Workshops

BORDERS has established joint ventures with international institutions engaged in border security and immigration research with the goal of building collaborative platforms for future investigations. In 2009, BORDERS began collaboration with Frontex, the European Union (EU) agency based in Warsaw tasked to coordinate the operational cooperation between member states in the field of border security.

Since 2010, BORDERS has contributed to a series of Frontex-sponsored workshops, experiments, and field tests for EU border guards. The purpose of these events was to introduce border guards to the AVATAR technology for deception detection and explore its potential utility for border management. Each event has built upon the outcomes of the previous one, and included participants from EU member states.

In this section, each of the events is summarized.

## WORKSHOP 1

### Artificial Intelligence for Screening and Decision Support at Border Crossings, Warsaw, Poland, 2010

In 2010, BORDERS and Frontex hosted their first joint workshop in Warsaw, Poland. The purpose was two-fold: (1) to help BORDERS researchers gain a better understanding of the issues and challenges facing border guards, and (2) to demonstrate the AVATAR system through a hands-on scenario constructed to demonstrate its capabilities in a "fake bomb" experiment and screening assessment.

#### "Fake bomb" experiment

In this experiment, participants were randomly assigned to one of two groups. The first group received a lead pipe with other components and told to assemble a fake improvised explosive device (IED) from printed instructions (Figure 8). After completion, they were instructed to pack the fake IED in a travel bag and to attempt to pass through a security screening that included being interviewed by the AVATAR system. The second group packed a travel bag only and went through the same screening process. During the automated interview, the AVATAR sensors measured changes in voice quality, eye gaze patterns, and pupil dilation for all persons in each group.

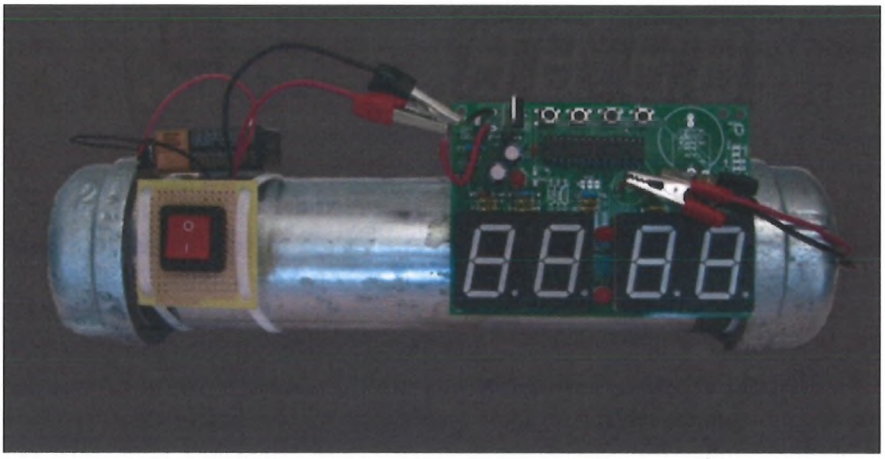

**Figure 8. Completed fake improvised explosive device (IED)**

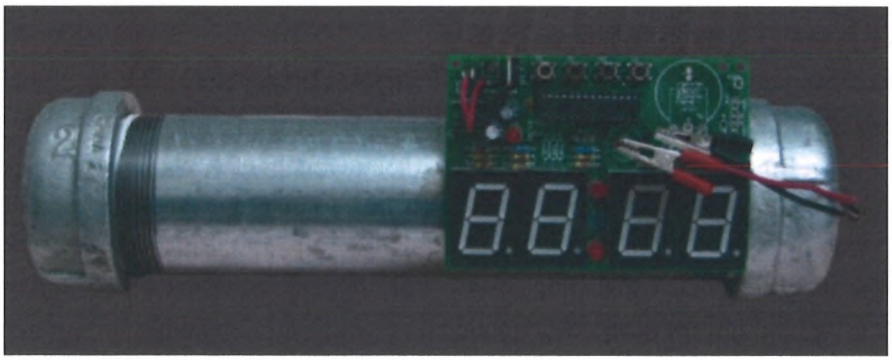

**Figure 9. Fake IED with missing switch**

During the automated interview, participants were shown three images showing, respectively, a keyboard, a sports car, and the fake IED (but without the switch attached; Figure 9). For each image, the AVATAR's eye-tracker sensor measured interviewees' gaze patterns (i.e., how long a person spent looking at a region of the image) and pupil dilation. The bomb makers ("guilty group") looked longer at the left side of the image (M=28.52%, SD=13.67%) where the detonator should have been, as they knew something was missing (Figure 10). The "innocent" group spent less time looking at the left side (M = 12.43%, SD = 6.46%) and focused on the more interesting right side containing the timer and electronics.

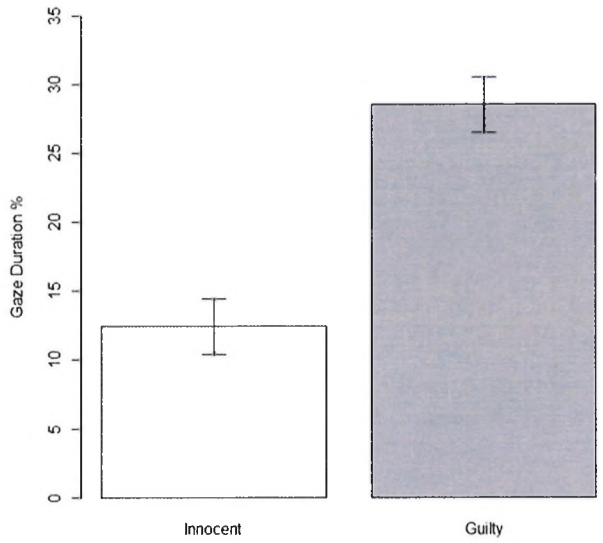


**Figure 10. Percentage of time spent fixated on the modified switch section of the bomb image for the two groups of interviewees**

The difference in an individual's pupil dilation—as measured while viewing the sports car image versus the image of the modified bomb—was also compared. It was expected that the guilty group would have a higher level of dilation when seeing the bomb due to arousal and familiarity (Figure 11). While not as distinct as the gaze differences, there were still significant differences between the two groups of interviewees (Guilty M=0.25, SD=0.21; Innocent M = 0.12mm, SD = -0.19mm).

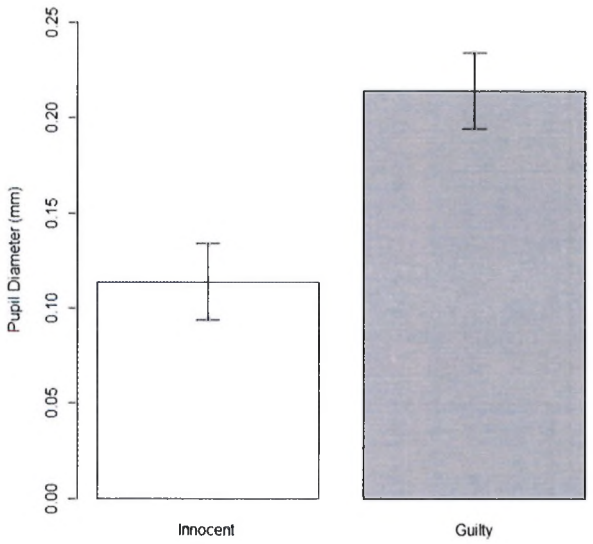


**Figure 11. Pupil dilation differences (in mm) as measured while viewing the sports car image versus the bomb image**

In addition to the eye gaze and pupil dilation examined above, the participants' voices were analyzed during the interaction. During the AVATAR interview, a stereophonic microphone was used to collect participants' voice samples in response to the question: "Are you carrying any prohibited items through this checkpoint today?" The guilty group had significantly greater fluctuations in vocal pitch when answering this question, indicating increased uncertainty or heightened cognitive load:

- Mean standard deviation for bomb maker group = 199 Hz
- Mean standard deviation for innocent group = 159 Hz

In terms of detection rate, the AVATAR had an overall accuracy rate of 87% in detecting bomb makers, with only four false positives and no false negatives.

Being able to use an image of the bomb is an overly simplistic representation of a real-world scenario, but the results of the experiment provided support for the feasibility of the use of eye gaze to detect deception or imposters. Further experiments are required to address more realistic interviewing contexts.

## "Improving judgment" experiment

After participating in the fake bomb study, workshop participants than took part in a second experiment investigating how human judgment can be improved by the AVATAR. In this experiment, participants viewed 20 videos of liars and truth tellers. The participants rated how truthful they found the individuals presented in each vid-

eo. For half of the videos, participants were provided with feedback (representative of the AVATAR's output) on the behaviors of the speakers. From this experiment, the performance of participants when detecting deception unassisted and assisted by the AVATAR were compared. Unaided by technology, participants judged truth and deception at chance levels (Overall = 54%, Deception detection = 48.9%, Truth detection = 48.9%). When aided by technology, participants improved their detection rate (Overall = 60%, Deception detection = 55.9%, Truth detection = 66.1%).

### Key workshop findings

- Based on a fake bomb experiment, the AVATAR identified 100% of the "guilty" participants with an overall accuracy of 87% (with false positives).
- In a second experiment, border guards were 7% more accurate in detecting deception when aided by technology providing behavioral analysis.
- While some workshop participants expressed skepticism about using the AVATAR to aid border guards, others expressed great interest in these technologies and in their future promise for improving the screening process.
- Important issues that participants identified as needing further study included incorporating culture and language considerations in the AVATAR, and the relative utility of these technologies in second-line vs. front-line screening.
- While terrorism is the major concern for the U.S. DHS, the EU's focus is on border control issues related to immigration, including document fraud.
- The five greatest threats to the EU according to participants are illegal immigration, false documents, organized crime, human trafficking, and smuggling.

## WORKSHOP 2

## Improving Border Checks with Next-Generation Artificial Intelligence and Advanced Sensor Technology: Decision Support for Assessment, Screening, and Interviewing, Warsaw, Poland, 2011

In the 2010 workshop, document fraud was identified as a major challenge for the border guards. For the 2011 workshop, a "document fraud" experiment was designed to test the AVATAR's ability to identify holders of false documents by analyzing their behavior and verbal responses during a simulated border check.

### "Document fraud" experiment

To conduct the "document fraud" experiment, some workshop participants were assigned the task of going through passport control using their own travel documents (innocent group), while participants in another group were instructed to use "falsified documents" to go through passport control (guilty group). During the AVATAR screening, each person's scanned document appeared on the screen and the person was

asked if the information was correct. Simultaneously, the sensors monitored the person's eye gaze behavior and vocal cues.

As part of the experiment and to investigate the ability of the AVATAR to be more accessible to a wide variety of cultures, the interviewee could choose to have the automated interview conducted in one of five languages: English, Spanish, French, German, and Polish. Interestingly, despite the option for most participants to choose to be interviewed in their native language, nearly all persons choose to have the interview conducted in English.

All workshop participants were given a Frontex "visa" document at registration to use as their travel document during the simulated border check, in addition to their original national passport. If participants were assigned to the innocent group, they waited in line for the border check and answered all questions posed by the AVATAR honestly. Participants in the guilty group were given instructions (and falsified documents) to impersonate someone else and to attempt to pass the border check to commit a terrorist act. Guilty participants were offered a reward if they remained undetected. The results of this study are summarized below and described in more detail in Elkins, Derrick, and Gariup (2012).

To build on the previous year's research into how to improve border guard decision making, some participants served in the role as human screener during the "document fraud" experiment. Some of the screeners were provided wireless tablet computers containing the results of the AVATAR's behavioral analysis of each interview. Figure 12 illustrates the output provided to screeners to aide their evaluation. There were also screeners that were not provided any AVATAR support. All screeners (who were blind to the condition of the participants, i.e., unaware of each participant's actual personal information) indicated which passengers they thought were suspicious or were lying.

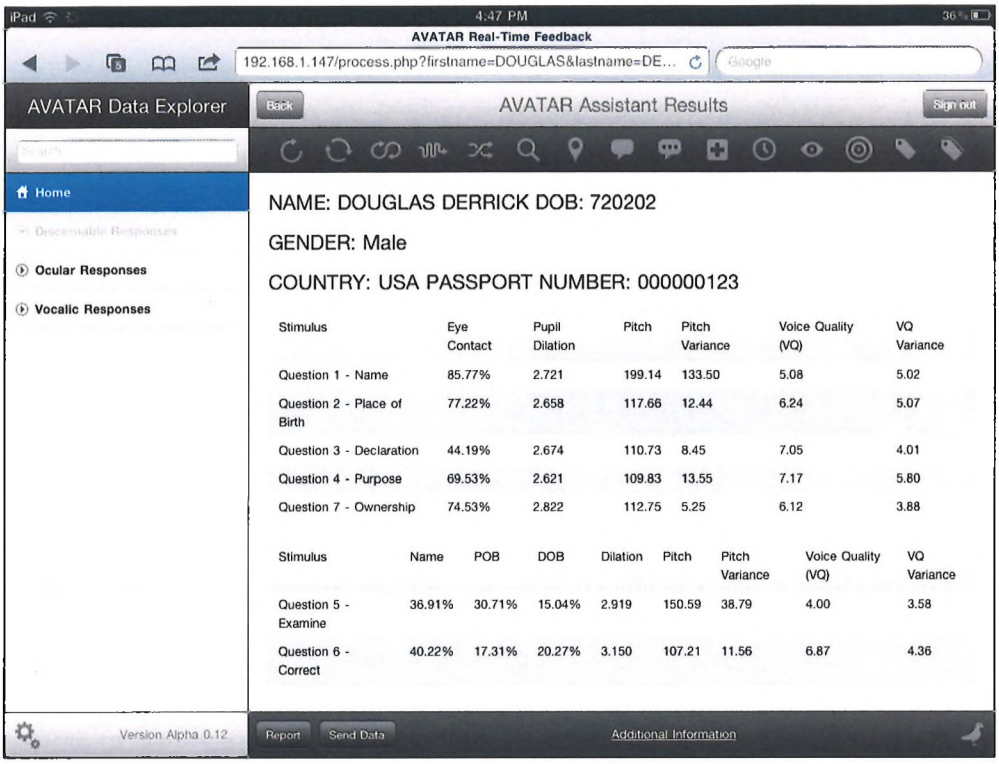


**Figure 12. AVATAR behavior analysis provided to screeners**

## Vocalic measures

An individual baseline was created for each participant using his or her response to a control question ("What is your name?") compared to the response to a "charged" question ("Is the information in this document correct?"). The results shown in Figure 13 indicate that guilty participants showed a significantly greater drop in vocal quality than innocent participants. This marked change is very likely due to the stress involved in maintaining the lie.

## Eye behavior

The differences in fixation on the date of birth (DOB) field, which was modified on the fraudulent documents, were compared. The results showed that imposters spent nearly twice as much time looking at the DOB than truth tellers. The results are shown in Figure 14.

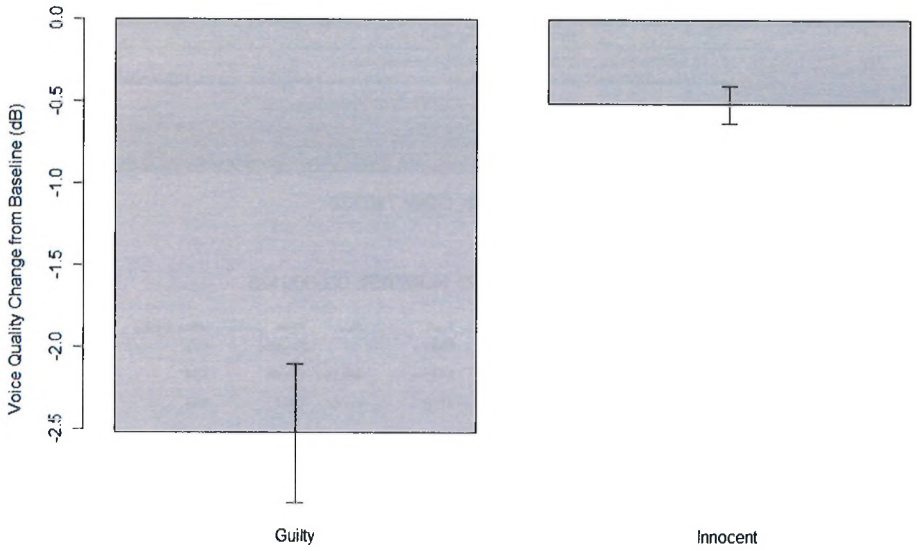


**Figure 13. Average change in voice quality for guilty and innocent groups for responses to "baseline" versus "charged" questions**

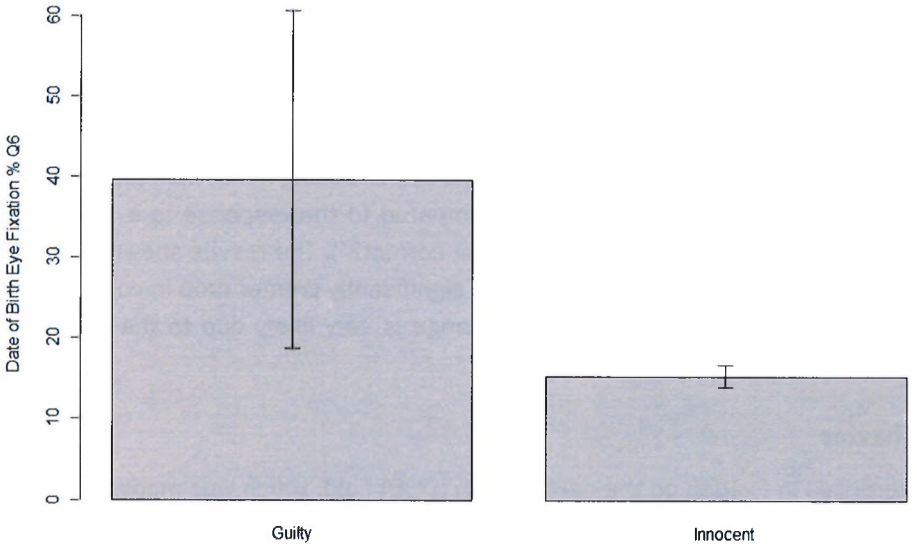


**Figure 14. Percentage of time spent looking at DOB region of identify documents for guilty and innocent groups**

## Combined

A merged classification model was created to measure the AVATAR's accuracy in detecting document fraud, as shown in Figure 15. An overall accuracy rate of 94.7%

was achieved. All guilty participants were successfully identified and three innocent participants were incorrectly classified as guilty, resulting in a false positive rate of 5.9%. (Note: "Eye Vote" in Figure 15 refers to how the classification model decided or "voted" on guilt or innocence based on eye behavior.)

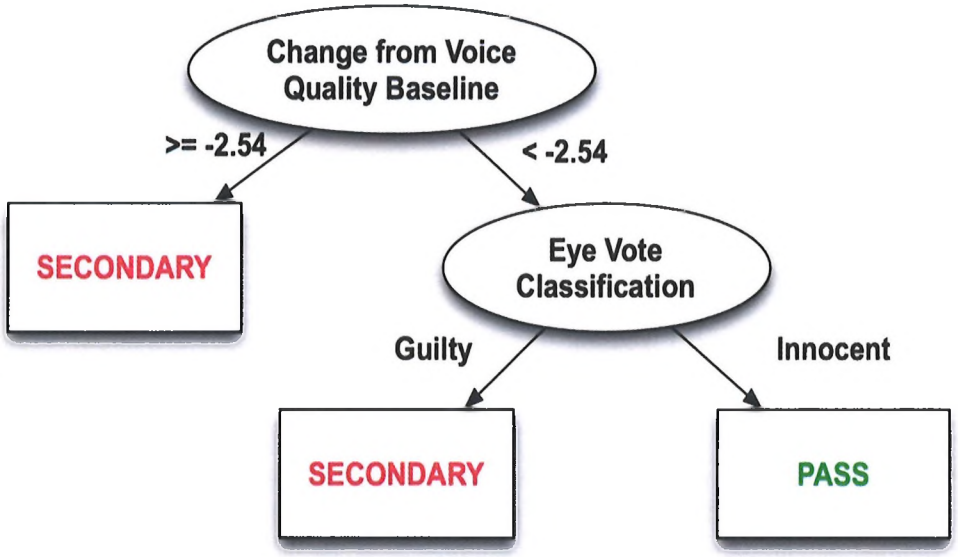


**Figure 15. Classification model for imposter experiment**

### "Improving screening" experiment

After participating in the document fraud experiment, all participants took part in a second experiment evaluating how they make decisions when supported by technology such as the AVATAR. The detailed results of this study were published in Elkins, Dunbar, Adame, and Nunamaker (2013). A key finding of this experiment was that advanced decision making technology such as the AVATAR has the potential to threaten border guards by challenging their expertise. When technology such as the AVATAR is implemented, it must be done so as a team member or collaborator, rather than as a means to simply provide guilty or innocent recommendations.

## Key workshop findings

- In the document fraud experiment, the AVATAR achieved an overall accuracy rate of 94.7% using fused vocalic and eye behavior sensors.
- Participants had mixed reactions to the assistance provided from the automated analysis of behavior. Some observed that there was a lot of information to process, and the ease of digesting the information would depend on the interface used to deliver it. That said, the scores indicated improvement as a result of exposure to more targeted information.