



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Human Inspired Technology Research Centre

PHD COURSE IN BRAIN, MIND AND COMPUTER SCIENCE

XXX CYCLE

LIE DETECTION IN THE FUTURE:

THE ONLINE LIE DETECTION VIA HUMAN-COMPUTER INTERACTION

Director: Ch.mo Prof. Giuseppe Sartori

Supervisor: Ch.mo Prof. Giuseppe Sartori

Co-Supervisor: Ch.mo Prof. Mauro Conti

PhD Candidate: Merylin Monaro

Acknowledgements

First, I am immensely grateful to my supervisor, prof. Giuseppe Sartori, for sharing his pearls of wisdom with me during the course of these three years of research. Thanks for giving me this opportunity and for following my training and progress with passion and interest.

I thank my computer science colleagues, who provided insight and expertise that greatly assisted the research. In particular, Riccardo Spolaor and Qian Qian Li, who collaborated to the development of some experiments, Mirko Polato and Nicolò Navarin for their assistance in machine learning issues.

I would also like to show my gratitude to all the bachelor's and master's students who gave a contribution to the collection of data, which are included in this manuscript (Angelica Bollini, Marta Businaro, Francesca Fugazza, Chiara Galante, Alessandra Guiotto, Sara Marcon, Ilaria Zampieri, Francesca Zecchinato).

Moreover, I thank the reviewers for their comments that, I am sure, will greatly improve this manuscript.

I thank Andrea Zangrossi for his important backup and all people of Sartori's group that contributed to make enjoyable the work environment.

Then, I'd like to thank my co-supervisor, prof. Mauro Conti, and prof. Luciano Gamberini for their valuable support, as well as all the people that have worked to improve my experience in BMCS PhD course.

Finally, the most special thanks are due to my family and friends that always supported my choices and encouraged me to get on my way.

Abstract

Half the people in the Planet Earth are now on internet, surfing the web, keeping connection with the outside world, using online services and interacting in social networks. However, the spread of internet is going hand in hand with the growing malicious use of it. Creating fake social network profiles, wide spreading fake news, posting fake reviews, identity theft to perpetuate online financial frauds are only few examples. To face these problems, all the big internet companies, like Google and Facebook, are now taking the direction towards the online lie detection research. The present work is a contribution to online deception detection through the study of computer-user interaction. After a brief review of the current lie detection methods, focusing on their advantages and disadvantages for online application, a series of proof of concept experiments are reported. Experiments were conducted measuring indices deriving from three different tools of human-computer interaction: reaction times on keyboard, keystroke dynamics and mouse dynamics. Two strategies were used to increase liars' cognitive load and facilitate the observation of distinctive features of deception: unexpected questions and complex questions. Experiments focused on the deception about identity, as it is a very hot issue and represents a current challenge for companies that provide online services. Participants were asked to respond lying or truth telling to questions that appeared on the computer screen, typing the response, clicking on it with the mouse or pressing one of two alternative keys on keyboard. Data collected from liars and truth-tellers' responses were analyzed and used to train machine learning classification models. Classification accuracies in distinguishing liars from truth-tellers ranged from 80% to 95%, depending on the deceptive task. Results have proved that it is possible to spot liars analyzing their interaction with the computer during the act of lie. In particular, we demonstrated that keystroke dynamics is a very promising tool for covert lie detection and it is easily integrable with the online existing applications. Moreover, we confirmed that the cognitive complexity of the deceptive task increases the possibility to detect deception.

Keywords

Online deception, lie detection, human-computer interaction, mouse tracking, keystroke dynamics

Contents

Acknowledgements	i
Abstract	ii
Keywords	ii
Chapter 1 Introduction	9
1.1 Fake Identities	9
1.1.1 Fake profiles	10
1.1.2 Fake chatting.....	11
1.1.3 Fake VISA	12
1.1.4 Online banking and other financial services.....	12
1.2 Fake News	13
1.3 Fake Reviews	14
Chapter 2 Related works	17
2.1 Linguistic Approach	18
2.2 Behavioral Approach	20
2.2.1 Mouse tracking.....	22
2.2.2 Keystroke dynamics.....	24
2.3 Strategies to Increase Cognitive Load	25
2.3.1 Unexpected questions.....	26
2.3.2 Complex questions	27
Chapter 3 Materials and Methods	29
3.1 Participants	29
3.2 Experimental Design and Procedures	29
3.3 Data Collection	29
3.3.1 Reaction times	30
3.3.2 Mouse tracking.....	31
3.3.3 Keystroke dynamics.....	33
3.4 Data Analysis	35
3.4.1 Feature selection	35
3.4.2 Descriptive and statistical analysis	35
3.4.3 Machine learning models	36
Chapter 4 Experiments	39
4.1 The Detection of Faked Identity with Unexpected Questions and Mouse Dynamics	39
4.1.1 Participants.....	39

4.1.2	Experimental procedure.....	39
4.1.3	Stimuli.....	40
4.1.4	Collected measures	41
4.1.5	Analysis of trajectories.....	42
4.1.6	Feature selection	44
4.1.7	Descriptive statistics	44
4.1.8	Machine learning models	46
4.1.9	Can we detect liars also when they respond truthfully?.....	49
4.1.10	Generalization to different cultures.....	51
4.1.11	The resistance to countermeasures.....	52
4.1.12	The use of changing labels.....	55
4.1.13	The use of negative questions.....	58
4.1.14	Discussion.....	61
4.2	The Detection of Faked Identity with Unexpected Questions and Choice Reaction Time.....	63
4.2.1	Participants.....	63
4.2.2	Experimental procedure.....	63
4.2.3	Stimuli.....	63
4.2.4	Collected measures	63
4.2.5	Features selection.....	64
4.2.6	Descriptive statistics	65
4.2.7	Machine learning models	65
4.2.8	Discussion.....	66
4.3	The Detection of Faked Identity with Unexpected Questions and Keystroke Dynamics	67
4.3.1	Participants.....	67
4.3.2	Experimental procedure.....	67
4.3.3	Stimuli.....	67
4.3.4	Collected measures	68
4.3.5	Feature selection	68
4.3.6	Descriptive statistics	69
4.3.7	Machine learning models	70
4.3.8	Can we detect liars also when they respond truthfully?.....	71
4.3.9	Analysis on normalized predictors.....	72
4.3.10	Countermeasures and alternative efficient models	73
4.3.11	Classification of liars using only data from truth-tellers	73
4.3.12	Application of the paradigm to online form.....	74
4.3.13	Discussion.....	76
4.4	The Detection of Faked Identity with Complex Questions and Choice Reaction Time	77
4.4.1	Participants.....	77

4.4.2	Experimental procedure.....	77
4.4.3	Stimuli.....	77
4.4.4	Collected measures	78
4.4.5	Descriptive statistics	79
4.4.6	Feature selection	80
4.4.7	Machine learning models	81
4.4.8	Analysis on normalized predictors.....	81
4.4.9	Analysis by stimuli.....	82
4.4.10	Discussion.....	83
4.5	The Detection of Faked Identity with Complex Questions and Mouse Dynamics	85
4.5.1	Participants.....	85
4.5.2	Experimental procedure.....	85
4.5.3	Stimuli.....	85
4.5.4	Collected measures	85
4.5.5	Analysis of trajectories.....	85
4.5.6	Feature selection	87
4.5.7	Descriptive statistics	87
4.5.8	Machine learning models	88
4.5.9	Countermeasures and alternative efficient models	88
4.5.10	Discussion.....	89
4.6	The Detection of False Autobiographical Events with Complex Questions and Mouse Dynamics	91
4.6.1	Participants.....	91
4.6.2	Experimental procedure.....	91
4.6.3	Stimuli.....	94
4.6.4	Collected measures	95
4.6.5	Analysis of trajectories.....	96
4.6.6	Feature selection	97
4.6.7	Descriptive statistics	98
4.6.8	Machine learning models	100
4.6.9	Discussion.....	100
4.7	Can mouse dynamics and complex questions generalize from a topic to another?	102
4.7.1	Detection of liars about autobiographical events using a model about identity.....	103
4.7.2	Detection of liars about identity using a model about autobiographical events.....	103
4.7.3	Discussion.....	104
Chapter 5	Conclusion.....	105
5.1	Main Results	105
5.2	General Conclusions	106
5.3	Covert Lie Detection	107

Contents

5.4	Limitations	108
5.5	Future Directions	109
	References	111
	Annex 1	125
	Annex 2	127
	Annex 3	131
	Annex 4	134
	Annex 5	137



**"Don't believe
everything you
read on the
internet just
because there's
a picture with a
quote next to it."**

-Abraham Lincoln

Chapter 1 Introduction

The International Telecommunications Union reports that on June 2017 the 51% of the world's population is on internet [1] that corresponds to 3,885,567,619 of users [2]. From 2010 to 2017, the global internet usage has grown by 976.4% and, at the same time, the number of users in social networks has exponentially increased. In June 2017, Facebook counted 2.01 billion monthly active users, whereas in 2010 they just achieved the threshold of 500 million of active people [3]. Five new Facebook profiles are created each second and every 60 seconds 510,000 comments are posted, 293,000 statuses are updated and 136,000 photos are uploaded [4]. A similar increment has been reported also for other socials, such as Twitter, Instagram, Snapchat, YouTube, Pinterest [5], [6].

But are all these activities related to real people and real experiences? The answer is no.

In 2012, in a report for the United States Securities and Exchange Commission, Facebook declared that about the 8.7% of the worldwide active users are actually duplicate or false accounts [7], [8]. During the last year Facebook, as well as Google, have intensified their actions against the fake online news [9]. To guarantee the economic interests of their millions of customers, companies like Amazon, TripAdvisor and Yelp are facing the problem of fake reviews [10]. Moreover, for a large number of others online services, such as e-commerce and online banking, the verification of the truthfulness of the information provided by the users is currently a key issue [11].

In short, the problem of online deception and its detection seems to be real and all the big companies that provide online services are moving in this direction. In the next paragraphs, a review of the most relevant issue about online deception is presented alongside their consequences for society and economy.

1.1 Fake Identities

In the current historical and cultural framework, the identity verification is an increasingly urgent problem [12]. Faked identities are used for a wide range of criminal purposes, both in the real world and in the internet environment. Oversimplifying, a faked identity may be defined as an alteration of demographic information in legal documents. For example, terrorists use false passports to enter European and US borders [13], [14], making this a big deal for physical security.

Concerning online security, the scenario of the fake identities becomes more intricate. Three main specific issues can be identified: the identity alteration, the identity theft and the identity fraud. The first consists in the modification of one or more user's demographic information, which is really common in social networks profiles [15] and it is not clearly legally regulated. For example, a man who declared to be a woman, a boy who claims to be younger than he actually is or an underage girl who says to be adult. Identity alteration can be done for different reasons, but in most cases it has

social purposes, such as attracting the attention [16] or the sympathy of the other people (e.g., grooming) [17] or avoiding the direct exposure to shameful situations [18]. To support the creation of fake accounts there are some online services, which automatically generate completely fake profiles, including address, employment, tracking numbers and financial information [19]. Completely invented accounts, especially fake email accounts, are also commonly used for economic aims and scams, such as spamming and phishing.

Conversely, the identity theft consists in the deliberate use of someone else's identity [20] and in most States it is punished with compensation or prison [21]. Motivations are usually social (e.g., pranks, cyberbullying) or related to business (e.g., to increase the visibility of posts and fan pages, manipulate votes or the number of visits, junk emails or advertisement inducing customers to pay for clicks) [22].

Finally, identity fraud is a crime where one person uses another person's personal data, without authorization, to deceive or defraud someone else [23]. Identity fraud can occur with or without identity theft, such as in the case where the fraudster has been given someone's identity information for other reasons but uses it to commit fraud. Motivations for online identity frauds are mainly financial crimes, such as open a credit card account without permission or stolen online banking credentials.

In the following, some statistics about the spread of the mentioned above phenomena are reported, describing their social and economic impact and quoting examples.

1.1.1 Fake profiles

As previously anticipated, in 2012 Mark Zuckerberg reported that around the 9% of Facebook accounts are fake or duplicate [7], that means about 83 million profiles [8]. Similar percentages were reported for Instagram, with the 10% of false profiles, and Twitter that counts approximately the 8.5% of fake accounts [24].

There are a mixture of innocent and malicious reasons for fake profiles, including professionals doing testing and research, and people who want to segment their Facebook use more than is possible with one account. Facebook has classified the fake accounts into three categories: duplicate accounts, misclassified accounts and undesirable accounts [8]. In 2012, duplicate profiles covered up to 4.8% (45.8 million) of Facebook's total active members. According to the social network's terms of service, users are not allowed to have more than one Facebook personal account or make profiles on behalf of other people (e.g., parents creating Facebook accounts for their young kids). Misclassified accounts are personal profiles that have been made for companies, groups or pets. Those types of profiles (22.9 million in 2012) are allowed on Facebook, but they need to be created as pages. For this reason, the company provides the opportunity to convert these accounts into approved pages without losing information. In 2012, Facebook estimated that 2.4% of its active accounts belonged to non-humans. Some pets, such as Boo, the "world's cutest dog", are typically categorized as public figures. The third category, undesirable accounts, includes fewer (just the 1.5% of all active accounts), but it is the most troublesome. On 2012, there were 14.3 million of undesirable accounts that Facebook believes have been created to violate the companies terms, like spamming.

Between malicious motivations of faked profiles, one of the most dangerous is the child grooming. A study conducted in Germany in 2011 on a sample of 518 students, aged between 10 and 16, reveals

that the 21.4% of participants, or rather two out of ten adolescents, had been cyber-grooming victims over the last year [25]. Most of the cyber-grooming victims report negative consequences, specially psychological, such as embarrassment, depression or self-harm [26]. They show difficulties in establishing new relationships, loss of trust in other people, afraid for ridicule and feeling of helplessness. Finally, they develop a sense of insecurity, given that the abuse has been perpetrated, through the network, within the home.

1.1.2 Fake chatting

The use of faked profiles to catch the attention of the other people is also spread amongst adults searching for other adults. This issue has been the focus of a very famous criminal case known as “tallhotblond” [27]. Tallhotblond is the story of Thomas Montgomery, a 46 years old man, and Mary Shieler a 45 years old woman, who met in a chatroom. He presented himself as a young 18 years old marine, whereas she stole the identity of her eighteen daughter. They had a chat-based love story for two years, until Mary discovered the real identity of Thomas and decided to break up with him and hang out another man. Montgomery became jealous and, finally, he killed the new lover of Mary. After the crime, he realized that the woman for whom he murdered did not really exist. This is only an example to demonstrate that what happens in the web, including relationships, may not correspond to the reality, and could have dangerous consequences.

Fake chatting refers to the phenomenon of have a chat conversation pretending to be a person with demographic characteristics different from the actual ones. Generally, it happens using a false chat account. But why people chat under false account? John Suler, professor of psychology at Rider University, speaks of “online disinhibition effect” [18]. In the virtual environment, we do not interface directly with other people but with accounts, that are usually usernames, sometimes associated with photos. The psychological distance, due the physical lack of the interlocutor, makes the user does not feel the restraints that he experiences when communicating in person. In the web-chat space, people say and do things that they would not say and would not do in face-to-face contexts. They relax and express themselves more openly. Anonymity, asynchronous communication, and empathy deficit contribute to online disinhibition. This disinhibition can be positive when it leads the user to display unusual kindness and generosity, whereas it become dangerous when it causes violent behaviors, such as foul language, demonstrations of hate, anger and threats.

The online disinhibition effect facilitates the phenomenon of sexting as well [28]. The term is a portmanteau of “sex” and “texting” and it consists in sending, receiving, or forwarding sexually explicit messages, photographs or images, via any digital device. A survey on 1,496 adolescents between 12 and 18 has shown that about the 10% of them have received sexual messages or videos via mobile phone, while the 6.7% sent these kind of messages to friends and adults, including strangers [29]. Two more recent investigation, which have been conducted by an Italian organization for the children’s rights and the American Academy of Pediatrics, reported that the 22% of the adolescent had sexting, also with strangers [30], in the last six months [31]. Moreover, it emerged that there is a positive correlation between sexting practice and risky sexual behaviors in real life, with a higher level of risk for those who shared photos in addition to text messages. Concerning photos, one of the most popular social networks based on photo sharing is Snapchat. The principal concepts of Snapchat is that pictures and messages are only available for a short time before they become inaccessible [32].

As consequence, users become more involved in exchanging images of all kinds, including sexual photos. However, a screenshot is enough to keep the image from the recipient and share it with other users. In fact, in Instagram there is the “#snapchat” hashtag where all the images taken by Snapchat are posted unbeknownst to the victim [33]. These images are also sold, fueling the pornography and child pornography market. To partially solve this problem, companies such as Instagram and Twitter are developing specific algorithms to ban inappropriate content (e.g., photos containing female nipples) [34].

1.1.3 Fake VISA

Consequences related to the use of false identities that are different from those described above concern VISA and passports. Numerous countries, including USA, provide travel VISA through electronic systems, such as the Electronic System for Travel Authorization (ESTA) [35]. International travelers are asked to compile an online form entering their biographical information and, after some checks of the databases, the access to the country is approved or denied. However, most of the criminals, including people working for terrorist organizations, are unknown and their information are not included in the databases [36].

Faked personal identity is a major issue in security in Europe as well [12]. Especially in the last years, a large number of migrants from the Middle East are entering Europe without documents, and sometimes is enough to self-declare the identity information to obtain a European passport [37]. Among them, there are often Islamist militants involved in terror attacks, as occurred recently in Paris, Brussels and Berlin [38], [39], [40], [41]. For example, the perpetrator of the Berlin attack on December 19, 2016 was an undocumented immigrant from Tunisia who entered Germany from Greece and then Italy using several false identities [39]. One of the terrorists involved in the Brussels airport suicide bombing on March 22, 2016 was using the identity of a former Inter Milan football player [13] to travel around Europe. The security measures adopted by the border patrol include mainly biometric passports and the cross-checks of information contained in databases (e.g., wanted lists, fingerprints). But in cases as those mentioned above, biometric identification tools (e.g., fingerprints) could not be applied as most of the suspects were previously unknown. So, despite the attempts by governments to increase security measures [42], the issue remains open.

1.1.4 Online banking and other financial services

Finally, the identity fraud is an urgent problem for financial online services, such as online banking and e-commerce. Even if online banking is an easy way to access bank services, it is enough to steal the credentials of a user to access his bank account.

As anticipated above, identity fraud is defined as the use of another person’s identity without authorization to deceive or defraud someone else (e.g., the bank) [23]. An example is the use another person’s information to open a credit card account without permission, and then charge merchandise to that account. Identity fraud does not occur when a credit card is simply stolen; that may be consumer fraud, but is not identity fraud. Conversely, identity fraud is a synonym of unlawful identity change. It refers to the specific crime activity of use the identity of another person or of a non-existing person fraudulently. In fact, identity fraud does not necessarily involve colluding or theft of another’s person information; it can involve the use of a fake name as well. The false identity is generally created

combining faked and actual information (e.g., a real social security number along with a fake address) [43]. The fraudster can then use the fake identity, to create a credit card or other accounts with financial purposes. The generation of false financial accounts is often very simple, as sometimes it is enough to scan the documents and send them to the bank. The scan documents may be easily falsified.

It is estimated that the economic losses due to this criminal activity in US are around billions each year [44]. Equifax reported that about the 80% of all credit card frauds are due to identity fraud accounts [45]. One famous case of identity fraud affected the PlayStation [46]. In 2011, the PlayStation Network was hacked into, and the information from everyone who had his credit card coordinates installed on the Network were stolen. The Sony needed three months to fix the problem. With approximately 77 million of compromised accounts, it was one of the largest data security breaches in history and cost hundreds of thousands of dollars to the Sony [47].

However, the losses are not just financial. Oftentimes, falsified or forged documents are used to purchase tobacco or alcohol as a minor, to acquire driver's licenses, as well as to continue playing on a certain sports team or organization when that person is really too old to compete [48]. Criminals can use the social security numbers of children to apply for government benefits, apply for loans or utility services, or rent a place to live. Children are targets of identity theft because the crime can go undetected for years, often until the child applies for his or her first credit card or mobile phone account. This phenomenon is called "runway" and it is estimated that between 140,000 and 400,000 children become victims of identity theft every year [49].

1.2 Fake News

Fake news are defined as articles drawn up with invented, misleading or distorted information, made public with the deliberate attempt to misinform or spread jokes through traditional media or via the internet, through the social media [50]. Fake news are often published with the intention of attracting the reader and finally to obtain financial or political gains. A separate global study published by Edelman in 2016 found that for news and information people trusted the search engines (63%) more than the traditional media, such as newspapers and television (58%) [51].

On December 2016 the CNBC listed the biggest fake news stories of 2016 [52], most notably the title "Pope Francis shocks world, endorses Donald Trump for president, releases statement". In few weeks the fake news reached 960,000 Facebook engagements. When, on November 8, 2016 Donald Trump became the new US president, Facebook was accused of have influenced the election outcome through the propagation of the fake news. To defend his Social, on November 10, 2016 Mark Zuckerberg released a declaration: "*Personally I think the idea that fake news on Facebook ... influenced the election in any way — I think is a pretty crazy idea. Voters make decisions based on their lived experience*". He also stated that more than 99% of what people see in his Social Network is authentic [53],

Other fake news titles of 2016 included "Trump offering free one-way tickets to Africa & Mexico for those who wanna leave America", "ISIS leader calls for American Muslim voters to support Hillary Clinton" and "FBI agent suspected in Hillary email leaks found dead in apparent murder-suicide" [52]. According to Buzzfeed [54], these stories boasted nearly two million Facebook engagements,

while in the same period the top performing Facebook story from the New York Times racked up just over 370,000. An investigation traced some of these fake publishers to Veles, a small town in Macedonia, where it has been discovered that over 140 fake news sites are based. In January 2017, Google and Facebook took the first concrete steps to curb the number of false news articles propagated across their sites [9].

1.3 Fake Reviews

Lust but not least, another serious problem related to online deception concern fake reviews [55], an issue that companies like Amazon, TripAdvisor and Yelp are now facing to guarantee the economic interests of their millions of customers [10]. Fake reviews refer to contents created by users and posted in internet where they express a false opinion about a product or a service. A 2013 European Consumer Centres' Network web survey showed that the 82% of respondents read consumer reviews before shopping [56].

An yearlong investigation has discovered different companies, which were located in Bangladesh, Philippines or Eastern Europe, where fake-review operators produced, for as little as a dollar, amazing comments for places that they had never seen in countries where they had never been [57]. The fraudulent reviews were posted on sites like Google, Yelp, Citysearch and Yahoo. A fake review of a restaurant may lead to a bad meal, which is disappointing, but the investigation uncovered a wide range of services buying fake reviews that could do more permanent damage, such as dentists, lawyers, even an ultrasound clinic.

Reviews can be categorized in three groups: positive reviews, negative reviews and neutral reviews. All these three types of reviews can be faked, especially positive and negative. In fact, a company could create a positive review to increase the number of its costumers or a negative review to decrease the number of the competitor's costumers.

In a recent study, the authors investigated the economic and business incentives to commit review fraud on the popular review platform Yelp [58]. They found that a restaurant is more likely to commit review fraud writing false positive posts when its reputation is weak (e.g., when it has few reviews, or it has recently received bad reviews), whereas it is more likely to leave reviews for competitors and when it faces increased competition. Chain restaurants, which benefit less from Yelp, are generally less likely to commit review fraud. Moreover, they estimated that about the 16% of the restaurants reviews on Yelp are faked.

The economic damage caused by faked reviews is huge, so much so that on October 2015 the European Parliament released a briefing on this specific issue, assuming some possible line of action to fight it [56]. The report stated that *“the problem of fake online reviews not only concerns individual consumers; it can lead to an erosion of consumer confidence in the online market, which can reduce competition. To deal with this issue, some guidelines have already been adopted by consumer enforcement bodies, regulators and other stakeholders, in the EU and internationally. Enforcement actions have also been taken. Fake online reviews should be taken seriously, as more and more consumers buy online, and the practice is becoming increasingly sophisticated”*.

In conclusion, is it possible to prevent the fake news posting and block it before they are write? Is it possible to identify fake reviews and black it out? Is it possible to detect people subscribing a website or a social with a fake identity? At date, these are not still effective possibilities, but research on online lie detection is going in this direction.

This work is a contribution to online lie detection. The main purpose is to investigate the possibility to detect deception analyzing the interaction between user and computer. In the next chapters, an exhaustive review of the current lie detection methods, focusing on their advantages and disadvantages for online application, is reported. Then, we present a series of proof of concept studies aimed to expand the scientific knowledge about human-computer interaction during the act of lie.

Chapter 2 Related works

The psychology of lying is one of the topics that have interested the researchers in the last century [59]. According to Abe, deception is a “*psychological process by which one individual deliberately attempts to convince another person to accept as true what the liar knows to be false, to gain some type of benefit or to avoid loss*” [60].

The first scientist who attempted to create a lie detector was Vittorio Benussi. First professor of experimental psychology at University of Padova, Benussi proposed to identify deceptive responses based on the subject’s psychophysiological measures, especially the breathe [61]. The instrument that he used to record the breath was called pneumograph, which consisted in a thoracic band that allowed to record respiratory movements and therefore to calculate the duration of inspiration and expiration. The basic idea was that when a person is lying his expirations become longer, whereas inspirations are shorter. Conversely, a truth-teller show an inverse inspiration-expiration duration pattern.

Lie detection techniques based on psychophysiological measures, like the pneumograph of Benussi, are known as emotion-based lie detection techniques [62]. Emotion-based lie detection highlights deception through the physiological reactions (arousal) that are induced by lying. These are the best known lie detection techniques and the most widely-used. In this category fall the famous polygraph [63], which is commonly associated with the Control Questions Technique (CTQ) [64] or the Guilty Knowledge Test (GKT) [65]. The emotion-based techniques are very diverse, according to the indices that are measured: heart rate analysis [66], eye tracking [67], thermography [68], voice stress analysis [69], facial expression analysis [70]. However, the link between deception and arousal has been questioned, in particular because an person’s physiological activation can be explained by many reasons and because not all the individuals are aroused when they produce deceptive responses [71]. More recently, researchers began to study the neural correlates of deception, using event-related potential (ERP) [72] or functional magnetic resonance imaging (fMRI) [73] as lie detectors. The basis of ERP techniques is the fact that recognition of infrequent and familiar events (e.g., crime scene details) modulates brain potentials such as the P300 or that the response conflict (e.g., the inhibition of an honest response while producing a deceptive one) modulates the amplitude of medial frontal negativities. The use of fMRI in the lie detection is aimed to obtain measurements of cerebral blood flow (a marker for neuronal activity) in individuals engaged in deception, showing differences in brain activity between deceptive and truthful responding.

As we are interesting in the problem of online deception, here we not further discuss the psychophysiological lie detection techniques. In fact, they are not currently usable to spot online lying, as they require very specific instrumentation and are not easily accessible to the average user who navigates in internet (they are very expensive, require a very specific expertise and take a lot of time to be set up).

Currently, the lie detection approaches that could be used to address the problem of online deception are essentially two: the linguistic approach and the behavioral approach. In the following paragraphs, a description of these approaches is reported, as well as the state of the art about their application to detect lies.

2.1 Linguistic Approach

Different studies have demonstrated that liars utilize language differently than truth-tellers and some linguistic cues can predict which speeches or texts may contain deception [74]. For example, Newman et al. [75] tried to distinguish true and false stories extracting linguistic features. They demonstrated that compared to truth-tellers liars showed lower cognitive complexity, used fewer self-references and other-references, and used more negative emotion words. According to these features, a computer-based text analysis program correctly classified liars and truth-tellers with an accuracy around 65%.

The verbal approach is widespread in the forensic environment to distinguish true from false statements made by crime victims [76]. One method, which falls under the verbal approach to lie detection, is the scientific content analysis (SCAN) [77]. The SCAN is the most frequently used verbal credibility assessment method [78]. This technique analyses the words that people use to try to determine if what they said is accurate. The basic idea is that liars and truth-tellers differ in the type of language that they use. Following this hunch, the SCAN proposes a list of linguistic criteria that could assist the examiner in differentiating between true and false statements [79]. However, at date there are no scientific evidence that confirm the discriminative power of the SCAN [76], [80].

Another linguistic technique to analyze the words that people say during their declarations is the statement validity assessment (SVA) [81]. It was originally designed to determine the credibility of child witness testimony in trials for sexual offences. The core of the SVA is the criteria-based content analysis (CBCA), which is aimed to distinguish true and false declarations [82]. The theory under the CBCA is that children's statements about true events differ in content and quality from their statements about fabricated events. Based on these differences, a list of criteria to evaluate the credibility of witness testimonies have been developed [83]. CBCA criteria are of two types: cognitive and motivational criteria. The first are likely to indicate true statements, as they are typically too difficult to fabricate (e.g., details about time and place). Conversely, motivational criteria refer to how the witness presents the statement: as liars focus on making a credible impression, they leave out from their stories the information that may cause damages (e.g., admitting lack of memory) [84]. However, numerous studies have shown that CBA is usefulness with adult victims and eyewitnesses [85], [76]. Vrij et al. [84] reported that the CBCA has an accuracy ranging from 55% to 90%, with an average accuracy of 70%.

A third verbal lie detection technique is the reality monitoring technique (RM) [76]. The rationale behind the RM is that memories of true events will differ in quality and content from fabricated memories [86]. The criteria to difference true and false statements include different aspects, such as realism, details about space and time, sensory information and clarity or vividness [87]. Some studies demonstrated that the RM has a comparable accuracy to that of CBCA, with percentages of accuracy ranging from 61% to 83%, with an average of 69% [78].

More recently, the linguistic approach has been applied also to online texts, especially reviews.

For example, Moilanen et al. [88] developed a software, named TheySay, which is able to measure the sincerity of a written text based on the textual sentiment analysis. Mihalcea and Strapparava [89] collected deceptive and truthful opinions on different issues, such as abortion and death penalty, from Mechanical Turk participants. In particular, for each topic, they solicited one truthful and one deceptive instance. Using a classifier based on psycholinguistic analysis, the authors classified true and false statements with an accuracy around 70%. Ott et al. [90] asked Mechanical Turk participants to generate a 400 positive faked reviews on a set of hotels. Then, fake reviews were combined with 400 positive truthful reviews from TripAdvisor on the same hotels. The text of true and false reviews were used to train a machine learning classifier based on three different approaches: genre identification, psycholinguistic analysis and text categorization using n-gram features. Results indicated that deceptive and truthful reviews were identify with an accuracy of 90%. Comparing the fake reviews obtained by Ott et al. [90] with the reviews classified as faked by the Yelp's filtering, Mukherjee et al. [91] observed hat the false statements produced by Turkers' participants were not representative of the real-life fake reviews. For this reason, they concluded that the linguistic approach is not really effective in the real-life setting, which is what has been shown by studies where participants are instructed to lie.

Some authors in literature have argued that lie detection would be most accurate if both verbal and nonverbal indicators of deception are taken into account [92]. Following this line of thought, several studies have focused on both linguistic and behavioral features, trying to improve the performance in classification of a text as genuine or faked. For example, Zhou [93] and Derrick et al. [94] investigated whether linguistic and behavioral indicators can be used for deception detection in instant messaging. Zohu [93] explored different nonverbal and verbal behaviors (participation level, discussion initiation, cognitive complexity and non-immediacy of sentences, frequency of spontaneous corrections, lexical and content diversity) during a chat discussion between participants, showing that these indices could significantly differentiate deceivers from truth-tellers. Derrick et al. [94] submitted the participants to a computer-mediated chat-based. They were instructed to be sincere or liars in response to each question according to a prompt given by the system. The system captured four kind of features: response time, number of edits (basic keystrokes, such as backspace and delete keys), word count and lexical diversity. Results showed that deception was positively correlated with the response time and the number of edits and negatively correlated to the word count.

The linguistic approach, combined with other behavioral elements, is now beginning to be used also from companies, like Google and Facebook to recognize the fake news [9]. One of the first online service that applied it to the detection of false online information was Yelp, which from 2005 use a filter to remove suspicious or fake reviews [95]. The Yelp's algorithm is trade secret, but recently some authors tried to speculate about it functioning [91], [96]. Darnell Holloway, senior manager at Yelp, declared: "*Yelp has software that evaluates every single review based on quality, reliability and user activity on Yelp*" [97]. Quality refers to the fact that reviews must have new, helpful and pertinent information. Reliability is a feature that concerns the user. When a user sets up his account, he has

the opportunity to give Yelp personal information, such as date of birth and city. The more information Yelp has about a user, the more “reliable” he is considered. Finally, the third parameter regards the activity performed by the user: users with less activity, fewer friends or fewer reviews are less likely to have their review recommended.

This declaration seems to coincide with the results of an exploratory study conducted by David Kamberer [96]. The author performed a content analysis of a subset of Yelp restaurant and religious organization reviews, unfiltered and filtered, exploring signals from the reviews or the reviewers that might explain the filtering process. The study found that factors intrinsic to the review itself are not related to filtering, but factors related to the reviewer are strong predictors. According to Kamberer, the Yelp system is much more likely to filter reviews from occasional, isolated reviewers than from prolific, socially connected reviewers. Mukherjee et al. [91] used a linguistic n-gram based approach to classify filtered and unfiltered Yelp’s reviews, discovering that it was not effective in detecting fake reviews. They observed that fake (filtered) and non-fake (unfiltered) reviews from the same user were linguistically similar, which explains why fake review detection using n-grams was not effective. However, they noticed that the spammers left behind some specific psycholinguistic footprints, which were indicators of deception. Then, authors have supposed that Yelp’s filtering algorithm is correlated with abnormal spamming behaviors, founding that the behavior analysis was highly effective in detecting fake reviews than the linguistic n-grams approach.

2.2 Behavioral Approach

The last emergent approach in the study of lie detection is the behavioral one. It consists in the observation of the nonverbal behavior of the suspect while he is producing a deceptive or truthful response [92]. For example, in online context it would be possible to observe the interaction between the user and the computer or the user’s behavior in the navigation space while he is posting a faked post on a social network.

One behavioral response feature, which is commonly used as lying index, is the reaction time (RT). RT-based lie detection techniques are based on the response latencies to a stimulus of interest [98]. In particular, it has been demonstrated that people manifest a lengthening of RT and an increase in error rate when they lie in response to questions [99]. This techniques find their roots on the cognitive load theory, according to which lying is cognitively more demanding than truth-telling and this higher cognitive complexity reflected itself in a number of indices of cognitive effort, including, for example, reaction times [100].

The deception production is a complex psychological process in which cognition plays an important role [101]. During the generation of a false response, the cognitive system does not simply elaborate a statement, but it carries out several executive tasks: it inhibits the true response and, subsequently, it produces a false response [102]. Moreover, the generation of a lie requires to monitor the reaction of the interlocutor and to adjust the behavior congruently to the lie [103]. All these mental operations cause an increase in cognitive load and, generally, a greater cognitive load produces a bad performance in the task the participant is carrying out, in terms of timing and errors [104]. This phenomenon has been observed by studying the RTs in double choice tasks: the choice between two alternatives becomes slower in the deceptive response than the truthful one [105].

According to the functioning of our cognitive system, behavioral-based lie detection tools have been proposed to identify liars, as the RT-based Concealed Information Test (CIT-RT) [106], the Timed Antagonistic Response Alethiometer (TARA) [107], the a Sheffield lie test [108], and the more recent autobiographical Implicit Association Test (aIAT) [109]. All these techniques are computerized task in which subjects are asked to respond to the stimuli, which are presented on the computer screen, pressing one of two alternative keys on keyboard.

The autobiographical Implicit Association Test (aIAT) is a variant of the IAT by Greenwald, McGhee, and Schwartz [110]. It is used to detect autobiographical memories encoded in the respondent's mind. In particular, it determines which one of two alternative memories is true and, consequently which one is false. During the task participants have to classify stimuli as quickly as possible in four different categories using two keys on keyboard:

- two target concept categories (e.g., China vs. Tuscany, example of stimuli: "I went in China for Christmas" vs. "I went in Tuscany for Christmas")
- Two attribute categories (true vs. false, example of stimuli: "I am in front of computer" vs. "I am climbing a mountain").

Then, in one combined block, two categories (one from the target concept and one from the attribute dimension) are mapped on the same response key. In a reversed combined block, participants have to classify the same four categories reversely paired, so that both target concept categories are paired with both attribute categories. The IAT effect is expressed as the difference between the combined and reversed combined blocks. In the block where two associated concepts require the same motor response, RT will be faster than in the block where the same two concepts require different motor responses [111].

The Sheffield lie test consists in presenting autobiographical questions to which participants have to provide "yes" or "no" responses using one of two different response keys [112]. Questions can appear in two different colors: participants are instructed to lie if the sentence is presented in the one color (lie-trials) and to tell the truth if the sentence is presented in the other color (truth-trials). Experiments that applied this paradigm found that liars had slower RT and made more errors on lie-trials compared to truth-trials.

The TARA requires subjects to classify a succession of mixed statements as true or false, as quickly and accurately as they can, by pressing one of two keys. Specifically, it requires to truth-tellers to complete two alternating tasks using the same strategy, but requires liars to complete them using contradictory strategies. The faster they do so, the more likely they are to be telling the truth; the slower they do so, the more likely they are to be lying. Experimental studies reached an accuracy rate around 85% in detecting liars [107].

Finally, the CIT-RT applies the theoretical framework of CIT (previously known as Guilty Knowledge Test) to reaction times. It consists in presenting the critical information within a series of very similar, noncritical sources of distractor information. For example, if the concealed knowledge about a murder weapon is under scrutiny, a gun (the known murder weapon) will be presented together with distractors that are also potential murder weapons (e.g., a knife, etc.). For the innocent

subjects, the response is expected to be similar to all stimuli. By contrast, for the guilty subject (with guilty knowledge for the weapon), longer responses for the critical item are expected (e.g., the gun) [113].

To the best of our knowledge, there are not online applications of lie detection which are based on RT. The first researchers who have proposed the RT-based lie detection for the web are Bruno Verschuere and Bennett Kleinberg [114]. Through a web platform, the authors applied the CIT to the identity detection [115]. Participants were asked to compile a first online form with their personal information. Secondly, they were asked to learn a false identity and to perform the CIT task pretending that the learnt identity was their real one. Results showed that the CIT identified the true identities of the participants with an accuracy ranging from 86% to 94%.

The efficiency of the behavioral lie detection techniques has been proved. However, there are some limits in the online practical application. First, all these methods require a prior knowledge about the information that has to be checked as true or false. In fact, they require that the true information (or the true memory) is available, while in most real applications this is unknown to the examiner. For example, if we want to check the truthfulness of the user identity, all these techniques require that the true identity is available. Nevertheless, in some cases, such as the Facebook's users, the true identity is unknown. This means that given two alternative information these techniques can say which is true and which one is false, but they can't say in absolute terms whether the reported information is true or false. Secondly, RT based techniques only study the response latency, therefore liars have just to check a unique parameter to fake the lie detector. Even though RT are implicit measures, during the aIAT or CIT examination the lie detection purpose is explicit (overt detection of deception). Both the science and the everyday practice, show that the best indices of lying are implicit behaviors that the subject puts into action automatically. For this reason, the ideal situation to detect deceptions is realized when the examinee is not aware about the lie detection purpose and about the parameters that are collected (covert detection of deception). Finally, all the above mentioned RT-based lie detection techniques are conceived for the application in court or in other face-to-face situations. Thus, they do not follow the natural flow of the online activities. Finally, their administration is pretty long, and sometimes the instruction to complete the test, such as in the case of aIAT, are very complex. In other words, they are difficult to be integrated in the current online services.

2.2.1 Mouse tracking

Latency measures can be collected not only using two alternative (yes/no) response buttons, but can be embedded in more complex measures. It has been proposed that kinematic indices or keystroke characteristics provide a clue to recognize the deceit during human-computer interaction [116], [117]. In this and in the next paragraph, a description of these techniques and their application in detecting lies are reported.

Recently, researchers found as a simple hand movement can be used to study the continuous evolution of the mind processes underlying a behavior [118]. Especially, the hand motor response has been tracked during computerized multiple-choice tasks to understand the dynamics of a broad range of psychological processes [119]. Results have shown that hand-tracking can provide an high-

fidelity real-time motor trace of the mind [120]. Differently from RT, hand movement is not a static index resulting from a cognitive operation, but it is an online measure of the entire process.

A very easy way to capture the hand movements during the subject's response on computerized tasks is to track mouse dynamics [121]. Mouse dynamics refers to all the information that describe a mouse movement in terms of spatial and temporal features. This procedure has recently been applied to a large number of fields and has proved to be useful in highlighting the cognitive complexity of the subjects' responses. Mouse tracking has been used to investigate the cognitive processes of negative sentence verification [122], racial attitudes [123], perceptions [124], prospective memory [125], and lexical decisions [126].

Applying this evidence to the study of lie, Duran, Dale & McNamara published the results of the first work in which hand movements were used to distinguish deceptive responses to the truthful ones [116]. During the task, participants were instructed to answer "yes" or "no" questions about autobiographical information. Questions appeared on a screen and participants were asked to respond using the Nintendo Wii controller. For each participant, half of the trials required to respond truthfully and the other half required a false response, according to a visual cue that appeared on the screen along with each question. Results indicated that deceptive responses could be distinguished from truthful ones based on several dynamic indices, such as the overall response time, the motor onset time, the arm movement trajectory, the velocity and the acceleration of the motion.

Hibbeln and colleagues studied the mouse movements in an insurance fraud online context [127]. During the task, participants were asked to claim damages to their car by compiling an online insurance claim form. In particular, they had to indicate the repair costs and the locations of the damage. Some participants were advised to declare a higher number of car damages in order to obtain a greater compensation by the insurance. Results suggested that being deceptive increased the normalized distance of movement, the speed of movement, the response time, and resulted in a higher number of left clicks.

Similar results have been obtained by Valacich et al. [128] who proposed a pilot study to identify guilty individuals involved in specific insider threat activities. They analyzed mouse movements while participants compiled an online survey similar to the CIT. Their preliminary observations showed that guilty insiders had a different mouse movement pattern when answering to key-items as compared to non-key-items.

At date, these are the only studies that investigate the effect of deception on mouse dynamics. There are also several studies in literature that applied mouse movement analysis to the problem of the user authentication or identification. However, the main limitation of these methods is that they require necessarily a certain level of knowledge about the alleged user, and a specific training, in order to be able to recognize the intruder. In other words, they are not lie detection techniques, but biometrics techniques similar to finger print and face recognition. Once again, these methods cannot be used to spot fake identities in absence of a ground truth.

2.2.2 Keystroke dynamics

Keystroke dynamics is the detailed timing information about human typing rhythm: it describes exactly when each key is pressed and released, while a person is typing at a computer keyboard, a mobile phone or a touch screen panel [129], [130]. It has been widely used as a biometric measure for user authentication [131], [132] but, similar to mouse dynamics authentication, it requires a prior knowledge about an alleged user to recognize him.

Concerning the application of keystroke dynamics to lie detection, only few studies focused on it, and the majority of these principally used a linguistic approach to deception detection or they considered only some simple features of the text rather than the rhythm of typing (see section 2.1).

The first authors who proposed to apply keystroke dynamics to were Grimes, Jenkins and Valacich [117]. They conceived the Keystroke Dynamics Deception Detection model to explain the relationship between deceptive behavior and keystroke dynamics. According to this model, the production of a falsehood may cause an increase both in emotional arousal and in cognitive load. These increments may result in a consequent change in the fine motor control, which in turn results in a deviation of the typing ability, affecting the keystroke dynamics personal baseline. The Keystroke Dynamics Deception Detection model was tested in a pilot study. Each subject shared three statements, two truthful and one falsehood, on a web page. Keystroke characteristics were captured by a JavaScript based web application. In this paper, the authors did not discuss the results, but they brought some limitations of the study that could be a good point for future developments.

Banerjee et al. analyzed different keystroke parameters to improve the performance of a classifier in distinguishing truthful and deceptive writers of online reviews and essays collected via Amazon Mechanical Turk [133]. Each Turker wrote a truthful and a deceptive text about three topics (restaurant review, gay marriage and gun control); the order of the true/false texts was balanced between subjects. Mouse and keyboard events (KeyUp, KeyDown and MouseUp) for all texts were captured. Moreover, the authors considered the following features: editing patterns (e.g., number of deletion keystrokes, number of MouseUp events, number of arrow keystrokes), temporal aspects as writing speed and pauses (e.g., timespan of entire document, average timespan of word plus preceding keystroke, average keystroke timespan, average timespan of spaces, average timespan of non-white spaces keystrokes, average interval between words) and writing speed variations over word categories (e.g., nouns, verbs, adjectives, function words, content words). They implemented a binary SVM classifier, which achieved a baseline average accuracy of 83.62% on linguistic features. Introducing keystroke features, they obtained a statistically significant improvement of the deception detection classifier, ranging from 0.7% to 3.5%.

To conclude, although keystroke dynamics are mouse dynamics are promising technique for online deception detection, the literature on these topics is not enough. For this reason, we have looked at the matter further through the experiments that are reported in Chapter 4.

2.3 Strategies to Increase Cognitive Load

In sections 2.1 and 2.2, we analyzed different methodologies (RT, mouse dynamics and keystroke dynamics) to record the distinctive features of the deception. The present paragraph focuses on how it is possible to increase the liars' identifiability.

We argued that there is evidence that the process of inhibiting the truthful response and substituting it with a deceptive response may be a complex cognitive task, which results in an increase of response time. However, in some instances, responding with a lie may be faster than truthfully responding [112]. In fact, distinct types of lies may differ in their cognitive complexity and may require different levels of cognitive effort. For example, the cognitive effort may be minimal when the subject is simply denying a fact that actually happened [134] or when the lie has been overlearned [135]. In other words, when the response is automated the liar's cognitive load remains intact and his response becomes indistinguishable from the response of a truth-teller. This evidence, anticipated by previous literature [135] [136], it was found also in our experiments (see paragraph 4.3.8).

To overcome this issue, authors in literature have proposed different strategies, which are usually used in police interrogations, and which allow increasing cognitive load of liars, keeping unaltered the cognitive load of truth-tellers [101].

A first strategy to increase the cognitive load during an interview consists in asking the subject to perform a second task at the same time as the interview. This allows minimizing the cognitive resources of the subject, which are destined to the lie, rising clues of deceit. Liars, who are already partially committed to lying, should find the additional task as particularly exhausting [137].

Similar to dual task, is the introduction of task switching. Task switch can be understood as continuously move from a task (e.g., lying) to another (e.g., counting) or switch up from lying to truth telling [138]. In this second sense, some authors have proposed to manipulate the proportion of lie/truth-trials across participants, to investigate how the continuous change from truth to lie telling would affect the cognitive load and, consequently, the subject performance [136]. A research has shown that the cognitive cost of deception decreases when people frequently respond deceptively, while it increases when people rarely respond deceptively. People who often responded deceptively are faster and made fewer errors than people who only occasionally responded deceptively.

Another efficient technique to increase the cognitive load is the time restriction of the response [139]. This strategy consists in ask the examinee to answer questions as quickly as possible. This limits opportunity for liars to self-monitor and control the response. The high cognitive load of rapid responding questions may increase the number of deception cues, such as voice pitch elevation, pupil dilation, reduced blinking, long response times, accidental blurting of the truth because of the and may increase.

Soliciting surprise drawings is another efficient strategy to spot liars. In fact, Vrij et al. observed that truth tellers' drawings of their workplaces contained more plausible details, especially those involving their coworkers, than liars doing the same [140].

Vrij et al. proposed to impose cognitive load asking participants to keep the eye contact with the examiner during an interview. The half of participants were asked to maintain the eye contact during the entire interview, whereas the other half responded to the examiner's questions without any specific instruction. Results showed that liars were more detectable when were obliged to keep eye contact than in the control condition.

Another technique used by Vrij and colleagues to improve cognitive load, is to ask subjects to recall their stories in an inverse order [141]. Asking interviewees to report their stories in reverse order is already practiced in police interviews in several countries. The basic principle under this strategy is that recalling an event starting from the end interrupts the prepared sequence of false events and requires activating the working memory to manipulate the recall and monitor its coherence. An experiment confirmed that instructing interviewees to recall their stories in reverse order facilitated detecting deception. Half of participants were asked to reports their stories in reverse and the other half in chronological order. They were video recorded and observed by police officers, who identified the liars with an accuracy above the chance level.

However, some authors tried to warn researchers against the possibility that in some situations these cognitive strategies could be challenging for truth-tellers as well. For example, recalling very distant memories or future intentions may be overloading in terms of cognitive effort. As consequence, the differences between liars and truth-tellers become minor [142].

In the following subsections, two more strategies to increase liars' cognitive load are describe in more detail, since they have been included in the experiment presented in Chapter 4.

2.3.1 Unexpected questions

The strategy of asking unexpected questions was pioneered by Vrij and co-workers [143]. Unexpected questions are questions to which liars cannot prepared their response in advance. On the contrary, expected questions are those questions to which liars have fabricated the response previously. Imaging a person who decided to lie about his identity on a social network. An example of expected questions is "which is your date of birth?", whereas an unexpected question could be "which is your zodiac?". As liars are not prepared to answer unexpected questions, they are forced to generate new details that were not part of their original script. In the specific example, when a user subscribes a website he knows that the date of birth is one of the information commonly asked. Conversely, the zodiac sign is not frequently asked. Thus, the liar has to fabricate the false response in real time, checking the congruency of the response with the other faked information and maintaining his credibility and consistency [144]. Liars give their planned responses to expected questions easily and quickly, but they need to fabricate plausible responses in the case of unexpected questions, and this yields an increase in the cognitive load. By contrast, truthful responses are not plagued by the side effects of the cognitive load as they are quite automatic and effortless for both expected and unexpected questions.

The technique of unexpected questions was originally applied to in investigative interviews [145]. The procedure commonly provides that the examiner initially ask anticipated questions and then he switch to unanticipated questions. Such questions may relate to particular spatial or temporal aspects of the recalled event. In a first experiment, Vrij et al. [143] interrogated couples of liars and truth-

tellers about a lunch that they had together at the restaurant. The couples of truth-tellers really ate out together at the restaurant, whereas liars were instructed to pretend to had lunch together. During the interview the examiner asked questions about temporal and spatial features of the lunch-event (e.g., who finished to eat first? Where was your table located?). Comparing the responses to unexpected questions of the two member of each couple, liars were identified with an accuracy of 80%. The classification was based on the discrepancies reported by the two persons at the same questions. Then, a second experiment was run by Lancaster et al. [146], investigating the power of unexpected questions in a between subjects experimental design. This time participants were not paired, but the single subject was asked to lie or tell the truth. Results showed a good classification rate for both truth-tellers (78%) and liars (83%). The authors observed that liars, with respect to truth-tellers, reported many more details to the expected questions versus the unexpected questions, and the lie detection could be based on this difference.

2.3.2 Complex questions

Unexpected questions are a powerful tool for uncovering deception, but they cannot be used in every condition. When responding to unexpected questions, liars have to process the information in the questions in real time as quickly as possible so that cognitive processing load is combined with time stress in the performance of the task. Within the cognitive load approach to lie detection, one unsolved problem is the identification of a liar when unexpected questions are not available. Typical conditions when unexpected questions cannot be used are in the so-called “lies of omission” [147], which consist in denying something that did happen (“I did not do it” type of lies). For example, if a guilty subject is denying any wrongdoing (in a “Did you do it?” type of question), it is difficult to create an unexpected question that efficiently uncovers his deception.

In this work, we present a new technique for detecting lies when unexpected questions cannot be crafted. The technique consists of presenting complex sentences. We have named “complex questions” the sentences that contain more than one information in the same phrase. For example, to investigate the identity one could ask a question about the name (e.g., Is Alice your name?) and a question about the place of birth (e.g., Were you born in Montréal?). A complex question encompasses both this information in the same sentence (e.g., Are you Alice born in April?).

Asking complex questions is very closed to increasing the number of response alternatives among which the liar has to choose. Williams et al. [148] stated that when questions involved more than one possible lie response, liars reveal a greater response latency. By contrast, the authors found that the number of alternatives did not significantly affect response times when individuals told the truth. In fact, in real life situations, the subject has to choose one lie in a range of endless possibilities, deciding which the better one is according to the context. The greater number of alternatives requires more cognitive effort by liars who need to monitor the plausibility of more than one information.

Our hypothesis is that complex questions require greater cognitive resources compared to simple questions, because they require subjects to analyze each information one by one and labelling it as true or false. In other words, the subject has to monitor the plausibility of more than one information and retain it in working memory to, finally, decide if the entire sentence is true or false. While truth-tellers can speedily carry out this sequence of mental operations, liars need more time to match the

plausibility of each information with the lie they told [102]. As result, we expected that liars have a bad performance, compared with truth-tellers, when they are involved in a decision task, making a greater number of errors and showing slower reaction times. This hypothesis is investigated in the experiments reported in sections 4.4, 4.5 and 4.

Chapter 3 Materials and Methods

In this chapter, a general description of methods and materials is reported. These were similar for all the experiments that are presented in next chapter, except when specified.

The general experimental procedure was approved by the Ethics Committee for psychological research – Padova University Psychology Department, and it was in accordance with the relevant guidelines and regulations [149].

3.1 Participants

A total number of 640 healthy participants took part to the experiments. They were volunteers, mostly recruited among students of Padova University. All participants were over 18 and provided the informed consent before the experiment. Subjects did not receive any compensation for participation.

Exclusion criteria have not been applied, except for language. In fact, it was required that the subject's mother tongue was the same of the experiment. It was to avoid an influence in response times due to comprehension difficulties. Moreover, each subject participated to just one experiment, to avoid that his performance could be compromised by a previous knowledge of the experimental procedure. Data of subjects who did not understand the task were removed before the analysis.

3.2 Experimental Design and Procedures

All the experiment were built according to a between subject experimental design. Participants were always randomly assigned to one of two experimental groups: liars or truth-tellers. Liars were participants instructed to lie about a topic (e.g., their identity), whereas truth-tellers were subjects who where asked to respond truthfully. Every experiment counted at least 40 participants, 20 for each experimental condition. The sample size is similar to other lie detection researches based on response latencies [111]. It has been calculated that a sample size = 40 is enough to have a statistical power $(1 - \beta) = 0.8$, given a significance level $(\alpha) = 0.05$ and a medium effect size $(d) = 0.5$ [150]. Moreover, for each experiment we confirmed that the two experimental groups were similar in terms of age, gender and schooling ($p > .05$ both for age, gender and schooling).

3.3 Data Collection

All the experiments took place in the laboratories of the Department of General Psychology – University of Padova.

Experiments were conducted measuring indices deriving from three different tools of human-computer interaction: reaction times (RT) on keyboard, keystroke dynamics and mouse dynamics. In the following, the materials used to collect data are summarized and all the indices derived from keystroke and mouse dynamics are described in detail.

3.3.1 Reaction times

Experiment measuring RT were programmed in E-Prime[®] 2.0 [151]. All the experiments were run on a single laptop ASUS K56C with a LCD 15.6" diagonal screen.

Stimuli appeared in the center of the computer screen and the two response labels were placed in the right and left upper corners.

To give their response, subjects were instructed to press the key "A" or "L" on the computer keyboard that corresponded respectively to the left and right response label. In Figure 3.1, an example of user-computer interaction during the experimental task is represented. All stimuli were presented randomly.

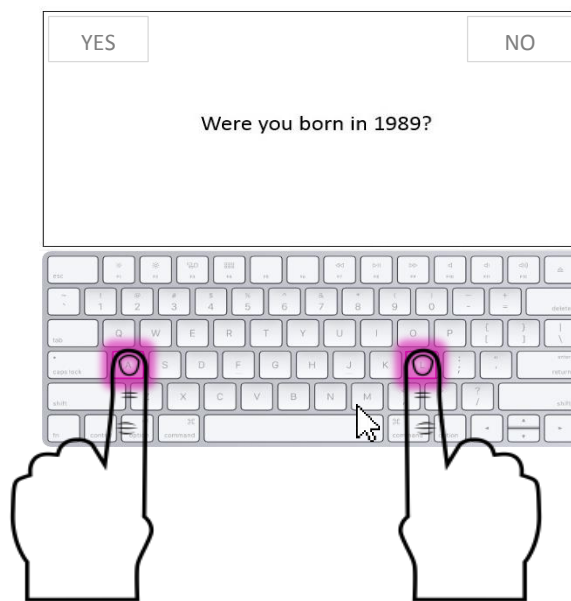


Figure 3.1: an example of user-computer interaction during the experimental task.

Each stimulus appeared automatically after the response to the previous one, so no actions were required to subjects to bring each new question. No temporal response limits were fixed and no feedbacks were provided for wrong responses.

During the task, we recorded RT to each stimulus and errors. For errors, we mean the wrong responses given by the subject according to the information that he reported, independently to the fact that he was a liar or a truth-teller. Then, for each participant, we averaged the RT and errors of all stimuli. We also calculated the Inverse Efficiency Score (IES), an index that combines speed and accuracy [152]. In fact, it is possible to increase the response speed but it usually leads to a higher percentage of error (PE). The IES takes into account the number of errors and increases proportionally the average RT of the subject according to the following formula:

$$\text{IES} = \frac{\text{RT}}{(1 - \text{PE})}$$

Equation 3.1: calculation of the Inverse Efficiency Score.

3.3.2 Mouse tracking

All the experiments measuring mouse dynamics were programmed using MouseTracker software [121]. Each experiment was run using the same laptop. For some experiments, we used an ASUS K56C with a 15.6" diagonal screen, for other experiments an ASUS UX303L with a 13.3" diagonal screen.

Stimuli appeared in the upper-central part of the computer screen and the two virtual response buttons were placed in the right and left upper corners. Each question appeared when participant clicked on the "START" virtual button that was located in the lower-central part of the screen. Then, subjects were asked to respond clicking with the mouse on one of the two alternative response buttons. Figure 3.2 reports an example of computer screen as appeared to the subject during the task. Questions were presented randomly.

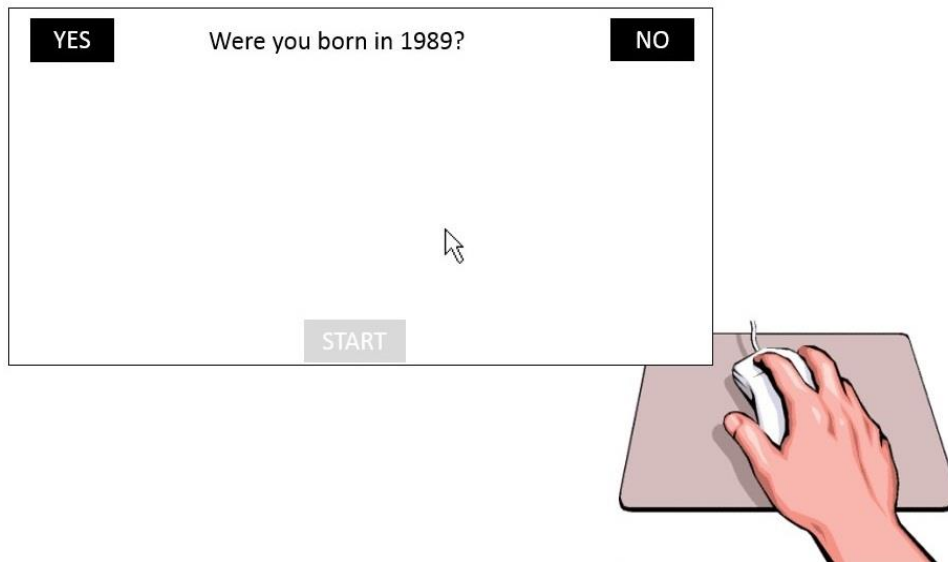


Figure 3.2: an example of computer screen as appeared to the subjects during the task.

No temporal limits were fixed to complete the response and no feedbacks were provided for wrong responses. However, it was important to ensure that participants were responding on-line (during the actual processing of lying) rather than off-line (once the lie had already been made). For this reason, we alerted participants when they were initiating mouse movement too late, setting up the threshold of initiation time to 2000 ms. When initiation time exceeded this threshold, a message appeared alerting participants that they initiated movement too late and encouraging to initiate future responses earlier. After each response, the software automatically relocated the mouse to the "START" button (origin).

The experiments operate in a coordinate space, where x value ranges from -1 to 1 and y value ranges from 0 to 1.5 (see Figure 3.3). For each response, the software records the mouse position from the origin (X_0, Y_0), which corresponds to the START button, to the click on the response button. To deal with the different length of the recorded trajectories and to permit the averaging and comparison across multiple trials with different numbers of coordinate pairs, the x, y coordinates of each trajectory are time-normalized. By default, MouseTracker performs a time normalization in 101 time frames using linear interpolation. As result each trajectory is described by 101 time frames and each time

frame corresponds to a specific x, y coordinate. In other words, the pair (X_n, Y_n) indicates the position of the mouse along the axis at the time n , where n ranges from 1 to 101.

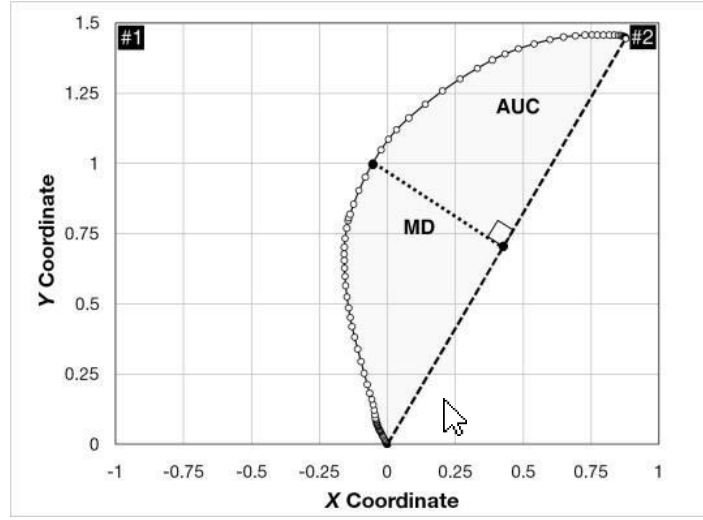


Figure 3.3: a representation of the coordinate space in which the experiments operates.

The software also describes the motor response in terms of spatial and temporal features, such as the onset, the duration, the shape, the stability and the direction of the trajectory. The space-time features recorded by MouseTracker are described in detail in Table 3.1, and graphically represented in Figure 3.3.

	Feature	Description
Temporal features	Initiation time (IT)	The time between the appearance of the question and the beginning of the mouse movement.
	Reaction time (RT)	The time from the appearance of the question to the click on the response box.
	Maximum deviation time (MD-time)	The time to reach the point of maximum deviation.
Spatial features	Maximum deviation (MD)	The largest perpendicular distance between the actual trajectory and the ideal trajectory.
	Area under the curve (AUC)	The geometric area between the actual trajectory and the ideal trajectory.
	x -flip	The number changes in direction along the x -axis.
	y -flip	The number changes in of direction along the y -axis.

Table 3.1: Description of spatial and temporal features recorded by MouseTracker software.

In addition, we calculated the average velocity (v) and acceleration (a) of the mouse movement between two time frames, respectively on x -axis and y -axis:

$$v_x = X_n - X_{n-1} \text{ and } v_y = Y_n - Y_{n-1}, 0 < n < 101$$

$$a_x = v_{x n} - v_{x n-1} \text{ and } a_y = v_{y n} - v_{y n-1}, 0 < n < 101$$

Equation 3.2: calculation of velocity and acceleration between two time frames on x and y -axis.

Finally, for every feature mentioned above, we computed the average value of the responses for each participant. We also calculated the average number of errors made by each subject in responding to questions as the ratio between the number of errors and the number of stimuli. For errors, we mean

the number of wrong responses given by the subject according to the information that he reported, independently to the fact that he was a liar or a truth-teller.

3.3.3 Keystroke dynamics

Experiments collecting keystroke dynamics were implemented in an online platform that we expressly designed using PHP, HTML and JavaScript. In particular, the recording of keystrokes dynamics and time intervals was programmed using JavaScript. All the experimental tasks are online accessible through this link: <https://truthorlie.math.unipd.it/>. Data was stored via MySQL Ver 14:14 Database.

During the task, stimuli were displayed in the central area of the screen. Participants were asked to respond each question typing the answer in an edit box. Figure 3.4 shows an example of the presentation screen and the location of the edit box filled by the participants. Stimuli appeared in random order.

After responding, participants were instructed to press ENTER to confirm the response and pass to the next question. There were no temporal limits to digit the response and no feedbacks were provided for wrong responses. A bar in the lower part of the computer screen indicated the percentage of task completion.

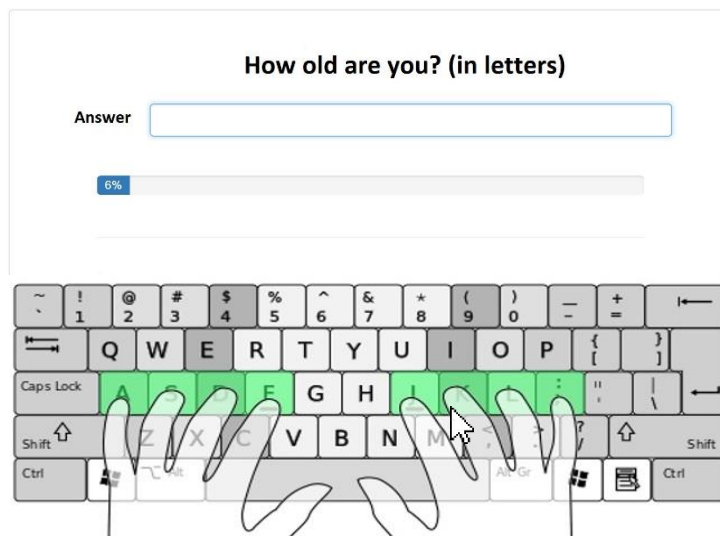


Figure 3.4: an example of the presentation screen as appeared to participants during the task.

During the presentation of each stimulus and the subject response, the website recorded the following time events: the onset of question, each key down, each key up and the push of ENTER. Starting from these events, the following features were collected for each trial:

- **Prompted-firstdigit:** it is the interval between the onset of the question on the computer screen and the first key pressed. In some experiments, this index was adjusted using a readability index for the Italian language (GULPEASE index). In other words, we refined the reaction time by weighting the latency of the response to the question for the difficulty of reading the latter. The GULPEASE is an index which takes into account two linguistic features to estimate the readability of a sentence: the length of the word and the length of the sentence in terms of number of

letters [153]. The index value ranges from 0 to 100, where 0 is low readability and 100 high readability.

- Prompted-enter: it is the total time from the stimulus onset to ENTER (pressed at the end of the response).
- Firstdigit-enter: it is the time between the first key press and ENTER.
- Time before enter key: it is the time between the first key press and ENTER.
- Answer length: it is the number of characters in the response.
- Writing time: average typing speed (firstgidit-enter /number of characters).
- Down time: it is the time stamp for each key pressing.
- Up time: it is the time stamp for each key releasing.
- Up and down time: it is the sum of down time and up time for each key.
- Press time: it is the time duration between each key down and key up.
- Flight time: it is the interleaving time between each key up and the next key down.
- Di-graphs: these are the sum of up time, down time or up and down time for two consecutive keys.
- Tri-graphs: these are the sum of up time, down time or up and down time for three consecutive keys.
- Frequency of use for special characters: it is the total number of key pressing for Shift, Del and Canc, Space and Arrows characters.

Moreover, for each feature, average, maximum, minimum, median, standard deviation and variance were computed. The final list of predictors counted 62 features, which are reported in Annex 1.

Finally, we calculated the total number of errors committed by each participant in responding questions. In detail, errors mean the number of box filled by entering the wrong information according to what the subject reported, independently to the fact that he was a liar or a truth-teller. Errors were calculated checking each response given by the subject with the conceptually correct information. An example of conceptual error was the response “Capricorn” to the question “Which is your zodiac?” when the participant’s date of birth was 20 April. We only considered conceptual errors for three reasons. Primarily, other types of errors such as typing errors were rarely detected because of the low number of words required by the responses; secondly, when found, such errors were minimal, not affecting the conceptual correctness of the answer. Finally, an indirect measure of typing errors was already given by the frequency of special characters, such as Del and Canc keys.

3.4 Data Analysis

To analyze data, we have followed the same workflow in all experiments. First, we performed a features selection to reduce the number of predictors and to select those to enter in machine learning (ML) models. Secondly, we ran some descriptive and statistical analysis on the features selected, to highlight the differences between the experimental groups. Then, the features selected were used to train different machine learning algorithms and build classification models that are able to predict whether a subject is a liar or a truth-teller. Finally, each model was tested on a new sample of participants to test the model generalization.

3.4.1 Feature selection

Features selection is a widely used procedure in machine learning models construction [154]. As we are interesting in elaborating a classification model that is able to detect liars as accurately as possible, the feature selection process is a very powerful mean. In fact, it permits to remove redundant and irrelevant features and to increase the model generalization by reducing overfitting [155] and noise in the data.

Citing Mark Hall, “*empirical evidence from the feature selection literature shows that, along with irrelevant features, redundant information should be eliminated as well. [...] A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other*” [154]. Following this methodology, in all experiments the non-redundant features have been extracted on the basis of their correlation with the dependent variable and their intercorrelation. In other words, we picked out the features that are more correlated with the class to predict (liar vs truth-teller) and less correlated one to each other.

This procedure has been performed manually or using a correlation based feature selector (CFS) [154], as implemented in WEKA 3.9 [156]. The CFS algorithm evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy with the other predictors. Subsets of features that are highly correlated with the class (the dependent variable) while having low intercorrelation are preferred. There are different methods that the algorithm can use to search the subset of predictors through the spaces of features. Here, a Greedy Stepwise search method has been chosen. It performs a greedy forward or backward search (in this case, a forward search has been used) through the space of predictors subsets. It may start with no/all attributes or from an arbitrary point in the space (we decided to start with no attributes). It stops when the addition/deletion of any remaining attributes results in a decrease in evaluation. Through running this algorithm, the best predictors were identified.

For each selected predictor, we have usually reported the point biserial correlation coefficient (r_{pb}), which relates to the correlation with the dependent variable, and the correlation matrix with the other selected features.

3.4.2 Descriptive and statistical analysis

Descriptive statistics and statistics on the difference between the experimental groups were run using R software version 3.2.2 [157]. Welch's t -test was run using ‘lsr’ package, which adjusts the number of degrees of freedom when the variances are thought not to be equal to each other [158]. The

ANOVA was calculated using ‘ez’ package [159]. Cohen’s d effect size was estimated using the ‘effsize’ package [160], which assess the d magnitude according to the thresholds provided by Cohen [161]: $d < 0.2$ is considered negligible, $d < 0.5$ is small, $d < 0.8$ is medium and $d > 0.8$ is large. For Bayes Factor calculation, the ‘BayesFactor’ package was used [162]. According to Kass and Raftery interpretation, Bayes Factor values from 1 to 3 are not worth more than a bare mention, a value between 3 to 20 is considered positive, values from 20 to 150 are strong and values greater than 150 are very strong [163].

3.4.3 Machine learning models

The predictors resulting from the feature selection were fed as input to a number of ML models in order to evaluate the accuracy in the subjects’ classification as liars or truth-tellers. Machine learning models were implemented using the data mining software WEKA 3.9 [156].

ML models were evaluated following a 10-fold cross-validation procedure [164]. The k-fold cross-validation is a technique used to evaluate predictive models by repeatedly partitioning the original sample (e.g., 40 participants) into a training set to train the model, and a validation set to evaluate it. Specifically, in 10-fold cross-validation the original sample is randomly partitioned into 10 equal-size subsamples, the folds (e.g., 10 subsamples of 4 participants each one). Of the 10 subsamples, a single subsample is retained as validation data for testing the model, and the remaining $10-1=9$ subsamples were used as training data. Such process is repeated 10 times, with each of the 10 folds used exactly once as validation data. The results from the 10 folds were then averaged to produce a single estimation of prediction accuracy.

Once the models were tuned, in most experiments we adopted a new set of participants to test the generalization of the performance on completely new data. This procedure allows us to estimate the generalization performances of the selected ML model in an unbiased way [165]. Precisely because the models are built to fit the data, it is important to know how an existing model fits new unseen data. The new group of participants (test set), was usually collected after the models were built, so the new subjects had never been seen from the ML classifiers. Data were collected by a different experimenter with respect to the one who built the models and subjects were randomly assigned to the experimental conditions, so there was not any a priori knowledge about how classifier work during the collection of the test set. In all the experiments, the sample size of the test group ranged from 10 to 20 subjects and it corresponded at least to the 25% of the training sample, a percentage that is usually regarded as satisfactory [166].

For each model, we report accuracy, recall (sensitivity or true positive rate) and precision. Accuracy corresponds to the percentage of subjects correctly classify as liars or truth-tellers. As we are interested in detecting lies, the recall is expressed as the percentage of liars who are correctly identified, whereas precision represents the fraction of correct liars among those identified as liars.

As stated above, we evaluated the classification accuracy of different ML models. This was to investigate whether the results were stable across classifiers and did not depend on the specific model assumptions. In fact, the algorithms that we have chosen are representative of different underlying classification strategies. These are the following:

- Logistic regression: it measures the relationship between the categorical dependent variable and the independent variables by estimating probabilities using a logistic function [167].
- Support Vector Machine (SVM): it is a binary linear classifier, which maps the space and divide the examples of the separate categories by a margin that is as large as possible [168], [169].
- Naïve Bayes: it is a probabilistic classifier based on Bayes' theorem that assumes the independence between the features [170].
- Random Forest: it is an ensemble learning method that operates by constructing a multitude of decision trees and combining their results [171].
- Logistic Model Tree (LMT): it combines logistic regression and decision tree learning [172].

ML models, such as some of those reported above, are difficult to interpret. Often, the mechanics that yield the algorithm to identify the single participant as a liar or truth teller is unclear. For this reason in some cases, to better understand the decision rules on which the classifications results are based on, we ran a tree model called J48 [173]. It is one of the simplest – if not the simplest – classifier in terms of transparency of the operations computed by the algorithm and it permits to easily highlighting the classification logic (even if not the most efficient) [174]. In other words, it was helpful to explain the operations performed by the algorithm on the data to obtain the classification output. Another tree model that we used with the same purposes is CART tree [175]. The CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample.

In all experiments, the parameters of the algorithms are those set by default in WEKA 3.9 [156]. For all the details on ML classifiers parameters, see Annex 2.

Chapter 4 Experiments

In this chapter a series of proof of concept experiments are presented. They investigate the human-computer interaction behavior during the production of lie. In particular, as anticipated in chapter 3, RT, keystroke dynamics and mouse dynamics are recorded. Then, these measures are used as input for predictive models, which are aimed to recognize the user as liar or truth-teller. To bring out the distinctive signs of deception, we employed two strategies to increase liars' cognitive load: the unexpected questions and the complex questions.

In order to compare the different methodologies (RT vs keystroke dynamics vs mouse dynamics, and unexpected questions vs complex questions), we have chosen to focus most of the experiments on the same topic, which is the deception about identity. As argued in the first chapter, it is a very hot issue and represents a current challenge for companies that provide online services.

4.1 The Detection of Faked Identity with Unexpected Questions and Mouse Dynamics

The general aim of the following experiments is to validate a computerized technique to spot people who declare false identity information asking unexpected questions and analyzing mouse dynamics.

Methods and results that are reported in this section have been partially published by Monaro, Gamberini and Sartori in PlosOne Journal [134] and conference proceedings [176], [177].

4.1.1 Participants

A first sample of forty participants was recruited and data were used as training set to build ML models. Twenty participants were assigned to the liars' group and the other twenty to the truth-teller condition. The demographic characteristics of the sample are reported in Table 4.1.

Then, a second sample of 20 participants (ten liars and ten truth-tellers) was collected and used as test set to assess the model generalization. Demographic information about participants are in Table 4.1.

Sample	N	Gender	Age	Education
Training set	40	M = 17, F= 23	M = 25, SD = 4.6	M = 17, SD = 1.8
Test set	20	M = 9, F= 11	M = 23, SD = 1.5	M = 17, SD = 0.8

Table 4.1: demographic information about training and test set. In the second column (N) the number of participants for each sample is reported. The third column shows the number of male and female in each sample. The fourth and the fifth columns report mean (M) and standard deviation (SD) of participants' age and education.

4.1.2 Experimental procedure

Participants assigned to liars' group were asked to learn a faked identity from a false Italian identity card (ID card). The ID card contained a photo of the subject, aside from the basic faked information about identity (name, surname, date of birth, place of birth, residence address, occupation and marital

status). An example of faked ID card is reported in Figure 4.1. There were not time restrictions to learn the new ID information, as the subjects were invited to take all the time they needed. When participants thought to be ready, they were asked to recall the faked identity twice. The examiner verified the correctness of the learned information and rectified any errors. All participants have recalled the identity information correctly within the second recall. Between the two recalls, they were required to perform some mental arithmetic as a distracting task. Finally, participants were instructed to complete the experimental task, responding to any questions according to the faked profile previously learned.

On the other hand, truth-tellers were asked to provide their identity information compiling an ID card on which their photo was attached (see Figure 4.1). After performing the distracting task, they completed the experimental task responding truthfully to all questions.

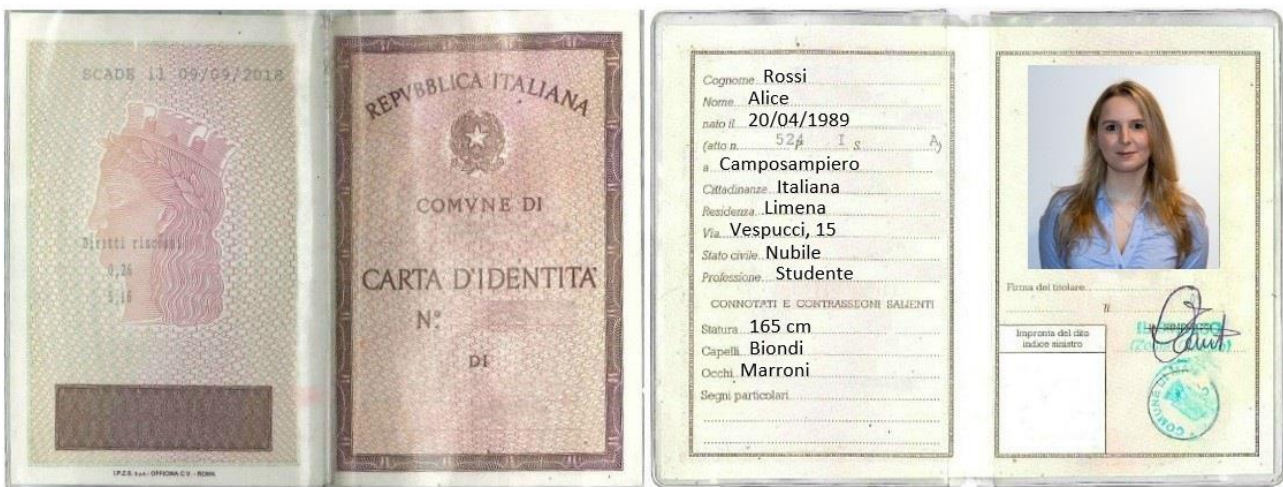


Figure 4.1: an example of faked ID card. The document reproduced an Italian standard identity card. It contains the following information: last name, first name, date of birth, city of birth, citizenship, city of residence, residence address, marital status, occupation, height, hair color, eye color. The photo posted on the ID card corresponds to the real face of the participant.

The experimental task was computerized and consisted in responding yes/no questions clicking with the mouse on one of the two alternative response labels. For more details about the modalities of presentation of the stimuli, see paragraph 3.3.2.

4.1.3 Stimuli

A total number of 32 questions were presented to each subject. Sixteen sentences required a “yes” response, and 16 required to respond “no”, for both liars and truth-tellers. The 32 experimental questions were preceded by 6 training questions (3 requiring a “yes” response and 3 requiring a “no” response) on issues related to the identity but not included in the experiment (e.g., “Is your weight 51 kg?”). Questions belonged to the following categories:

- 12 expected questions. Expected questions concerned information that was explicitly rehearsed before the experiment by both truth-tellers and liars. Liars responded to questions according to the fake identity profiles that the experimenter had assigned to them. Truth-tellers responded according to their true identities. Expected questions were about name, surname, year of birth, month of birth, town of residence and street of residence.

- 12 unexpected questions. The unexpected questions included information closely related to the false identities but not explicitly rehearsed before the experiment by either the truth-tellers or the liars. In this case, the liars responded according to the information related to the fake identities assigned to them, while the truth-tellers responded according to their true identities. Unexpected questions were about age, zodiac, region of birth, province of birth, region of residence and capital town of the region of residence.
- 8 control questions. Control questions included personal information that could not be hidden to the examiner supervising the test (e.g., the gender). For this reason, both liars and truth-tellers had to respond truthfully to these questions. Control questions were about gender, skin color, hair color and citizenship. For example, “Are you male?” (for a male subject) required a “yes” response, whereas “Are you a female?” (for a female subject) required a “no” response.

To sum up, both liars and truth-tellers responded to 16 expected, unexpected, and control questions that required to respond “yes” and to 16 expected, unexpected, and control questions that required “no” responses. Expected, unexpected and control questions were presented randomly and inter-mixed. An example of questions is reported in Table 4.2.

Type of question	Question that requires “yes” response by both liars and truth-tellers	Question that requires “no” response by both liars and truth-tellers
Expected	Is Alice your name?	Is Maria your name?
	Is Rossi your last name?	Is Bianchi your last name?
	Were you born in 1989?	Were you born in 1986?
	Were you born in April?	Were you born in August?
	Do you live in Limena?	Do you live in Caserta?
Unexpected	Do you live at Vespucci street?	Do you live at Marconi street?
	Are you 27 years old?	Are you 23 years old?
	Is Aries your zodiac sign?	Is Leo your zodiac sign?
	Were you born in Veneto?	Were you born in Campania?
	Were you born in the province of Padova?	Were you born in the province of Caserta?
	Do you live in Veneto?	Do you live in Campania?
	Is Venezia the capital of the region where you live?	Is Napoli the capital of the region where you live?
Control	Are you female?	Are you male?
	Is your skin white?	Is your skin brown?
	Do you have blond hair?	Do you have black hair?
	Are you an Italian?	Are you a French?

Table 4.2: the table reports an example of the 32 expected, unexpected and control questions presented to participants and related to a truth or faked identity.

It should be noted that liars told lies only in the expected and unexpected “yes” responses. In fact, for the liars, the expected and unexpected questions regarding their faked identities were actually “no” responses that, because they were lying, required “yes” responses. In other words, only the questions with expected and unexpected “yes” responses differentiated the two experimental groups because the truth-tellers responded sincerely, while the liars cheated. For all of the other questions (control “yes”, control “no”, expected “no”, unexpected “no”), both liars and truth-tellers responded truthfully.

4.1.4 Collected measures

During the subjects’ response, mouse dynamics were recorded by MouseTracker software. For more details about data collection, see section 3.3.2.

In addition to the errors and the spatial-temporal features extracted by default by the MouseTracker software (IT, RT, MD, AUC, MD-time, x-flip, y-flip, see paragraph 3.3.2), we analyzed the position of the mouse along the x and y -axis in search of points of maximum difference between the trajectories of liars and truth-tellers. The two groups had a maximum difference in the first half of the trajectory along the y -axis (see Figure 4.2). The points of maximum difference were Y18, Y29 and Y30, with maximum separation located at time frame Y29. Then, we calculated the velocity between these time frames: (Y30-Y29) and (Y29-Y18).

The final set of predictors included 13 independent variables, which mapped the various dimensions of the response: number of errors, initiation time (IT), reaction time (RT), maximum deviation (MD), area under the curve (AUC), maximum deviation time (MD-time), x-flip, y-flip, Y30, Y29, Y18, Y30–Y29, and Y29–Y18. For each variable, we computed the average value of the 32 responses for each participant.

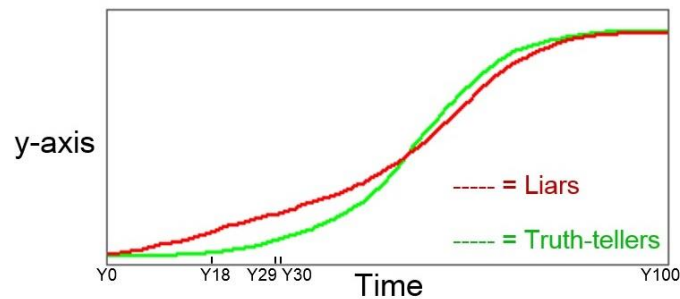


Figure 4.2: the figure reports the points of maximum difference between trajectories of truth-tellers and liars on y -axis over the time.

4.1.5 Analysis of trajectories

First, we visually compared the mouse trajectories of liars and truth-tellers. Figure 4.3 reports an example of the prototypical trajectory of a truth-teller and Figure 4.4 shows the prototypical trajectory of a liar. Each trajectory corresponds to a single question response. As can be noticed, the truth-teller's responses resulted in a more direct trajectory connecting the starting point with the correct response. By contrast, the liar showed trajectories that were more erratic. He initially deviated toward his default correct response and later changed his trajectory to press the false response button. Furthermore, the liar spent more time moving on the y -axis in the initial phase of the response than the truth-teller.

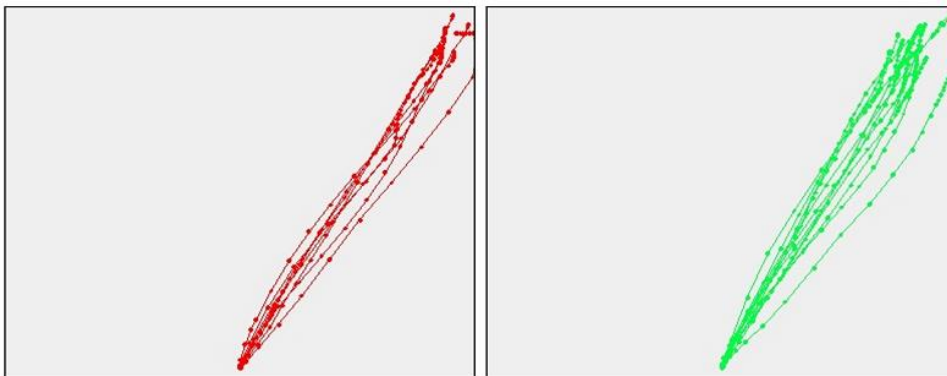


Figure 4.3: the prototypical trajectory of a truth-teller. In red the response trajectories to control questions and in green the trajectories of unexpected questions.

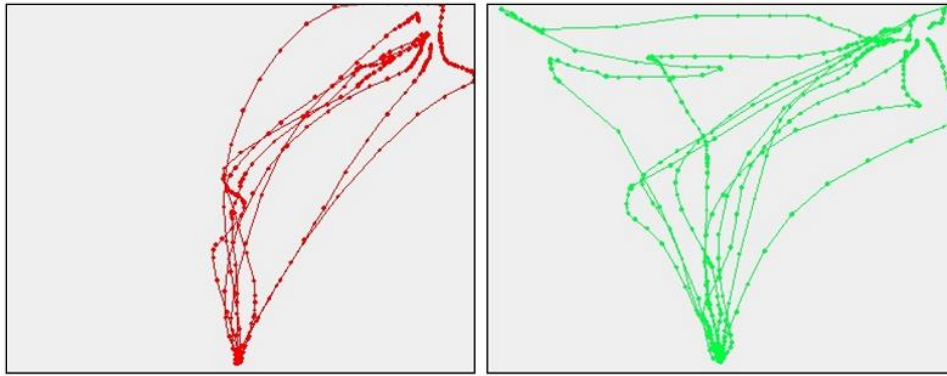


Figure 4.4: the prototypical trajectory of a liar. In red the response trajectories to control questions and in green the trajectories of unexpected questions. Note that this liar is responding truthfully to control questions. Nonetheless, his response diverges from the direct trajectory that ideally characterizes a truthful response (see Fig 4.3). This generalization of the liar mindset when the liar is responding to questions that require truthful responses is discussed in the following paragraphs.

These observations have been confirmed at the group level. Figure 4.5 reports the average trajectories for liars and truth-tellers responding “yes” to expected and unexpected questions (the only questions to which the liars responded deceitfully). As is clear from the figure, the two experimental groups differed in both the AUC and MD parameters.

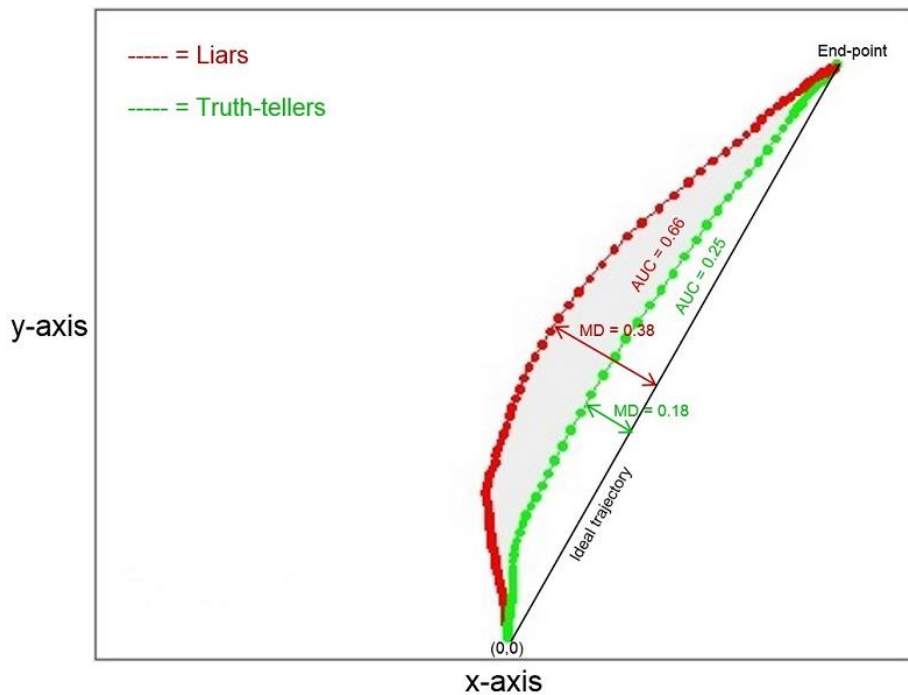


Figure 4.5: the figure represents the average trajectories between the subjects, respectively for liars (in red) and for truth-tellers (in green), to the expected “yes” and unexpected “yes” questions. Expected and unexpected questions that require a “yes” response are those to which the liars lied. The values of the MD and AUC for the two groups are reported. The grey area represents the difference in the AUC parameter between the liars and truth-tellers.

Finally, we plotted the mouse trajectories separately for control, expected, and unexpected questions (see Figure 4.6). Trajectory of liars and truth-tellers in control questions are almost overlapping. The maximum difference in trajectory is observed, again, in response to unexpected questions.

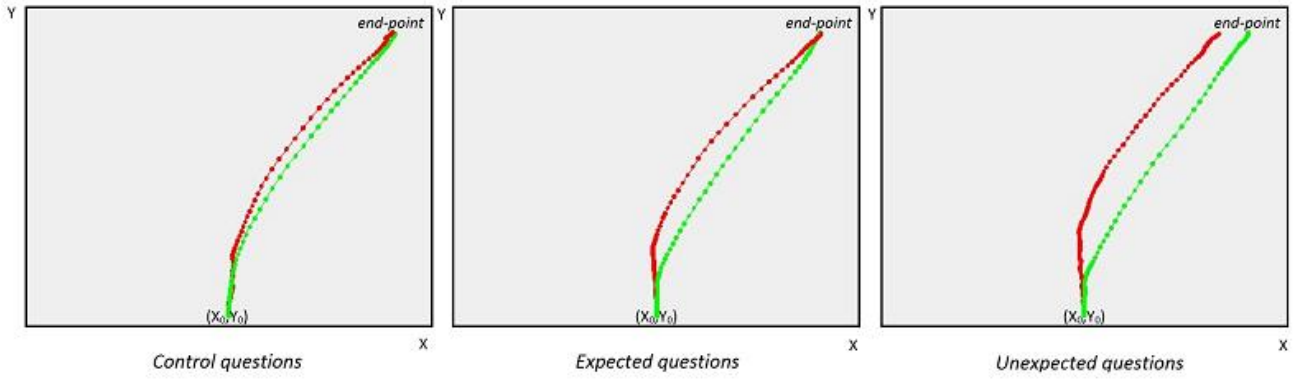


Figure 4.6: mouse trajectories for control questions (left figure), expected questions (central figure), and unexpected questions (right figure) comparing liars (red line) and truth-tellers (green line).

4.1.6 Feature selection

A feature selection was performed in order to remove redundant and irrelevant features and to select those that improved the models accuracy and generalization. In particular, as reported in paragraph 3.3.2, a correlation based feature selection was run to highlight the independent variables that had the maximum correlation with the dependent variable (truth-tellers vs. liars) and the minimum correlation across them.

From the 13 independent variables that were entered into the correlation analysis, the following features were selected: number of errors ($r_{pb} = 0.69$), AUC ($r_{pb} = 0.54$), MD-time ($r_{pb} = 0.46$), and Y29 ($r_{pb} = 0.43$). In Table 4.3, the correlation matrix between features is reported, as well as the correlation value between the dependent and independent variables (r_{pb}).

	Errors	AUC	MD-time	Y29	Condition
Errors	1.00	0.26	0.11	0.18	0.69
AUC	0.26	1.00	0.22	0.67	0.54
MD-time	0.11	0.22	1.00	0.39	0.46
Y-29	0.18	0.67	0.39	1.00	0.43
Condition	0.69	0.54	0.46	0.43	1.00

Table 4.3: the table reports the correlation matrix for the four features that were selected and their correlation value with the dependent variable.

4.1.7 Descriptive statistics

Feature selection isolated, from the original set of 13 predictors, four independent variables: errors, AUC, MD-time, and Y29. Table 4.4 reports the descriptive statistics for these features, as well as the analysis of the difference between truth-tellers and liars (t -test, Cohen’s d and Bayes Factor).

Seeing as number of errors is the most correlated feature with the dependent variable (liars vs truth-tellers), we investigated the errors distribution in more detail (see Table 4.5). Both liars and truth-tellers did not make errors to control questions and only 2/240 errors to expected questions. The difference between the two groups arises from unexpected questions, where truth-tellers made a total 5/240 errors and liars 82/240. In other words, in responding to unexpected questions the average liar makes 16 times the number of errors with respect to the average truth-teller. It is worth noting that liars make more errors to unexpected “yes” (60/120), which are the questions where they lie, rather

than unexpected “no” (22/120), where they respond truthfully ($t = -4.59, p < .01$; Cohen’s $d = 1.60$; $BF = 16.42$).

Feature	Group	M (SD)	t -test (t, p -value)	Cohen’s d	BF
Errors	Liars	0.13 (0.09)	5.83, < .01	1.84	> 150
	Truth-tellers	0.01 (0.02)			
AUC	Liars	0.60 (0.38)	3.09, < .01	1.23	70.41
	Truth-tellers	0.22 (0.21)			
MD-time	Liars	1264.33 (377.78)	3.19, < .01	1.01	13.52
	Truth-tellers	951.92 (219.62)			
Y29	Liars	0.27 (0.23)	2.93, < .01	0.92	7.66
	Truth-tellers	0.11 (0.08)			

Table 4.4: descriptive statistics for errors, AUC, MD-time and Y29. Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers). The last three columns report statistics about the difference between the two groups. In particular, the table shows the results of the independent t -test (t -value and p -value are reported), effect-size (Cohen’s d) and Bayes Factor (BF). For the interpretation of Cohen’s d and Bayes Factor, see paragraph 3.4.2.

Type of question		Liars (N = 20)	Truth-tellers (N = 20)
Control n = 320	Total number of errors / 160 stimuli	0 / 160	0 / 160
	Errors mean	0.000	0.000
	Errors SD	0.000	0.000
Expected n = 480	Total number of errors / 240 stimuli	2 / 240	2 / 240
	Errors mean	0.008	0.008
	Errors SD	0.091	0.091
Unexpected all n = 480	Total number of errors / 240 stimuli	82 / 240	5 / 240
	Errors mean	0.341	0.020
	Errors SD	0.475	0.143
Unexpected “yes” n = 240	Total number of errors / 120 stimuli	60 / 120	5 / 120
	Errors mean	0.500	0.042
	Errors SD	0.502	0.201
Unexpected “no” n = 240	Total number of errors / 120 stimuli	22 / 120	0 / 120
	Errors mean	0.183	0.000
	Errors SD	0.389	0.000

Table 4.5: number of errors in control, expected and unexpected questions that were committed by liars and truth-tellers. Errors in unexpected questions were also explored separately for “yes” and “no” questions.

Concerning the other three selected features, we investigated whether there is a difference between the questions to which subjects responded by moving the mouse to the right (questions requiring a “no” response) and questions to which subjects responded moving a mouse to the left (questions requiring a “yes” response). A t -test on the whole sample was carried out in order to compare left and right responses. Results showed that the trajectories in the two types of responses did not differ. In fact, we did not find any statistically significant difference both for MD-time ($t = 1.63; p = 0.1$; Cohen’s $d = 0.2$; $BF = 0.57$) and Y29 ($t = 0.1; p = 0.9$; Cohen’s $d = 0.01$; $BF = 0.17$). For AUC, we obtained the following results: $t = -2.09$ and $p = 0.04$, but the Cohen’s d value showed a small effect size ($d = -0.33$), and the Bayes Factor approached ($BF = 1.2$). In Figure 4.7, trajectories of the left and right responses are reported. It can be noted that the two curves follow a very similar, albeit specular, trajectory.

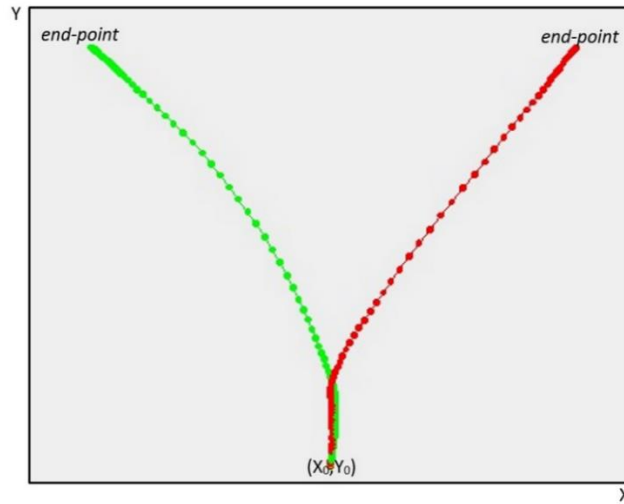


Figure 4.7: the figure reports the trajectories to the left response button (in green) and to the right response button (in red).

4.1.8 Machine learning models

The four selected features (errors, AUC, MD-time, Y29) were entered in different ML algorithms: logistic regression, SVM, LMT, random forest. Models were evaluated following a 10-fold cross-validation procedure, as described in paragraph 3.4.3. Results obtained by 10-fold cross-validation are reported in Table 4.6. All models reached an accuracy around 90% (36/40 subjects correctly classified) or higher in classifying subjects as liars and truth-tellers. Logistic classifier reached an accuracy of 95% (38/40 participants correctly classified).

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	95% (15.8)	0.950	0.950	95%	0.909	1.000
SVM	90% (17.5)	1.000	0.800	95%	1.000	0.900
Naïve Bayes	90% (17.5)	0.944	0.850	85%	0.769	1.000
Random forest	92.5% (12.1)	0.947	0.900	95%	0.909	1.000
LMT	92.5% (12.1)	1.000	0.850	95%	0.909	1.000

Table 4.6: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in training and test set. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

Then, we tested the generalization of the models performance on the new set of 20 participants (see section 4.1.1). Results, in Table 4.6, confirmed that all the models generalized very well on unseen data, reaching accuracies ranging from 85% to 95%.

About the rate of false positive and false negative, the confusion matrix showed that the number of liars and truth-tellers misclassified is not equal for all the algorithms. Logistic regression failed in detecting one liar and one truth-teller in the training set, whereas it misclassified a liar in the test set. The SVM was completely unbalanced towards the false negatives, misclassifying four liars in training set and one liar in test set. Naïve Bayes failed in classifying one truth-teller and three liars in cross-validation and three truth-tellers in the test set. Random forest misclassified two liars and one truth-

teller in training set and only one truth-teller in the test set. Finally, LMT failed in recognizing three liars in cross-validation and one truth-teller in the test set.

To have an insight on how the ML models distinguish between liars and truth-tellers we have run a tree model. It can help in the interpretation by providing an easy way to understand classification rules. The model used is CART and its output is reported in Figure 4.8. In words, the decision tree may be interpreted as follows:

- if the mean number of errors per subject is below 0.0469
and $AUC < 0.78$ then the subject is a truth-teller (subjects in this condition were 20/23 are classified correctly)
or $AUC > 0.78$ the subject is a liar (3 subjects are in this condition and all correctly classified)
- if errors are > 0.0469 the subject is a liar (14 participants fall in this leaf and all are classified correctly)

This tree model indicates that errors are the most important basis in classification and AUC (mouse trajectory) contributed to fine tune the classification.

```
error < 0.0469
| AUC < 0.7805500000000001: truthteller(20.0/3.0)
| AUC >= 0.7805500000000001: liar(3.0/0.0)
error >= 0.0469: liar(14.0/0.0)

Number of Leaf Nodes: 3

Size of the Tree: 5
```

Figure 4.8: output of the CART tree from the 10-fold cross-validation. The output describes the model on which is based the decision about the subject's classification as liar or truth-teller.

In order to study more in depth the relative weight of the predictors, we re-run the classifiers eliminating the four predictors one by one. Results (see Table 4.7) revealed that the major contribution in prediction accuracy comes from errors to unexpected questions with mouse dynamic features fine tuning an already good classification. In fact, eliminating the number of errors from the predictors, the classification accuracy dropped around to 75% in cross-validation and 70% in the test set. On the other hand, the prediction based solely on errors yielded an average accuracy of 82% in the cross-validation and around 99% in test set. After dropping AUC from the predictors, the classification accuracy remained stable in the test set and fell to 90% during cross-validation. Similar results were obtained removing MD-time from predictors (on average, 88.5% of accuracy in cross-validation and 91% in test set). Finally, after discharging Y29 from predictors, the accuracy both in the training (around 92.5%) and the test sets (around 95%) decreased slightly. Briefly, the relative importance of the independent variables indicated that the total number of errors gave the major contribution in correctly distinguishing liars from truth-tellers, followed by the MD-time, the AUC, and the position of the mouse along the y-axis on the 29th time frame.

Experiments

Predictors	ML classifier	Training set (10-fold cross-validation)			Test set		
		Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Errors	Logistic	82.5% (20.6)	0.933	0.700	100%	1.000	1.000
	SVM	80% (19.7)	1.000	0.600	95%	1.000	0.900
	Naïve Bayes	85% (21.1)	1.000	0.700	100%	1.000	1.000
	Random forest	77.5% (21.9)	0.789	0.750	100%	1.000	1.000
	LMT	85% (21.1)	1.000	0.700	100%	1.000	1.000
AUC, MD-time, Y29	Logistic	77.5% (21.9)	0.789	0.750	70%	0.667	0.800
	SVM	75% (20.4)	0.813	0.650	65%	0.667	0.600
	Naïve Bayes	75% (20.4)	0.778	0.700	55%	0.571	0.400
	Random forest	67.5% (31.3)	0.667	0.700	65%	0.615	0.800
	LMT	75% (20.4)	0.778	0.700	70%	0.700	0.700
Errors, MD-time, Y29	Logistic	95% (10.5)	0.950	0.950	95%	0.909	1.000
	SVM	85% (21.1)	1.000	0.700	95%	1.000	0.900
	Naïve Bayes	87.5% (21.2)	0.941	0.800	95%	0.909	1.000
	Random forest	90% (13.0)	0.900	0.900	95%	0.909	1.000
	LMT	90% (17.5)	1.000	0.800	100%	1.000	1.000
Errors, AUC, Y29	Logistic	90% (17.5)	0.944	0.850	95%	0.909	1.000
	SVM	87.5% (21.2)	1.000	0.750	85%	0.818	0.900
	Naïve Bayes	85% (24.1)	0.850	0.850	85%	0.769	1.000
	Random forest	90% (17.5)	0.944	0.850	95%	0.909	1.000
	LMT	90% (17.5)	0.944	0.850	95%	0.909	1.000
Errors, AUC, MD-time	Logistic	95% (11.0)	1.000	0.900	95%	0.909	1.000
	SVM	92.5% (12.0)	1.000	0.850	85%	0.818	0.900
	Naïve Bayes	92.5% (12.0)	0.947	0.900	90%	0.833	1.000
	Random forest	87.5% (17.7)	0.895	0.850	95%	0.909	1.000
	LMT	92.5% (12.0)	1.000	0.850	95%	0.909	1.000

Table 4.7: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in training and test set, using different set of predictors. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

Finally, we were interested investigating the contribution of control, expected, and unexpected questions in the classification. For this reason we run three separate models. Table 4.8 reports classification results obtained by training and testing the models separately for each type of question. It is confirmed that it is not possible to efficiently distinguish liars from truth-tellers solely based on control questions. The same is true also for expected questions although, in this case, the trajectories of the two groups seem to be more separated (see Figure 4.6). The major contribution derives from unexpected questions. In fact, using only unexpected questions, classification accuracy reaches 90%, both in cross-validation and test set. This result confirms that the cognitive load of liars, due to unexpected questions, is at the origin of the difference between the two groups.

Type of question	ML classifier	Training set (10-fold cross-validation)			Test set		
		Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Control	Logistic	60% (31.6)	0.611	0.550	60%	0.667	0.400
	SVM	55% (25.8)	0.667	0.200	55%	0.600	0.300
	Naïve Bayes	62.5% (31.7)	0.727	0.400	55%	0.600	0.300
	Random forest	57.5% (20.6)	0.579	0.550	50%	0.500	0.700
	LMT	50% (31.2)	0.500	0.500	60%	0.625	0.500
Expected	Logistic	67.5% (29.0)	0.684	0.650	50%	0.500	0.500
	SVM	65% (21.1)	0.875	0.350	55%	0.571	0.400
	Naïve Bayes	60% (24.2)	0.667	0.400	65%	0.667	0.600
	Random forest	62.5% (35.9)	0.632	0.600	55%	0.545	0.600
	LMT	65% (26.9)	0.688	0.550	55%	0.556	0.500
Unexpected	Logistic	95% (10.6)	0.909	1.000	95%	0.909	1.000
	SVM	92.5% (12.1)	1.000	0.850	90%	0.900	0.900
	Naïve Bayes	92.5% (12.1)	0.947	0.900	85%	0.818	0.900
	Random forest	87.5% (17.7)	0.875	0.900	90%	0.900	0.900
	LMT	95% (10.6)	1.000	0.900	90%	0.900	0.900

Table 4.8: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in training and test set, separately for control, expected and unexpected questions. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.1.9 Can we detect liars also when they respond truthfully?

As mentioned above, liars lied only when they responded “yes” to expected and unexpected questions. In all the other questions (expected “no”, unexpected “no”, control “yes”, control “no”), they responded truthfully (see paragraph 4.1.3). An interesting question is whether the liars could also be spotted from their truthful responses. In section 4.1.5, we compared the trajectories of the two groups to expected and unexpected questions that required a “yes” response (see Figure 4.5).

In Figure 4.9 we have compared the trajectories of the two groups when they responded truthfully (expected and unexpected questions that required a “no” response and control questions). Although the difference between liars and truth-tellers is reduced in comparison with the trajectories of questions where the liars were lying (Figure 4.5), a difference in MD and AUC is still detectable.

In order to evaluate whether this difference is statistically significant, we have run an independent t -test on the four predictors, showing that the liars’ response styles may be identified even when they responded truthfully. Results are reported in Table 4.9.

The accuracy rates in identifying liars and truth-tellers on the sole basis of responses to questions to which the liars responded truthfully are reported in Table 4.10. It should be noticed that using expected “no”, unexpected “no” and control questions, accuracy is around 80% in cross-validation and around 75% in the test set, whereas classification accuracy based only on “yes” responses to expected and unexpected questions ranges from 85% to 95% both in training and test set.

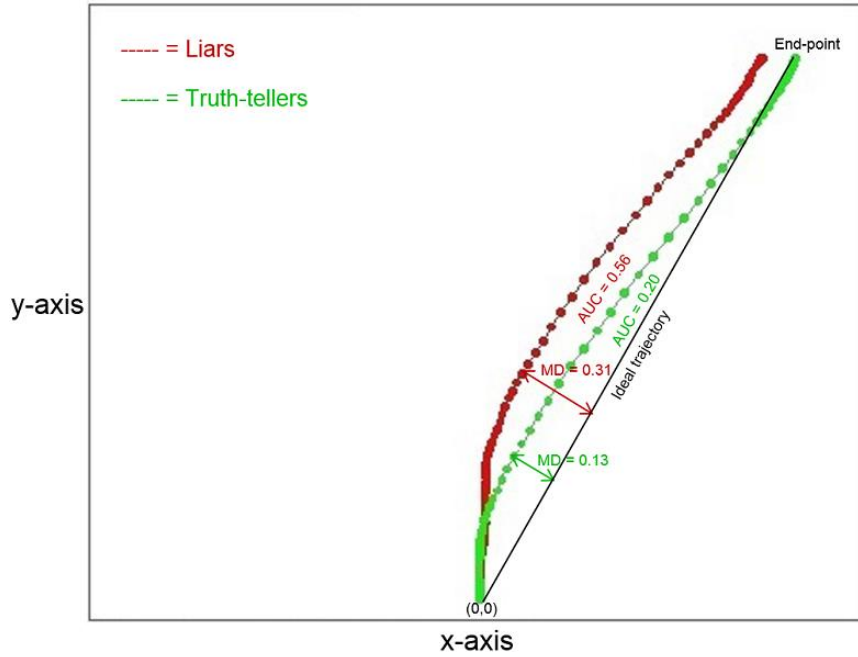


Figure 4.9: the figure represents the average trajectories between the subjects, respectively for liars (in red) and for truth-tellers (in green), to the expected “no”, unexpected “no” and control questions. Expected “no”, unexpected “no” and control questions are those to which the liars responded truthfully. The values of the MD and AUC for the two groups are reported. The grey area represents the difference in the AUC parameter between the liars and truth-tellers.

Feature	Expected and unexpected “yes” questions	Expected and unexpected “no” questions and control questions
Errors	$t = 6.06, p < 0.01, d = 1.91, BF > 150$	$t = 3.44, p < 0.01, d = 1.09, BF = 23.11$
AUC	$t = 3.46, p < 0.01, d = 1.09, BF = 24.46$	$t = 3.36, p < 0.01, d = 1.06, BF = 19.63$
MD-time	$t = 3.42, p < 0.01, d = 1.08, BF = 22.03$	$t = 2.65, p < 0.02, d = 0.83, BF = 4.37$
Y29	$t = 2.63, p < 0.02, d = 0.83, BF = 4.26$	$t = 2.98, p < 0.01, d = 0.94, BF = 8.51$

Table 4.9: comparison between liars and truth-tellers in questions where liars responded truthfully (third column) and questions where liars responded lying (second column). In particular, the table shows the results of the independent t -test (t -value and p -value are reported), effect-size (Cohen’s d) and Bayes Factor (BF) for errors, AUC, MD-time and Y29. For the interpretation of Cohen’s d and Bayes Factor (see paragraph 3.4.2)

Predictors	ML classifier	Training set (10-fold cross-validation)			Test set		
		Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Expected and unexpected “yes” questions	Logistic	92.5% (12.1)	0.947	0.900	85%	0.769	1.000
	SVM	90% (17.5)	1.000	0.800	95%	1.000	0.900
	Naïve Bayes	87.5% (13.2)	0.895	0.850	85%	0.818	0.900
	Random forest	85% (12.9)	0.889	0.800	95%	0.909	1.000
Expected and unexpected “no” and control questions	LMT	92.5% (12.1)	0.947	0.900	85%	0.769	1.000
	Logistic	80% (19.8)	0.800	0.800	75%	0.629	0.900
	SVM	80% (23.0)	0.929	0.650	70%	0.833	0.500
	Naïve Bayes	80% (23.0)	0.833	0.750	75%	0.692	0.900
and control questions	Random forest	70% (25.9)	0.722	0.650	75%	0.692	0.900
	LMT	77.5% (18.4)	0.824	0.700	75%	0.778	0.700

Table 4.10: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in training and test set, separately for questions to which liars responded truthfully and questions to which they lie. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is reported as well. Recall is the percentage of liars who are correctly identified and precision.

To sum up, both statistical analysis and ML analysis have shown that liars could be identified even when they are responding truthfully, but with lower accuracy. What is interesting here from a cognitive point of view is that, in the experimental design, the mind-set of the liars extended its effects to questions to which they responded truthfully. To our knowledge, this pattern of results has never been reported before and could be an indication of the level of sensitivity of the mouse-movement analysis.

4.1.10 Generalization to different cultures

To address the effects of culture on the results generalization, we tested a new sample of 20 German native speakers. Participants were recruited at University of Düsseldorf. The sample included 10 truth-tellers and 10 liars. Demographic information about participants are in Table 4.11. The experimental procedure was identical to that followed for the Italian sample (see section 4.1.2), but the ID card on which the false identity was presented was a German ID Card. Moreover, stimuli (see section 4.1.3) have been adapted to German culture (e.g., Italian regions have been replaced with German Länder) and translated in German language.

Sample	N	Gender	Age	Education
Training set (40 Italian participants)	40	M = 17, F= 23	M = 25, SD = 4.6	M = 17, SD = 1.8
Test set (20 German participants)	20	M = 9, F= 11	M = 29, SD = 8.9	M = 16, SD = 2.2

Table 4.11: demographic information about training set and test set. Training set included 40 Italian participants, whereas 20 German subjects compose the test set. In the second column (N) the number of participants for each sample is reported. The third column shows the number of male and female in each sample. The fourth and the fifth columns report mean (M) and standard deviation (SD) of participants' age and education.

Results from this group were evaluated using the models originally trained on the 40 Italian participants. Classification accuracies are reported in Table 4.12. It is remarkable that accuracies are very similar to those obtained from the Italian test set (see Table 4.6), confirming that the models can efficiently classify participants from different cultures.

ML classifier	Training set (10-fold cross-validation) 40 Italian participants			Test set 20 German participants		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	95% (15.8)	0.950	0.950	100%	1.000	1.000
SVM	90% (17.5)	1.000	0.800	90%	0.900	0.900
Naïve Bayes	90% (17.5)	0.944	0.850	85%	0.769	1.000
Random forest	92.5% (12.1)	0.947	0.900	95%	0.909	1.000
LMT	92.5% (12.1)	1.000	0.850	95%	0.909	1.000

Table 4.12: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in a training set of 40 Italian participants and in a test set of 20 German subjects. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

An analysis of the errors indicated that Italian and German participants made an equal number of errors, with statistical results for liars of $t = -1.4$, $p = 0.17$ (Cohen's $d = -0.49$, $BF = 0.64$) and statistical results for truth-tellers of $t = 0.66$, $p = 0.52$ (Cohen's $d = 0.28$, $BF = 0.43$). Table 4.13 reports the proportion of errors for the training sample (40 Italian participants) and the test set (20 German participants).

Sample	Condition	Average number of errors	Total number of errors
Training set	Liars	0.13 / 32	84 / 640
(40 Italian participants)	Truth-tellers	0.01 / 32	7 / 640
Test set	Liars	0.17 / 32	5 / 320
(20 German participants)	Truth-tellers	0.006 / 32	2 / 320

Table 4.13: number of errors for the training sample of 40 Italian participants and for the test sample of 20 German participants. The average number of errors and the total number of errors are reported, respectively in the third and fourth column.

4.1.11 The resistance to countermeasures

Resistance to countermeasures is a central issue in lie detection for all currently available techniques. In fact, if the subject is aware about the indices that are measured by the lie detector, he could apply some strategies to beat it [178]. For example, simple countermeasures, such as silently counting backward or pinching a finger, are effective in altering the results in detection of deception via fRMI [179]. In the RT-based techniques, such as CIT and aIAT, the intentional alteration of RT is enough to drop their efficiency [180].

In this paragraph, we have tested the resistance to countermeasure of the paradigm reported above, which combines unexpected questions and mouse dynamics. Our hypothesis was that this technique is promising also as regards resistance to countermeasures, for the following reasons:

- Errors to unexpected questions are diagnostic of lying and the subjects should respond errorless in order to cheat the test, and this seems impossible.
- The parameters used to encode mouse dynamics are high in number. It is unlikely that the responder succeeds in implementing countermeasures that simultaneously keep under voluntary control all the possible efficient predictors.
- There are a number of different set of predictors, which are roughly equivalent in terms of classification accuracy. Again, it is unlikely that the faker may keep under control voluntarily all these set of parameters.

We firstly observed that there are no possible countermeasures to the number of errors. In fact, the only way to avoid errors is to know all the questions in advance. Countermeasures, therefore, are limited to the mouse dynamics. As the number of mouse movement parameters is very high, it is possible to create different subset of predictors that efficiently classify the participants. To highlight this point, we tested an alternative classification model, which is based on a new set of predictors: errors, MD, RT, y-flip and (Y29-18). It should be remembered that the predictors entered in the original model included errors, AUC, MD-time and Y29. The correlation matrix of the new predictors and their correlation with the dependent variable are reported in table 4.14.

	Errors	RT	MD	y-flip	Y29-Y18	Condition
Errors	1.00	0.02	0.18	0.20	0.27	0.69
RT	0.02	1.00	0.38	0.31	0.22	0.40
MD	0.18	0.38	1.00	0.38	0.83	0.46
y-flip	0.20	0.31	0.38	1.00	0.30	0.31
Y29-Y18	0.27	0.22	0.83	0.30	1.00	0.41
Condition	0.69	0.40	0.46	0.31	0.41	1.00

Table 4.14: the table reports the correlation matrix for the four new features plus errors and their correlation value with the dependent variable.

Results from ML models indicate that the new set of predictors efficiently classify participants. Classification accuracies are reported in Table 4.15.

ML classifier	Training set	Test set				
	(10-fold cross-validation)	Precision	Recall	Accuracy	Precision	Recall
	Average accuracy (SD)					
Logistic	90% (12.9)	0.900	0.900	85%	0.769	1.000
SVM	90% (17.5)	1.000	0.800	90%	1.000	0.800
Naive Bayes	87.5% (24.3)	0.895	0.850	95%	0.909	1.000
Random forest	87.5% (17.7)	0.895	0.850	95%	0.909	1.000
LMT	92.5% (12.1)	0.947	0.900	100%	1.000	1.000

Table 4.15: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in training and test set using an alternative set of predictors. The features included in the models were errors, MD, RT, y-flip and (Y29-18). The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

To verify whether it is possible to keep under control the mouse dynamics during the task, and in order to evaluate the resistance of the method to countermeasures, we ran a further experiment instructing participants to cheat the test. We identified three different strategies that can be apply to beat the lie detector, according to the original set of predictors entered in ML models (errors, AUC, MD-time and Y29) and that were used to classify subjects as liars or truth-tellers. The countermeasure strategies that we identified were the following:

- **Velocity strategy:** it consisted in responding to questions as quick as possible, to avoid to be identify as liar on the basis of the MD-time and Y29 parameters.
- **Straightness strategy:** it consisted in thinking to the answer and then, moving the mouse to the response button as straight as possible. This strategy was thought to prevent the identification as liar based on the AUC parameter.
- **Slowness strategy:** this strategy consisted in slowing down the response time to easy questions (expected and control questions) to harmonies the response times of control and expected questions to those of unexpected questions. Once again, having more time to respond to critical questions, the mouse movement should result in a more direct trajectory connecting the starting point with the correct response.

Sixty coached liars (called also fakers) took part to the experiment, twenty for each countermeasure strategy. To balance the number truth-tellers participants with the number of fakers, we also collected thirty additional truth-tellers. The demographic features of the final sample are reported in Table 4.16.

Sample	N	Age	Education
Fakers velocity strategy	20	M = 22, SD = 1.6	M = 16, SD = 1.4
Fakers straightness strategy	20	M = 23 SD = 2.1	M = 17, SD = 2.1
Fakers slowness strategy	20	M = 22 SD = 2.2	M = 16, SD = 1.6
Truth-tellers	60	M = 23 SD = 3.1	M = 16, SD = 1.7

Table 4.16: demographic information about training set and test set. Training set included 40 Italian participants, whereas 20 German subjects compose the test set. In the second column (N) the number of participants for each sample is reported. The third column and the fourth columns report mean (M) and standard deviation (SD) of participants' age and education.

The experimental procedure was the same described in paragraph 4.1.2, except for the instruction given to fakers participants in order to cheat the lie detector. Firstly, all faker participants were instructed to reduce the number of errors as much as possible and were advised about the possibility of unexpected questions. For this reason, they were given five minutes to think to possible unanticipated questions and were permitted to search information on internet. Then, they were explicitly informed about the classification logic of the lie detector and instructed to beat it with one of the three strategies mentioned above.

An analysis was run training the ML models on the sixty fakers and the sixty truth-tellers. Predictors were the same of the original experiment (errors, AUC, MD-time, Y29). Classification accuracies are reported in Table 4.17. Results indicate that liars using a variety of countermeasures are identified with an accuracy of about 80% by the same classifiers used in the original experiment.

ML classifier	10-fold cross-validation		
	Average accuracy (SD)	Precision	Recall
Logistic	80.8% (14.2)	0.825	0.783
SVM	82.5% (8.3)	0.855	0.783
Naïve Bayes	75.8% (11.4)	0.804	0.683
Random forest	82.5% (13.3)	0.810	0.850
LMT	78.3% (14.8)	0.804	0.750

Table 4.17: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation. The accuracy obtained by the 10-fold cross-validation is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

To investigate which is the most efficient countermeasure, we re-run the classifications separately for the three strategies. Thus, we trained the classifiers on three different sample as follows: 20 fakers velocity strategy vs 20 truth-tellers, 20 fakers straightness strategy vs 20 truth-tellers, 20 fakers slowness strategy vs truth-tellers. Truth-tellers were always the same participants, that is the twenty participants used as training set in the original experiment (see section 4.1.1). Classification accuracies for the three samples are reported in Table 4.18.

Sample	ML classifier	10-fold cross-validation		
		Average accuracy (SD)	Precision	Recall
Fakers velocity strategy vs truth-tellers	Logistic	72.5% (21.8)	0.696	0.800
	SVM	72.5% (24.9)	0.765	0.650
	Naïve Bayes	77.5% (14.2)	0.824	0.700
	Random forest	77.5% (21.8)	0.789	0.750
	LMT	75% (11.8)	0.727	0.800
Fakers straightness strategy vs truth-tellers	Logistic	80% (15.8)	0.800	0.800
	SVM	75% (23.6)	0.708	0.850
	Naïve Bayes	70% (15.9)	0.643	0.900
	Random forest	82.5% (12.1)	0.810	0.850
	LMT	77.5% (21.8)	0.789	0.750
Fakers slowness strategy vs truth-tellers	Logistic	80% (19.7)	0.833	0.750
	SVM	80% (19.8)	0.833	0.750
	Naïve Bayes	80% (15.9)	0.833	0.750
	Random forest	87.5% (13.2)	0.895	0.850
	LMT	82.5% (16.9)	0.842	0.800

Table 4.18: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation, separately for the three different faking strategies. The accuracy obtained by the 10-fold cross-validation is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction

It can be noticed that the weaker faking strategy is slowing down the response time to easy questions. In fact, these fakes are detect with an accuracy ranging from 80% to 87.5%. Conversely, the most efficient countermeasure is to increase response speed. In fact, fakers using the velocity strategy are detectable with an accuracy ranging from 72.5% to 77.5%. Intermediate results have been obtained for the straightness strategy (moving the mouse to the response button as straight as possible), as these fakers are correctly classified with an accuracy ranging from 70% to 82.5%.

4.1.12 The use of changing labels

In the previous experiment, participants were asked to respond questions about identity choosing between two response labels, one containing the response “yes” and the other one containing the response “no”. A third experiment was run to investigate the effect of changing these two response labels.

Forty participants took part to the experiment, 20 male and 20 female, with average age = 24 (SD = 5.4) and average education = 17 (SD = 1.7).

The experimental procedure was identical to that describe above (see section 4.1.2).

The stimuli presented to subjects were the same of the previous experiment (see paragraph 4.1.3), but questions that required a yes/no response were intermixed with questions requiring to choose between two different alternative responses (e.g., question: how old are you?, response labels: 27 or 31). We refer to this kind of stimuli as questions with changing response labels.

The experimental task consisted of 100 double-choice questions, 20 control questions, 32 expected questions and 48 unexpected questions. Between them, the half of the questions requested a “yes” or

“no” response, while the other required a response according to changing response labels (e.g., to the question “Which is your gender?”, the possible response labels were “male” or “female”). Within the entire task, the correct responses, which are the answers congruent with the participant ID card, were presented on right position in the 50% of trials and on the left in the other 50%. Some examples of the 100 questions included in the experimental task are reported in Table 4.19.

Type of question	Question	Correct response	Incorrect response
Expected	Were you born in April?	Yes	No
	Were you born in October?	No	Yes
	What is your year of birth?	1987	1984
	What is your city of birth?	Verona	Milano
Unexpected	Is your residence city near Abano Terme?	Yes	No
	Is your residence city near Saturnia Terme?	No	Yes
	Which is your zodiac sign?	Aries	Capricorn
	What is your zip code?	35142	36125
Control	Are you female?	Yes	No
	Are you male?	No	Yes
	How tall are you?	160 cm	190 cm
	What is your skin color?	White	Black

Table 4.19: the table reports some example of the expected, unexpected and control questions presented to participants and related to a truth or faked identity. The third and fourth columns report respectively the correct and the incorrect response to the question. It should be noticed that half of responses were in form of yes/no, and the other half had changing response labels.

The features collected were those extracted by default by the MouseTracker software (IT, RT, MD, AUC, MD-time, x-flip, y-flip, see paragraph 3.3.2). In addition, we computed the average velocity (v) and acceleration (a) along the x and y -axis ($v_{(x)}$, $v_{(y)}$, $a_{(x)}$, $a_{(y)}$).

The Figure 4.10 represents the mouse trajectories of liars and truth-tellers for control, expected and unexpected questions. The trajectories of the two groups are totally superimposed in control questions and mostly overlapping in expected questions. Thus, liars and truth-tellers seems differ only in unexpected questions.

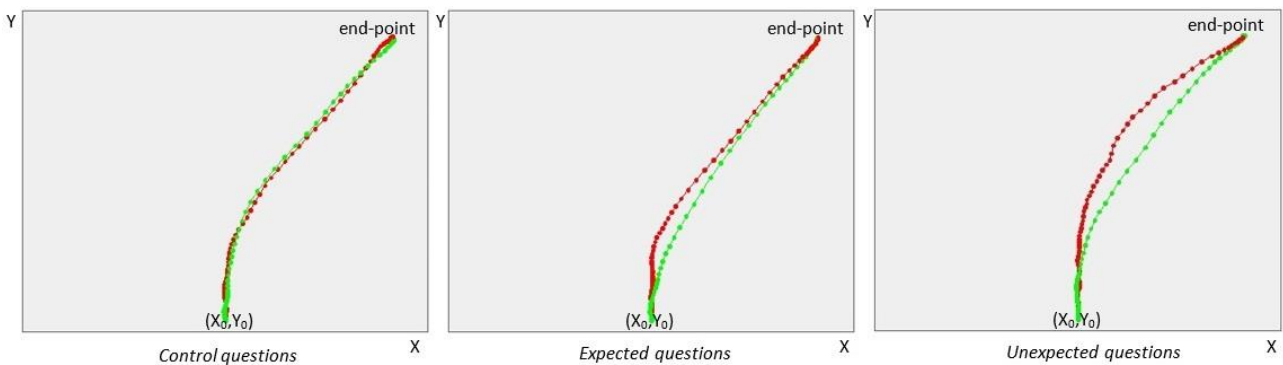


Figure 4.10: mouse trajectories for control questions (left figure), expected questions (central figure), and unexpected questions (right figure) comparing liars (red line) and truth-tellers (green line).

In order to confirm whether the difference between liars and truth-tellers trajectories in unexpected questions is statistically significant, we run an independent t -test. Results showed that liars’ responses significantly differ from truth-tellers’ ones in AUC ($t = 3.13$, $p < .0042$), RT ($t = 3.61$, $p < 0.0042$), and average velocity along x -axis ($t = -7.62$, $p < 0.0042$). Finally, liars make a higher number of errors

compared to truth-tellers ($t = 9.70, p < 0.0042$) (to avoid the multiple testing problem the correction of Bonferroni has been apply and the p -value has been set to 0.0042). Finally, we tested the difference between liars and truth-tellers also for expected and control questions, confirming that none of the measures considered reached the statistical significance in the independent t -test.

According to graphical observations and statistical analysis, we used only unexpected questions data to train different machine learning classifiers. The CFS algorithm selected the following features: number of errors ($r_{pb} = 0.84$) and $v_{(x)}$ ($r_{pb} = 0.78$). Given the limited number of features, to increase the number of predictors and thus, the resistance to countermeasure, we chose to enter in ML model RT ($r_{pb} = 0.51$) and AUC ($r_{pb} = 0.45$), MD ($r_{pb} = 0.41$) and MD-time ($r_{pb} = 0.40$) as well.

Machine learning models, trained with a 10-fold cross-validation, gave results comparable to those of the previous experiment (see Table 4.6), with accuracies ranging from 87.5% to 95%. Therefore, in the first instance, it seems that changing labels do not contribute to improve the classification accuracy.

To dig deeper this point, we re-run the ML models considering only questions with changing response labels. These new results are reported in Table 4.20. It should be noticed that the classification accuracy improved from 2.5% to 12.5% when we considered only questions requiring a response with changing labels.

ML classifier	10-fold cross-validation considering all questions			10-fold cross-validation considering only questions with changing response labels		
	Accuracy (SD)	Precision	Recall	Accuracy (SD)	Precision	Recall
Logistic	87.5% (13.2)	0.895	0.850	92.5% (12.1)	0.947	0.900
SVM	95% (10.5)	1.000	0.900	97.5% (7.9)	1.000	0.950
Naïve Bayes	95% (10.5)	0.950	0.950	97.5% (7.9)	0.952	1.000
Random forest	87.5% (17.7)	0.941	0.800	100% (0.00)	1.000	1.000
LMT	90% (12.9)	0.944	0.850	100% (0.00)	1.000	1.000

Table 4.20: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in 10-fold cross-validation. The second column reports the accuracy obtained considering both questions requiring a yes/no response and questions with changing response labels, whereas the third column reports the accuracy using only questions with changing response labels. It should be remembered that we entered in the models only the unexpected questions. The 10-fold cross-validation accuracy is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

In other words, we reached a better classification performance in distinguishing liars from truth-tellers when participants respond to questions according to changing response labels. Our hypothesis is that the continuous change of the response categories results in an increment of liars’ cognitive load. In fact, it is possible that using only “yes” or “no” fixed labels, after some trials the label processing becomes partially automated and does not require any mental effort. Conversely, in a task where label change away, the true label is familiar to truth-tellers, whereas both the true and the false labels are unfamiliar to liars, especially in the case of unexpected questions. For this reason, liars may require more cognitive effort to process labels and implement the correct response. As consequence, they have a poorer performance and the discrimination between the two experimental groups becomes more accurate.

4.1.13 The use of negative questions

Another issue in lie detection regards the need, in some situations, to introduce negative sentences to test the suspect. For example, imagine that one wants to test deception about recent drug use. There are only two way to formulate the question: “did you take drugs?” or “did you not take drugs?”. Recent experiments demonstrated that the RT-based based lie detectors (e.g., the aIAT), lose their accuracy when using stimuli in negative form [181]. For example, it has been shown that the use of negative sentences leads to a detrimental effects in the accuracy of the autobiographical Implicit Association Test (aIAT), which drops from 90% to 60% [182].

A large number of linguistic studies demonstrated that a negative sentence have a more complex syntax structure than an affirmative one [183] and, as consequence the human brain activate different area and take more time to process it [184]. Given this evidence, here we propose an experiment aimed to investigate the effect of negation on the detection of faked identities using complex questions and mouse dynamics.

Forty participants were recruited, 28 female and 12 male, with average age = 23 (SD = 1.8) and average education level = 17 (SD = 1.5).

The experimental procedure was identical to that describe above (see section 4.1.2).

The stimuli presented to subjects were similar to those in the original experiment (see paragraph 4.1.3), but affirmative sentences were intermixed with negative sentences.

The experimental task consisted of 72 yes/no questions, 36 affirmative and 36 negative. Between these 36 questions, eight were control questions, 14 expected and 14 unexpected. Half of the 72 questions requested a “yes” response and the other half a “no” response. An example of questions for each typology is reported in Table 4.21.

Type of question	Sentence form	Question	Required response	Truth-tellers	Liars
Expected	Affirmative	I was born in April	Yes	True	Lie
	Negative	I was not born in April	No	True	Lie
	Affirmative	I was born in October	No	True	True
	Negative	I was not born in October	Yes	True	True
Unexpected	Affirmative	34074 is my zip code	Yes	True	Lie
	Negative	34074 is not my zip code	No	True	Lie
	Affirmative	6893 is my zip code	No	True	True
	Negative	6893 is not my zip code	Yes	True	True
Control	Affirmative	I am female	Yes	True	True
	Negative	I am not female	No	True	True
	Affirmative	I am male	No	True	True
	Negative	I am not female	Yes	True	True

Table 4.21: the table reports some example of the expected, unexpected and control questions, affirmative and negative, presented to participants and related to a truth or faked identity. The fourth column reports the required response according to the true or false ID card. It should be noticed that the required response for liars and truth-tellers is the same. The sixth column contain the information about where liars effectively lied and where they told he truth. Truth-tellers told the truth in all questions (fifth column).

It should be noted that liars told lies only in the expected and unexpected questions, both in form of affirmation and in form of negation, that correspond to the data in the ID card. Table 4.21 reports a simplification where liars lied and where they told the truth.

During the task, the features collected were those extracted by default by the MouseTracker software (IT, RT, MD, AUC, MD-time, x-flip, y-flip, see paragraph 3.3.2). In addition, we computed the maximum velocity and acceleration along the x and y -axis ($\max v_{(x)}$, $\max v_{(y)}$, $\max a_{(x)}$, $\max a_{(y)}$).

The Figure 4.11 represents the mouse trajectories of liars and truth-tellers for control, expected and unexpected questions. Visually, the trajectories of the two groups differ in terms of MD and AUC in all type of questions (control, expected, unexpected). This could mean that for liars the presence of negations affects the complexity of the entire task.

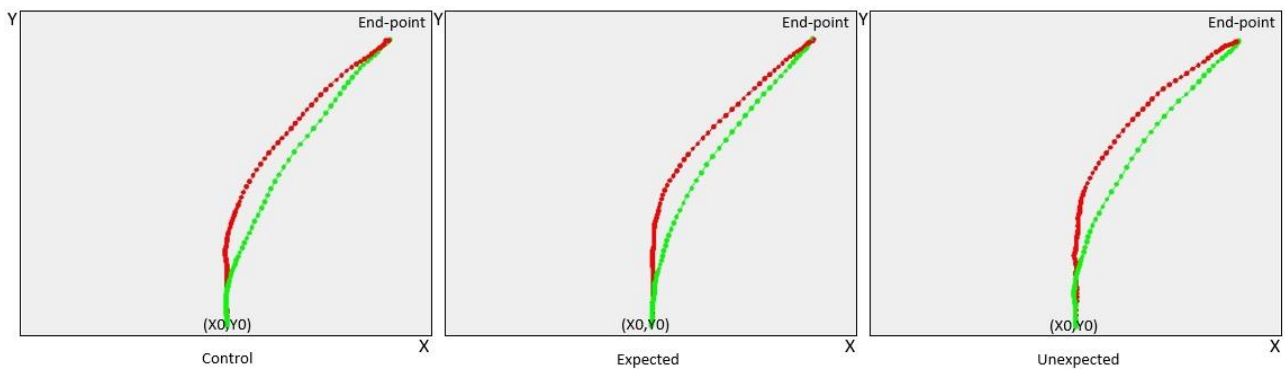


Figure 4.11: mouse trajectories for control questions (left figure), expected questions (central figure), and unexpected questions (right figure) comparing liars (red line) and truth-tellers (green line).

To deepen the effect of negative sentence, we plotted separately the trajectories related the responses to affirmative questions and to negative questions, respectively for liars and truth-tellers (see Figure 4.12). The figures show that both liars and truth-tellers had larger trajectories in response to negative questions respect to affirmative questions. In other words, the visual analysis confirm the evidence in literature that negations are more difficult to process than affirmations.

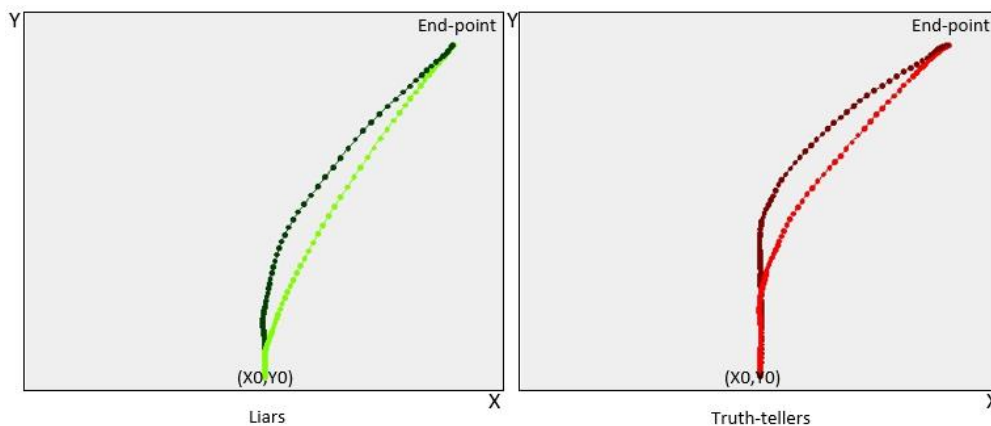


Figure 4.12: the left figure reports the average mouse trajectory of truth-tellers in response to positive questions (light green curve) and negative questions (dark green line). The right figure represents the average mouse trajectory of liars in response to positive questions (light red line) and negative questions (dark red curve).

Statistical analysis have corroborated this result as well. In fact, a factorial mixed ANOVA indicated that:

- Liars made a greater number of errors than truth-tellers [$F(1,38)=41.91$, $p<0.01$]. Overall, both liars and truth-tellers made more errors in unexpected questions than expected and control questions [$F(2,76)=85.99$, $p<0.01$], and in negative questions than affirmative [$F(1,38)=35.57$, $p<0.01$]. The interactions condition (liars vs truth-tellers) X type of question (control, expected, unexpected) showed statistically significant results [$F(2,76)=49.74$, $p<0.01$], as well as the interaction type of question X questions formulation (affirmation vs negation) [$F(2,76)=6.17$, $p<0.01$].
- Liars had greater MD and AUC than truth-tellers, respectively [$F(1,38)=7.71$, $p<0.01$] for MD and [$F(1,38)=5.84$, $p<0.01$] for AUC. Both liars and truth-tellers had larger MD and AUC in unexpected questions than expected and control questions, respectively [$F(2,76)=5.65$, $p<0.01$] for MD and [$F(2,76)=5.44$, $p<0.01$] for AUC. Moreover, both liars and truth-tellers showed wider MD and AUC in negative questions than affirmative questions, respectively [$F(1,38)=74.37$, $p<0.01$] for MD and [$F(1,38)=58.86$, $p<0.01$] for AUC.
- Liars had slower IT, RT and MD-time than truth-tellers, respectively [$F(1,38)=6.68$, $p<0.01$] for IT, [$F(1,38)=8.86$, $p<0.01$] for RT and [$F(1,38)=5.12$, $p<0.01$] for MD-time. Both liars and truth-tellers had increased IT, RT and MD-time in unexpected questions compared to expected and control questions, respectively [$F(2,76)=11.83$, $p<0.01$] for IT, [$F(2,76)=61.42$, $p<0.01$] for RT and [$F(2,76)=56.82$, $p<0.01$] for MD-time. Moreover, both liars and truth-tellers showed slower IT, RT and MD-time in negative questions than affirmative questions, respectively [$F(1,38)=18.71$, $p<0.01$] for IT, [$F(1,38)=35.13$, $p<0.01$] for RT, and [$F(1,38)=59.46$, $p<0.01$] for MD-time. Moreover, in RT and MD-time, the interactions condition (liars vs truth-tellers) X type of question (control, expected, unexpected) showed statistically significant results, respectively [$F(2,76)=21.36$, $p<0.01$] for RT and [$F(2,76)=15.06$, $p<0.01$] for MD-time.
- Finally, liars and truth-tellers differ for the number of x-flip and y-flip, respectively [$F(1,38)=7.88$, $p<0.01$] for x-flip and [$F(1,38)=11.27$, $p<0.01$] for y-flip.

To sum up, the effect of negation is obvious for both liars and truth-tellers. Moreover, concerning the number of errors liars made more errors than truth-tellers in negative sentences.

Last but not least question that we wanted to answer with this experiment is whether the introduction of negation in the task stimuli drops the classification accuracy.

Considering the responses both to affirmative and negative questions, the CFS algorithm selected the following features: number of errors ($r_{pb} = 0.74$), max $a_{(x)}$ ($r_{pb} = 0.38$), max $a_{(y)}$ ($r_{pb} = 0.46$) and x-flip ($r_{pb} = 0.41$). Taking onto account only the response to negative questions, the CFS algorithm selected the following features: number of errors ($r_{pb} = 0.65$), max $a_{(x)}$ ($r_{pb} = 0.30$), max $a_{(y)}$ ($r_{pb} = 0.38$) and RT ($r_{pb} = 0.39$).

Machine learning models, trained with a 10-fold cross-validation, gave results ranging from 77.5% to 87.5% of accuracy considering responses both to affirmative and negative questions, whereas the accuracies decreased (they ranged from 75% to 82.5%) entering in the classifiers only responses to

negative questions (see Table 4.22). To conclude, the introduction of negation influences the accuracy of the lie detector, which drops by 7.5% to 12.5%. In fact, the classification accuracies in cross-validation in the original experiment ranged from 90% to 95% (see section 4.1.8), whereas in the present experiment accuracies range from 77.5% to 87.5%.

ML classifier	10-fold cross-validation affirmative and negative questions			10-fold cross-validation only negative questions		
	Average accuracy (SD)	Precision	Recall	Average accuracy (SD)	Precision	Recall
Logistic	85% (17.4)	0.818	0.900	82.5% (16.9)	0.810	0.850
SVM	85% (17.5)	0.889	0.800	80% (19.7)	0.875	0.700
Naïve Bayes	87.5% (13.2)	0.895	0.850	77.5% (18.4)	0.824	0.700
Random forest	77.5% (21.9)	0.789	0.750	75 % (23.6)	0.778	0.700
LMT	77.5% (24.9)	0.789	0.750	80% (19.7)	0.883	0.750

Table 4.22: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in 10-fold cross-validation considering the response to either affirmative and negative questions or only to negative questions. The accuracy in 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.1.14 Discussion

To conclude, the use of unexpected questions and mouse dynamics recording is an accurate paradigm to detect people declaring faked identities. In particular, we demonstrated that:

- The mouse trajectories of liars differ from those of truth-tellers, especially for responses to unexpected questions. Liars take more time than truth-tellers to compute their response and they make a greater number of errors.
- The difference between the two experimental groups is more evident for unexpected questions than expected and control questions.
- The classification accuracy of a subject as liar or truth-teller is around 90-95%.
- The number of errors is the predictor that gives the major contribution in correctly distinguishing liars from truth-tellers, with the other mouse tracking parameters fine-tuning the classification.
- Liars can be identified even when they are responding truthfully, but with lower accuracy (around 75%). In other words, liars show a kind of expansion of their mind-set to questions to which they respond truthfully.
- The accuracy of the classification models remain stable when the paradigm is apply in different cultures.
- The paradigm is relatively resistant to countermeasures. In fact, liars that are trained to beat the lie detectors with different strategies are detected with the 80% of accuracy. A different number of efficient alternative classification models can be trained, and subjects are in trouble to concurrently monitor all the response parameters.
- The use of changing response labels slightly improves the efficacy of the paradigm (2.5%-5% of increment in the classification accuracy).

- The introduction of stimuli in form of negation slightly affects the accuracy of the technique (7.5%-12.5% of decrease in the classification accuracy).

4.2 The Detection of Faked Identity with Unexpected Questions and Choice Reaction Time

In this experiment, we replicated the paradigm proposed in chapter 4.1, but recording RT instead of mouse dynamics. The general aim is to verify whether RT, along with the technique of unexpected questions, are enough to efficiently spot people who declare false identity information.

4.2.1 Participants

Fourty subjects participated in the experiment, and their data were used as training set to build ML models. Twenty participants were assigned to the liars' group and the other twenty to the truth-teller condition. The demographic characteristics of the sample are reported in Table 4.23.

Then, a second sample of 10 participants (five liars and five truth-tellers) was collected and used as test set to assess the model generalization. Demographic information about participants are in Table 4.23.

Sample	N	Gender	Age	Education
Training set	40	M = 17, F= 23	M = 22, SD = 1.4	M = 16, SD = 1.1
Test set	10	M = 4, F= 6	M = 24, SD = 3.3	M = 17, SD = 1.6

Table 4.23: demographic information about training and test set. In the second column (N) the number of participants for each sample is reported. The third column shows the number of male and female in each sample. The fourth and the fifth columns report mean (M) and standard deviation (SD) of participants' age and education.

4.2.2 Experimental procedure

The experimental procedure is the same reported in paragraph 4.1.2.

4.2.3 Stimuli

Each participant responded to 78 questions, 18 control, 20 expected and 40 unexpected (for an explanation about the type of questions see section 4.1.3). Half questions required a "yes" response, and the other half required to respond "no", for both liars and truth-tellers. Questions were presented in form of affirmation. Questions are reported in Table 4.24. For more details about the modalities of presentation of the stimuli, see paragraph 3.3.1.

4.2.4 Collected measures

During the task, RT and errors are collected. Then, for each participant we computed the average RT and the average number of errors, separately for control, expected and unexpected questions. Moreover, we calculated the average RT for questions where participants made errors (in other words, RT to wrong responses) and questions where participants did not make errors (RT in right responses). The Inverse Efficiency Score (IES) was also calculated for control, expected and unexpected questions. The final list of predictors is the following: RT control, RT expected, RT unexpected, RT control right responses, RT expected right responses, RT unexpected right responses, RT control wrong responses, RT expected wrong responses, RT unexpected wrong responses, errors control, errors expected, errors unexpected, IES control, IES expected, IES unexpected.

Type of question	Question that requires “yes” response by both liars and truth-tellers	Question that requires “no” response by both liars and truth-tellers	
Expected	My name is Alice	My name is Maria	
	My last name is Rossi	My last name is Bianchi	
	I was born in 1989	I was born in 1986	
	I was born in April	I was born in August	
	I was born on 20 th	I was born on 13 th	
	I was born in Mestre	I was born in Capri	
	I live in Limena	I live in Caserta	
	I live at Vespucci street	I live at Marconi street	
	I am single	I am married	
	I am a student	I am a professor	
	Unexpected	I am 27 years old	I am 23 years old
		My zodiac is Aries	My zodiac is Leo
		I was born in Veneto	I was born in Campania
		I was born in the province of Venice	I was born in the province of Napoli
I live in Veneto		I live in Campania	
I live in the province of Padova		I live in the province of Caserta	
Venezia is the capital of the region where I live		Napoli is the capital of the region where I live	
Venezia is the capital of the region where I was born		Napoli is the capital of the region where I was born	
My first name contains double letters		My first name is without double letters	
The initials of my name are A.R.		The initials of my name are M.B.	
I already celebrated the birthday this year		I have yet to celebrate the birthday this year	
My last name contains double letters		My last name is without double letters	
My age minus one year is 26		My age minus one year is 25	
The city where I was born is just north of Bologna		The city where I was born is just south of Roma	
My zip code is 35142	My zip code is 7863		
My telephone are code is 049	My telephone are code is 062		
I live near the sea	I live near the mountains		
I live in the same region where I was born	I live in a different region than where I was born		
Control	I live between Treviso and Rovigo	I live between Lucca and Arezzo	
	I was born near Venice	I was born near Torino	
	I am female	I am male	
	My skin is white	My skin is brown	
	I have a ring on my finger	My fingers are without rings	
	I have light eyes	I have dark eyes	
	I wear glasses	I am without glasses	
	I am wearing a green t-shirt	I am wearing a blu t-shirt	
	I am 160 cm high	I am 190 cm high	
	I am attending the university	I am attending the high school	
I am wearing pants	I am wearing a skirt		

Table 4.24: the table reports an example of the 78 expected, unexpected and control questions presented to participants and related to a truth or faked identity.

4.2.5 Features selection

Feature selection was performed using the CFS algorithm, as reported in section 3.4.1. The CFS algorithm gave the following output: RT wrong expected ($r_{pb} = 0.51$), RT wrong unexpected ($r_{pb} = 0.19$), IES expected ($r_{pb} = 0.54$), IES unexpected ($r_{pb} = 0.77$). Table 4.25 reports the correlation matrix of the four selected features and their correlation with the dependent variable (liar vs truth-teller).

	RT wrong expected	RT wrong unexpected	IES expected	IES unexpected	Condition
RT wrong expected	1.00	0.23	0.62	0.55	0.51
RT wrong unexpected	0.23	1.00	0.44	0.53	0.19
IES expected	0.62	0.44	1.00	0.70	0.54
IES unexpected	0.55	0.53	0.70	1.00	0.77
Condition	0.51	0.19	0.54	0.77	1.00

Table 4.125: the table reports the correlation matrix for the four selected features and their correlation value with the dependent variable.

4.2.6 Descriptive statistics

Table 4.26 reports the descriptive statistics for the four selected features.

Feature	Group	M (SD)
RT wrong expected	Liars	1236.62 (1134.05)
	Truth-tellers	195.9 (481.10)
RT wrong unexpected	Liars	3134.65 (1103.06)
	Truth-tellers	2613.94 (1609.95)
IES expected	Liars	1896.42 (392.57)
	Truth-tellers	1451.06 (288.54)
IES unexpected	Liars	4463.56 (1325.12)
	Truth-tellers	2195.06 (332.37)

Table 4.26: descriptive statistics for errors, AUC, MD-time and Y29. Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers). The last three columns report statistics about the difference between the two groups. In particular, the table shows the results of the independent *t*-test (*t*-value and *p*-value are reported), effect-size (Cohen’s *d*) and Bayes Factor (*BF*). For the interpretation of Cohen’s *d* and Bayes Factor, see paragraph 3.4.2.

An ANOVA indicated that overall, the wrong responses of liars were longer than those of truth tellers [$F(1,38) = 7.80, p < .01$]. In addition, both liars and truth-tellers had longer RT in responding to unexpected questions compared to expected questions [$F(1,38) = 77.31, p < .01$]. It should be noticed that truth-tellers had very short RT in giving wrong responses to expected questions. It means that when a truth-tellers fails in responding to expected questions, this is probably to the impulsivity in the response. On the other hand, the errors of the liars are probably due to the incapacity to retrieve the correct information.

Concerning IES, the ANOVA suggested that liars had a greater IES than truth-tellers [$F(1,38) = 51.06, p < .01$]. Moreover, both liars and truth-tellers had greater IES in responses to unexpected questions compared to expected questions [$F(1,38) = 151.60, p < .01$]. Finally, the interaction condition X type of question (expected vs unexpected) was statistically significant [$F(1,38) = 47.04, p < .01$].

4.2.7 Machine learning models

The four selected features (RT wrong expected, RT wrong unexpected, IES expected, IES unexpected) were entered in five different ML algorithms. Models were evaluated following a 10-fold cross-validation procedure, as described in paragraph 3.4.3. Results obtained by 10-fold cross-validation are reported in Table 4.27.

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	90% (12.9)	0.900	0.900	80%	0.800	0.800
SVM	87.5% (17.7)	0.826	0.950	90%	0.833	1.000
Naïve Bayes	90% (17.5)	1.000	0.800	90%	1.000	0.800
Random forest	97.5% (7.9)	0.952	1.000	90%	1.000	0.800
LMT	95% (10.5)	0.909	1.000	90%	1.000	0.800

Table 4.27: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in training and test set. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

Then, we tested the generalization of the models performance on the new set of 10 participants (see section 4.2.1). Results, in Table 4.27, confirmed that all the models reached an accuracy around 90% in classifying subjects as liars or truth-tellers, both in training and test. Random forest and LMT showed a better performance in 10-fold cross-validation respect to the other classifiers. However, this improvement does not generalize to the unseen data of the test set.

About the rate of false positive and false negative, the confusion matrix showed that the number of liars and truth-tellers misclassified is not equal for all the algorithms. Logistic regression produced a balanced number of false positive and false negative, failing in detecting two liars and two truth-teller in the training set and one liar and one truth-teller in the test set. The SVM was completely unbalanced towards the false negatives, misclassifying four liars and one truth-teller in training set and one liar in the test set. Naïve Bayes had an opposite performance, failing the classification of four truth-tellers in the cross-validation and one truth-teller in the test set. Random forest misclassified one liar in training set and only one truth-teller in the test set. Finally, LMT failed in recognizing two truth-tellers in cross-validation and one truth-teller in the test set.

4.2.8 Discussion

To conclude, it is possible spot liars declaring faked identities asking unexpected questions and measuring RT with similar accuracy of mouse dynamics recording. However, it is reasonable to think that RT are less resistant to countermeasure, as the subject has to monitor just one predictor.

4.3 The Detection of Faked Identity with Unexpected Questions and Keystroke Dynamics

In the previous experiments, we demonstrated that the technique of unexpected questions together with the mouse dynamics or reaction times recording is efficient in detecting liars about identity. However, the paradigm that we have proposed has a yes/no structure, and it requires a preliminary crafting of the questions by the experimenter. For this reason, the online application of this technique may be problematic. In order to overcome such caveats, here we report an experiment in which participants will respond to similar questions entering their response in an edit box using the keyboard. In the present work, we asked unexpected questions to participants but, differently from the previous studies, we recorded the subjects' typing pattern on the keyboard (keystroke dynamics). The main advantage of keystroke dynamics is that, differently from mouse tracking and reaction times, it can be adopted also in the situations in which is not possible to formulate ended "yes or no" questions, such as in some online contexts (e.g., a website subscription form).

The goal of these experiments is to validate a computerized technique to spot people who declare false identity information asking unexpected questions and analyzing keystroke dynamics.

Parts of methods and results that are reported in this section are now under review for publication.

4.3.1 Participants

A first sample of forty participants was recruited and data were used as training set to build ML models. Twenty participants were assigned to the liars' group and the other twenty to the truth-teller condition. The demographic characteristics of the sample are reported in Table 4.28.

Then, a second sample of 20 participants (ten liars and ten truth-tellers) was collected and used as test set to assess the model generalization. Demographic information about participants are in Table 4.28.

Sample	N	Gender	Age	Education
Training set	40	M = 12, F= 28	M = 23, SD = 1.9	M = 17, SD = 1.8
Test set	20	M = 6, F= 14	M = 22, SD = 1.7	M = 16, SD = 1.6

Table 4.28: demographic information about training and test set. In the second column (N) the number of participants for each sample is reported. The third column shows the number of male and female in each sample. The fourth and the fifth columns report mean (M) and standard deviation (SD) of participants' age and education.

4.3.2 Experimental procedure

The experimental procedure was the same reported in section 4.1.2.

4.3.3 Stimuli

The task required to answer 18 open-ended questions relating to identity. Both liars and truth-tellers were instructed to response each question typing the answer in an edit box (see Figure 3.4). For more details about the modalities of presentation of the stimuli, see paragraph 3.3.3. Before starting the experiment, participants completed three training questions (data from training questions were not further analyzed). The 18 experimental questions, randomly presented to subjects, belonged to the following categories: four control questions, eight expected questions, eight unexpected questions

(for an explanation about the type of questions see section 4.1.3). The complete list of questions is reported in Table 4.29.

It should be noted that liars told lies in both expected and unexpected questions, whereas they responded truthfully in control questions.

4.3.4 Collected measures

During the subjects' response, keystroke dynamics were recorded. For more details about data collection, see section 3.3.3. A total of 62 attributes were calculated, averaging each variable over the 18 responses given by each subject.

Type of question	Question text
Expected	What is your name?
	What is your last name?
	In which year were you born?
	In which month were you born?
	In which city were you born?
	In which city do you live?
	What is your home address?
Unexpected	What is your e-mail address?
	How old are you? (in letters)
	Which is your zodiac?
	In which region were you born?
	In which province were you born?
	In which region do you live?
	Which is the capital town of your residence region?
Control	What is your gender?
	What is the color of your skin?
	What is the color of your hair?
	What is your nationality?

Table 4.29: the table reports the entire list of 18 questions (expected, unexpected and control) presented to participants and related to a truth or faked identity.

4.3.5 Feature selection

Feature selection was performed as reported in paragraph 3.3.2, that is selecting the subset of predictors which had maximum correlation with the dependent variable and minimal intercorrelation between features. Firstly, we selected the predictors that showed the maximum correlation with the experimental condition (liar vs truth-teller): number of errors ($r_{pb} = 0.85$), prompted-firstdigit adjusted for the GULPEASE index ($r_{pb} = 0.71$) (for an explanation of the GULPEASE index see section 3.3.3), prompted-firstdigit ($r_{pb} = 0.70$), prompted-enter ($r_{pb} = 0.65$), firstdigit-enter ($r_{pb} = 0.46$), writing time ($r_{pb} = 0.50$), time before enter key down ($r_{pb} = 0.43$). Then, we looked at the intercorrelation between these seven features. Two of the seven predictors (prompted-firstdigit and prompted-enter) showed a very high correlation value, respectively with prompted-firstdigit adjusted GULPEASE ($r_{pb} = 0.99$) and with firstdigit-enter ($r_{pb} = 0.89$). Thus, these features have been excluded, in order to avoid redundancy. In Table 4.30, the correlation matrix between features is reported, as well as the correlation value between the five final attributes and the dependent variable (r_{pb}).

	Errors	Prompted-firstdigit GULPEASE	Firstdigit- enter	Writing time	Time key before enter down	Condition
Errors	1.00	0.51	0.25	0.46	0.44	0.85
Prompted-firstdigit GULPEASE	0.51	1.00	0.66	0.6	0.54	0.71
Firstdigit-enter	0.25	0.66	1.00	0.67	0.52	0.46
Writing time	0.46	0.6	0.67	1.00	0.67	0.5
Time key before enter down	0.44	0.54	0.52	0.67	1.00	0.43
Condition	0.85	0.71	0.46	0.50	0.43	1.00

Table 4.30: the table reports the correlation matrix for the four features that were selected and their correlation value with the dependent variable.

4.3.6 Descriptive statistics

Feature selection isolated, from the original set of 62 predictors, five independent variables: errors, prompted-firstdigit adjusted for the GULPEASE index, firstdigit-enter, writing time and time before enter key down. Table 4.31 reports the descriptive statistics for these features, as well as the analysis of the difference between truth-tellers and liars (*t*-test, Cohen's *d*).

Feature	Group	M (SD)	<i>t</i> -test (<i>t</i> , <i>p</i> -value)	Cohen's <i>d</i>
Errors	Liars	0.22 (0.08)	10.57, < .01	3.34
	Truth-tellers	0.01 (0.02)		
Prompted-firstdigit GULPEASE	Liars	4932.43 (1466.47)	6.48, < .01	2.05
	Truth-tellers	2406.06 (843.70)		
Firstdigit-enter	Liars	4968.79 (1875.60)	3.33, < .01	1.05
	Truth-tellers	3456.55 (564.41)		
Writing time	Liars	543.20 (174.98)	3.07, < .01	1.17
	Truth-tellers	376.06 (82.50)		
Time key before enter down	Liars	1413.12 (703.11)	3.04, < .01	0.96
	Truth-tellers	880.10 (320.35)		

Table 4.31: descriptive statistics for errors, prompted-firstdigit adjusted for the GULPEASE index, firstdigit-enter, writing time and time before enter key down. Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers). The last three columns report statistics about the difference between the two groups. In particular, the table shows the results of the independent *t*-test (*t*-value and *p*-value are reported) and effect-size (Cohen's *d*). For the interpretation of Cohen's *d*, see paragraph 3.4.2.

As the number of errors is the feature most correlated with the dependent variable, we analyzed the error rate separately for control, expected and unexpected questions. Results are reported in Table 32. It should be noticed that the error rate of the two groups is similar when they respond to control and expected question. By contrast, when responding to unexpected questions liars produce 27 times more errors than truth-tellers.

Type of question	Liars (N = 20)	Truth-tellers (N = 20)
Control (n = 80)	0 / 80	0 / 80
Expected (n = 160)	0 / 160	3 / 160
Unexpected (n = 120)	3 / 120	81 / 120

Table 4.32: number of errors in control, expected and unexpected questions that were committed by liars and truth-tellers.

4.3.7 Machine learning models

Five different classifiers (logistic, SVM, Naïve Bayes, random forest and LMT) were trained via 10-fold cross-validation procedure, using data from the first 40 participants as training set (see section 3.4.3). Then, in order to evaluate generalization of the results on completely new data, models were tested on the 20 new participants never used in the learning phase (see paragraph 4.2.1). Accuracies obtained by the classifiers during training and testing are reported in Table 4.33. All the classifiers reached at least 90% of accuracy in distinguishing liars from truth-tellers.

About the rate of false positive and false negative, the confusion matrix showed that the number of liars and truth-tellers misclassified is not equal for all the algorithms. Logistic regression failed in detecting one truth-teller and three liars in the training set, whereas in the test set all subjects all correctly detected. False positive and false negative were balanced in SVM and Naïve Bayes classification on training set (it misclassifies one liar and one truth-teller), whereas in the test set SVM failed in classify two liars and Naïve Bayes misclassified one truth-teller. Random forest misclassified two liars and one truth-teller in training set and only one liar in the test set. Finally, LMT failed in recognizing one liar in cross-validation and two liars in the test set.

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	90% (12.9)	0.864	0.950	100%	1.000	1.000
SVM	95% (10.5)	0.950	0.950	90%	0.833	1.000
Naïve Bayes	95% (10.5)	0.950	0.950	95%	1.000	0.900
Random forest	92.5% (12.1)	0.905	0.950	95%	0.909	1.000
LMT	97.5% (7.9)	0.952	1.000	90%	0.833	1.000

Table 4.33: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation and test set. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

In order to highlight the relative importance of predictors, we eliminated one by one the features recalculating the classification accuracy. Results are reported in Table 4.34. To sum up, eliminating errors from predictors classification accuracy decreases around 80% in the cross-validation and around 65% in the test. When prompted-firstdigit adjusted GULPEASE is eliminated, the overall accuracy remains substantially high (around 92% for training and around 90% in the test). Eliminating the firstdigit-enter variable, the accuracy remains high (around 95% both in cross-validation and test set). The same occurs by removing the writing time and time key before enter down. In short, errors are the single most important predictor in identifying subjects as liars or truth-tellers. Furthermore, the variables related to the response latency (prompted-firstdigit adjusted GULPEASE), the writing time (firstdigit-enter and writing time) and the interval between the last key press and the confirmation of the response (time before enter key down) are also contributing significantly.

Removed predictor	ML classifier	Training set (10-fold cross-validation)			Test set		
		Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Errors	Logistic	82.5% (16.9)	0.810	0.850	65%	0.600	0.900
	SVM	82.5% (16.9)	0.783	0.900	65%	0.600	0.900
	Naïve Bayes	80% (19.7)	0.750	0.900	60%	0.571	0.800
	Random forest	82.5% (16.9)	0.842	0.800	70%	0.643	0.900
	LMT	82.5% (12.1)	0.842	0.800	65%	0.600	0.900
Prompted-firstdigit adjusted for the GULPEASE index	Logistic	92.5% (12.1)	0.905	0.950	90%	0.833	1.000
	SVM	92.5% (12.1)	0.870	1.000	90%	0.833	1.000
	Naïve Bayes	92.5% (12.1)	0.905	0.950	90%	1.000	0.800
	Random forest	95% (10.5)	0.950	0.950	100%	1.000	1.000
	LMT	92.5% (12.1)	0.870	1.000	90%	0.833	1.000
Firstdigit-entert	Logistic	92.5% (12.1)	0.905	0.950	100%	1.000	1.000
	SVM	95% (10.5)	0.950	0.950	90%	0.833	1.000
	Naïve Bayes	95% (10.5)	0.950	0.950	95%	1.000	0.900
	Random forest	92.5% (12.1)	0.947	0.900	95%	0.909	1.000
	LMT	97.5% (7.9)	0.952	1.000	90%	0.833	1.000
Writing time	Logistic	92.5% (12.1)	0.905	0.950	100%	1.000	1.000
	SVM	95% (10.5)	0.950	0.950	90%	0.833	1.000
	Naïve Bayes	95% (10.5)	0.950	0.950	95%	1.000	0.900
	Random forest	90% (12.9)	0.900	0.900	95%	0.909	1.000
	LMT	97.5% (7.9)	0.952	1.000	90%	0.833	1.000
Time before enter key down	Logistic	90% (12.1)	0.900	0.900	85%	0.818	0.900
	SVM	95% (10.5)	0.950	0.950	90%	0.833	1.000
	Naïve Bayes	95% (10.5)	0.950	0.950	95%	1.000	0.900
	Random forest	90% (12.9)	0.900	0.900	95%	0.909	1.000
	LMT	97.5% (7.9)	0.952	1.000	90%	0.833	1.000

Table 4.34: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in training and test set, using different set of predictors. The first column indicates the attribute that has been removed from the original set of five predictors (see paragraph 4.3.2). The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.3.8 Can we detect liars also when they respond truthfully?

All the analysis reported above, were run taking into account the responses to all three types of questions (control, expected and unexpected questions). We specifically analyzed control questions alone as both liars and truth-tellers are required to respond truthfully to control questions. All classifiers yielded a classification around chance level to this type of questions (47.5% as regards the cross-validation and to 50% in the test (see Table 4.35). This result indicates that responses to control questions of the two groups are virtually indistinguishable. In other words, using this paradigm, liars and truth-tellers are not detectable also when liars responded truthfully.

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	40% (24.1)	0.400	0.400	60%	0.571	0.800
SVM	45% (15.9)	0.458	0.550	45%	0.471	0.800
Naïve Bayes	47.5% (27.5)	0.481	0.650	45%	0.471	0.800
Random forest	55% (35.0)	0.550	0.550	40%	0.417	0.500
LMT	45% (15.8)	0.474	0.900	50%	0.500	1.000

Table 4.35: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation and test set using analyzing only control questions. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.3.9 Analysis on normalized predictors

Normalized predictors are features that are less influenced by inter-individual and environmental variables. In fact, the analysis reported above were conducted on raw data using two groups of subjects (liars and truth-tellers) similar in age, cultural level and typing skills. One could argue that keyboard dynamics are modulated by a number of different variables such as age, cultural level and typing skills. In order to render the results generalizable it would be interesting to see whether similar results hold not only for raw data but also for normalized predictors. To overcome this limitation, we run again the classification models using only normalized indexes. These indices were the following:

- Average number of errors = (number of errors / total number of questions)
- Writing time = (firstdigit-enter / answer length)
- Firstdigit time = (Prompted-firstdigit – prompted-enter)
- ([Writing time / prompted-firstdigit] – prompted-enter)

In addition to these four new normalized predictors, errors were included in the models, as errors do not depend on typing skills, age or cultural level. Results from the five classifiers using the normalized predictors are reported in Table 4.36. In short, it is confirmed also for normalized predictors the high accuracy in classifying truth-tellers and liars.

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	90% (12.9)	0.900	0.900	100%	1.000	1.000
SVM	92.5% (12.1)	0.870	1.000	90%	0.833	1.000
Naïve Bayes	90% (17.5)	0.900	0.900	95%	1.000	0.900
Random forest	95% (10.5)	0.950	0.950	100%	1.000	1.000
LMT	90% (12.9)	0.833	1.000	90%	0.833	1.000

Table 4.36: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation and test set using normalized predictors. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.3.10 Countermeasures and alternative efficient models

We did not directly tested resistance to countermeasures of this technique, but a number of reasons indicate that coaching subjects could be difficult, and specifically:

- 1) Errors to unexpected questions are diagnostic of lying and the subjects should respond without errors in order to cheat the test. There are no easy countermeasures to the number of errors. In fact, the only way to beat the test avoiding errors is to know the unexpected questions in advance. Countermeasures, therefore, are limited to the keystroke dynamics.
- 2) Parameters used to encode keystroke dynamics are high in number and only some of them have been used in building the original model. It is unlikely that the cheater succeeds in implementing countermeasures that simultaneously keep under voluntary control all possible efficient predictors. To highlight this points, we have tested a new model that uses as predictors errors ($r_{pb} = 0.85$), prompted-firstdigit ($r_{pb} = 0.70$), prompted-enter ($r_{pb} = 0.65$), time before enter key flight ($r_{pb} = 0.43$), and di-graph down time average ($r_{pb} = 0.38$). For an explanation about these predictors, see paragraph 3.3.3. Note that the predictors used in the original analysis were errors, prompted-firstdigit adjusted GULPEASE, firstdigit-enter, writing time and time before enter key down.

Classification accuracies using the new set of predictors are reported in Table 4.37. These results clearly show that there are other sets of predictors (different from those originally used) that can be used to efficiently classify the participants, making it hard to apply countermeasures.

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	92.5% (12.1)	0.905	0.950	100%	1.000	1.000
SVM	95% (10.5)	0.950	0.950	90%	0.833	1.000
Naive Bayes	97.5% (7.9)	1.000	0.950	90%	1.000	0.800
Random forest	90% (12.9)	0.864	0.950	90%	0.833	1.000
LMT	97.5% (7.9)	0.952	1.000	90%	0.833	1.000

Table 4.37: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation and test set using a new set of five predictors. The new set of predictors, different from those originally used, included errors, prompted-firstdigit, prompted-enter, time before enter key flight and di-graph down time average. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.3.11 Classification of liars using only data from truth-tellers

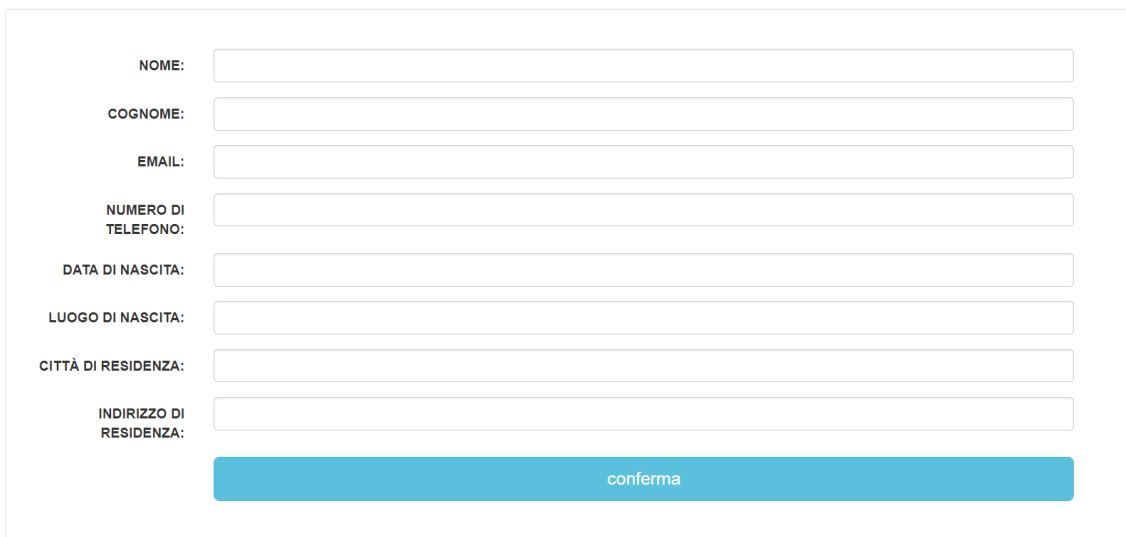
While liars are instructed to lie about their identity, truth-tellers are required to respond freely without no specific instructions. Under this view, liars are responding in an anomalous way with respect to truth-tellers. Normally, in a real situation, the majority of the subjects report true identities and only few of them provide false information and show an anomalous pattern of response. In order to evaluate whether liars may still be identified based on their anomalous response style we have applied a machine learning technique called anomaly detection [185]. Anomalies are data that have different patterns from normal instances. The detection of anomalies provides significant information and finds application in many fields. For example, detection of anomalies in credit card transaction, anomalies

in an astronomy image or in a nuclear plant functioning. Anomaly detection techniques classify subjects after a training limited to the most frequent group, in our experiment the truth-tellers [186]. At prediction, new instances with unknown class labels can either belong to the target class (the class learned during training, truth-tellers) or to a new class that was not available during training (in our case the liars). This type of learning problem is known as one-class classification. Following this logic, we tested whether a one-class classifier [185] may classify liars satisfactorily even if the model is trained only using data from truth-tellers. This ML algorithm has been trained using logistic on the 20 original truth-tellers' data and tested on the test set of 20 participants (10 liars and 10 truth tellers). The one-class algorithm classified correctly the 85% of instances (precision = 1.000, recall = 0.700); specifically it classified correctly the 70% of the truth-tellers as target and the 100% of the liars as outliers. If we run the test on a group of 30 liars and 10 truth-tellers results are 29/30 liars correctly classified and 7/10 truth tellers correctly classified. This result indicated that the classifier trained only on truth-tellers might identify the liars with high accuracy.

4.3.12 Application of the paradigm to online form

As anticipated above, lie detection via keystroke dynamics is more suitable than mouse tracking for the application in online contexts (e.g., to verify the authenticity of information typed by the user during and online subscription). Moreover, this setting permits a covert lie detection (a lie detection procedure that the respondent is unaware of it) preserving the usability and the user experience of the web surfing.

To argue this point, we have adapted the experiment to the situation of online subscription. An online form has been presented to subjects, asking them to compile it with the goal to subscribe for a new chat system. The subscription form, as appeared to the subject, is shown by Figure 4.13.



The image shows a web form for registration. It contains the following fields from top to bottom:

- NOME:
- COGNOME:
- EMAIL:
- NUMERO DI TELEFONO:
- DATA DI NASCITA:
- LUOGO DI NASCITA:
- CITTÀ DI RESIDENZA:
- INDIRIZZO DI RESIDENZA:

At the bottom of the form is a blue button labeled "conferma".

Figure 4.13: the figure reports the online form that the subjects were asked to compile in order to subscribe a new chat website.

After they successfully compiled the form, four unexpected questions appeared rapidly. Subjects were asked to answer these questions to confirmed their registration. An example of the screen containing unexpected questions to which subjects were asked to respond after filling the form, is reported in Figure 4.14. Unexpected questions were about age (“How old are you?”), ZIP code (“Which is your

ZIP code? ”), zodiac (“Which is your zodiac?”), and the capital town of the residence region (“Which is the capital town of your residence region?”).

Figure 4.14: an example of unexpected question as appeared to the participants during the task.

One-hundred participants took part to the experiment, 50 liars and 50 truth-tellers. The sample consisted in 43 male and 57 female, with average age = 23 (SD = 2.4), and average education = 15 (SD = 1.5). The experimental procedure that was followed to instruct liars to respond according to a learned faked identity is the same reported in section 4.1.2.

The keystroke dynamics features were collected only for responses to the four unexpected questions. The features were the same reported in paragraph 4.2.4.

Feature selection was performed using the CFS algorithm, as reported in section 3.4.1. The selected features were the following: number of errors ($r_{pb} = 0.52$), prompted-firstdigit ($r_{pb} = 0.57$), prompted-enter ($r_{pb} = 0.62$), writing time ($r_{pb} = 0.39$) and the maximum value of di-graph up and down time ($r_{pb} = 0.62$). For an explanation about these predictors, see paragraph 3.3.3.

An independent t -test confirmed that liars and truth-tellers statistically differ in all the five selected variables ($p < .01$).

Using five different ML classifiers, we run a 10-fold cross-validation on the entire sample of participants. Classification accuracies are reported in Table 4.38, and are around 85%. This means that the recording of keystroke dynamics during the response to only four unexpected questions is enough to detect liars with a high accuracy. This result also demonstrates that the identity detection system that we propose here is easy to integrate in the existing web application (e.g., registration or login form), without radically modifying the user experience.

ML classifier	10-fold cross-validation		
	Average accuracy (SD)	Precision	Recall
Logistic	84% (7.0)	0.886	0.780
SVM	87% (8.2)	0.894	0.840
Naïve Bayes	87% (6.8)	0.911	0.820
Random forest	85% (10.9)	0.872	0.820
LMT	86% (9.7)	0.891	0.820

Table 4.38: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation on the sole base of four unexpected questions. The accuracy in 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.3.13 Discussion

Through the two last experiments, we demonstrated that keystroke dynamics, along with unexpected questions, is an accurate technique to spot faked identities. Results are similar to those obtained recording mouse dynamics or RT. We also showed that this paradigm, compared to mouse dynamics and RT paradigms, is more suitable for online applications, as it is more promising in terms of integration with the existing online applications.

The most interesting predictors are the number of errors, which are due to the presence of unexpected questions, and the response latencies, such as the prompted-firstdigit index, the firstdigit-enter, the writing time and the time before enter key down. In other words, the high classification accuracy seems based on the total time taken to elaborate and produce the response, rather than the writing pattern in terms of up and down of each key during the response typing.

Although the results of the experiments hitherto presented are astonishing, the high classification accuracy is largely due to the effect of the unexpected questions. In the experiments that we presented in the previous sections, the subjects spent few minutes to learn a faked identity. It is true that in some real cases (e.g., terrorists traveling with a false passport) the subject have more time to learn more in detail the new false identity. However, it is also true that in other real cases (e.g., a user that try to subscribe a website with a faked identity) liars are not well prepared. Since the method is based on asking unexpected questions, the time taken to learn the information is not so crucial. Indeed, what is crucial to beat the test is to be prepared to respond with unexpected information. This imply that the subject should be aware in advance about the underlying logic of the task, and that he should know in advance the possible questions. Asking unexpected questions is a complex crafting process that requires identifying what are the unexpected questions for the examinee. If an unexpected question is not really unexpected (as the subject may prepare himself in advance) it becomes an expected question and loses its efficacy. Moreover, though unexpected questions can be varied in content (e.g., postal code could be included and so on), they are difficult to apply in all deception detection situations. For example, in crimes that consist on having or not put in place an action (e.g., I dealt /I didn't deal drug in last few months), it is extremely difficult to fabricate unexpected questions. Such type of lies are known as lie of omission and substantially consist in deny an action [147]. Furthermore, in some cases, the crime details are unknown and the investigators have no elements to build unexpected questions. To overstep the limits of unexpected questions, we proposed an alternative strategy to increase the liars' cognitive load: the use of complex questions. To guarantee comparable results with the previous experiments, we have followed the same experimental procedure focusing on the detection of faked identities through mouse dynamics and RT recording.

4.4 The Detection of Faked Identity with Complex Questions and Choice Reaction Time

The general aim of this experiment is to validate a computerized technique to spot people who declare false identity information asking complex questions and analyzing RT.

As anticipated in section 2.3.2, the use of complex questions is a technique to increase liars' cognitive load, facilitating the detection of false responses. It consists in asking questions that are composed by more than one target information. For example, investigating the subject identity, a complex question may be composed by the information about the name and the date of birth (e.g., "Are you Mary born in April 1987?").

Methods and results that are reported in this section are under review for publication.

4.4.1 Participants

Data from a first sample of 40 participants were collected and used as training set to build ML models. Twenty participants were assigned to the liars' group and the other twenty to the truth-teller condition. Then, ten new participants (five liars and five truth-tellers) were recruited and used as test set to assess the ML models generalization. Demographic information about participants are reported in Table 4.39.

Sample	N	Gender	Age	Education
Training set	40	M = 15, F= 25	M = 22, SD = 1.5	M = 16, SD = 1.4
Test set	10	M = 2, F= 8	M = 22, SD = 2.5	M = 16, SD = 1.4

Table 4.39: demographic information about training and test set. In the second column (N) the number of participants for each sample is reported. The third column shows the number of male and female in each sample. The fourth and the fifth columns report mean (M) and standard deviation (SD) of participants' age and education.

4.4.2 Experimental procedure

The experimental procedure was similar to that reported in section 4.1.2. In other words, liars were asked to learn a faked identity and to respond questions according to the information previously learned. On the contrary, no specific instruction were given to truth-tellers, who completed the task responding according to their actual identity.

4.4.3 Stimuli

Sixty questions were presented to participants in form of affirmation. Thirty sentences required a "yes" response, and 30 required to respond "no", for both liars and truth-tellers. The experimental questions were preceded by 10 training questions (5 requiring a "yes" response and 5 requiring a "no" response) to allow the subject to familiarize with the task (data from training questions were not included in the analysis). Questions belonged to the following categories:

- 20 control questions. Control questions were sentences to which both truth-tellers and liars had to respond truthfully. These sentences were unrelated to identity and referred to the experimental condition. Half of the control sentences required a "yes" response (e.g., "I am sitting in front of a computer") and half a "no" response (e.g., "I am climbing a mountain"). Both liars and truth tellers were required to respond truthfully to all control questions.

- 20 simple questions. Simple questions were related to the identity. Truth-tellers responded were asked to respond according to their true identities, whereas liars responded according to the faked identity previously learned. Half of the simple sentences required to respond “yes” (e.g., “My name is John”) and the others required a “no” response (e.g., “My name is Antony”). For liars, a “yes” response corresponded to a lie.
- 20 complex questions. Complex questions were sentences that included two or three information about the identity (e.g., “I am Mary, a 29 years old girl from Venice”). Participants were instructed to respond “yes” when all of the information in the sentence was true, whereas they responded “no” when at least one of the information included in the sentence was false. In other words, participants had to respond “yes” when the entire sentence was true, and “no” when there was one or more pieces of false information in the sentence. Complex sentences that required to respond “yes” were composed as follows: five sentences contained two identity information and five sentences contained three identity information. Complex sentences requiring a “no” response were organized as follows: five sentences were composed by two identity information (in this case the false information was always in the second place). The other five sentences had three identity information (three of them with only one false information in the last place of the sentence and two with two false information in the second and third places of the sentence). An example of questions is reported in Table 4.40.

To sum up, both liars and truth-tellers responded to 30 control, simple and complex questions that required to respond “yes” and to 30 control, simple and complex questions that required “no” responses. Control, simple and complex questions were presented randomly and intermixed. For more details about the modalities of presentation of the stimuli, see paragraph 3.3.1.

It should be noted that liars told lies only in the simple “yes” and complex “yes” responses. In fact, for the liars, the simple and complex questions regarding their faked identities were actually “no” responses that, because they were lying, required “yes” responses. In all of other questions (control “yes”, control “no”, simple “no”, complex “no”), both liars and truth-tellers responded truthfully.

4.4.4 Collected measures

For more details about data collection, see section 3.3.1. During the subjects’ response, RT and errors were recorded. For each participant, we averaged the RT and errors belonging to different type of questions (control, simple, complex, also separately for yes and no). Then, the Inverse Efficiency Score (IES) for control, simple and complex questions was computed. The final list of the 23 predictors that have been taken into account for ML analysis is reported in Annex 3.

Type of question	Question that requires “yes” response by both liars and truth-tellers	Question that requires “no” response by both liars and truth-tellers
Control	I am in front of a computer I am standing in front of a computer I'm using a computer I am responding with a keyboard Now, I'm sitting on a chair Right now, I'm sitting I'm doing an identity verification test I am participating in a test I'm reading sentences I am responding to phrases	I'm swimming in the sea I'm climbing a mountain I'm traveling by airplane I am aboard an airplane Now, I am on the beach I'm taking in the sun at the beach I'm eating at the restaurant I'm having lunch at the restaurant I'm playing football I'm watching a football match
Simple	My name is Merylin I was born in Trieste I live in Monfalcone I am single My last name is Monaro I am a student I was born in the province of Trieste I live in Via Timavo 47 I was born on 20 th April I was born in 1987	My last name is Zurri. My name is Greta. I was born on 08.15.1990. My birthday is in August. I'm married. My city of birth is Ortona. I live in Lanciano. I am an engineer. I was born in the province of Chieti I live in via Postojna.
Complex	I am <u>Merylin</u> , born in <u>Trieste</u> . My name is <u>Merylin Monaro</u> . I was born in <u>Trieste</u> , and I live in <u>Monfalcone</u> . I was born in <u>April 1987</u> in <u>Trieste</u> . I am a <u>student</u> , and I live in <u>Via Timavo 47</u> . In the <u>1987</u> , I was born in <u>April</u> in <u>Trieste</u> . I am <u>Monaro</u> , a <u>single student</u> . My name is <u>Merylin</u> , I am <u>single</u> and I live in <u>Monfalcone</u> . I am the <u>student Merylin</u> , born on <u>20.04.1987</u> . I am the <u>student Merylin</u> , born in <u>April</u> .	I am <u>Merylin</u> , and I live in Lanciano . I was born on <u>20th April 1987</u> in Ortona . I was born in <u>Trieste</u> , and I live in Lanciano . I am a <u>student</u> , and I am married . I live in <u>Monfalcone</u> in via Postojna 65 . I am <u>Merylin</u> , a single engineer . I am <u>Merylin</u> , a <u>student</u> born in August . <u>Merylin Monaro</u> was born in August 1990 . I am <u>Merylin Zurri</u> , and I am married . I am <u>Monaro</u> , a married woman of Ortona .

Table 4.40: the table reports an example of the 60 control, simple and complex questions presented to participants and related to a truth or faked identity. Information about the identity in complex questions are underlined. In bold the false identity information according to which participants had to respond “no”.

4.4.5 Descriptive statistics

Before the features selection, we run some descriptive analysis on RT and errors in control, simple and complex questions, to highlight the differences between the two experimental groups. In Tables 4.41 and 4.42, average number of errors and average RT in control, simple and complex questions are reported, taking into account also the difference between “yes” and “no” responses. It can be noticed that liars, in responding “no” to complex sentences, are 30% slower than the average RT (second column), whereas truth tellers in the same stimuli are only 10% slower. It is worth noting that the number of errors that liars made, on average, was 5.6 times the number of errors of truth tellers.

An ANOVA indicated that:

- overall, the responses of liars were longer than those of truth tellers [$F(1,38)=38.39$ $p<0.001$].
- Complex questions were slower than simple questions for both liars and truth-tellers [$F(2,76)=147.45$ $p<0.001$], but complex sentences were much slower, with respect to simple

sentences, in liars than in truth-tellers [F(2,76)=25.22, p<0.01]. In fact, the difference in RT between complex and simple sentences was 848 ms for liars and only 463 ms for truth-tellers.

- There is not a main effect of the response type (yes/no) [F(1,38)=2.34, p>0.01]. The interactions group X response type, question type X response type and group X question type X response type do not show statistically significant results (respectively [F(1,38)=1.88, p>0.01], [F(2,76)=4.62, p>0.01] and [F(2,76)=2.91, p>0.01]). It means that generally, both liars and truth-tellers, have the same RTs when responding yes or no questions. It excludes the possibility that the effect observed in the complex sentences is due to the act of negating rather than the lie itself.

Group	Total M (SD)	Control M (SD)		Simple M (SD)		Complex M (SD)	
		YES	NO	YES	NO	YES	NO
Liars	1974 (321.66)	1491	1570	1796	1748	2644	2596
		(302.2)	(353.14)	(328.78)	(292.41)	(711.76)	(580.58)
Truth-tellers	1389 (273.36)	0.75	0.79	0.90	0.88	1.33	1.31
		1283	1251	1201	1238	1842	1521
		(266.74)	(246.41)	(337)	(270.18)	(396.82)	(286.72)
		0.92	0.90	0.86	0.89	1.32	1.10

Table 4.41: descriptive statistics for RT related to control, simple and complex questions requiring a “yes” or “no” response. Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers). Under RT mean and sd, the ratio between the average RT for the specific type of question and the overall RT in all the task is reported.

Group	Total M (SD)	Control M (SD)		Simple M (SD)		Complex M (SD)	
		YES	NO	YES	NO	YES	NO
Liars	0.093 (0.092)	0.05	0.065	0.1	0.1	0.08	0.165
		(0.20)	(0.22)	(0.10)	(0.10)	(0.13)	(0.19)
Truth-tellers	0.014 (0.013)	0.53	0.69	1.07	1.07	0.86	1.77
		0.015	0.005	0.005	0.01	0.025	0.025
		(0.05)	(0.02)	(0.02)	(0.03)	(0.05)	(0.04)
		1.07	0.35	0.35	0.71	1.78	1.78

Table 4.42: descriptive statistics for errors related to control, simple and complex questions requiring a “yes” or “no” response. Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers). Under errors mean and sd, the ratio between the average number of errors for the specific type of question and the overall number of errors in all the task is reported.

4.4.6 Feature selection

Data collection produced a set of 23 predictors. Feature selection have been performed using a correlation based feature selector (CFS) (see section 3.4.1). Running this algorithm, the following predictors were selected: Simple Yes RT ($r_{pb} = 0.67$), Complex Tot RT ($r_{pb} = 0.73$), Complex No RT ($r_{pb} = 0.77$), Mean Total errors ($r_{pb} = 0.55$) and Mean Simple Tot errors ($r_{pb} = 0.66$). For a detailed explanation of these variables see Annex 3. In Table 4.43, the correlation matrix between features is reported, as well as the correlation value between the five final attributes and the dependent variable (r_{pb}).

	Simple Yes RT	Complex Tot RT	Complex No RT	Mean Total errors	Mean Simple Tot errors	Condition
Simple Yes RT	1.00	0.85	0.84	0.47	0.51	0.68
Complex Tot RT	0.85	1.00	0.93	0.42	0.43	0.73
Complex No RT	0.84	0.93	1.00	0.48	0.51	0.77
Mean Total errors	0.47	0.42	0.48	1.00	0.88	0.55
Mean Simple Tot errors	0.51	0.43	0.51	0.88	1.00	0.66
Condition	0.68	0.73	0.77	0.55	0.66	1.00

Table 4.30: the table reports the correlation matrix for the four features that were selected and their correlation value with the dependent variable

4.4.7 Machine learning models

Five ML classifiers (logistic, SVM, Naïve Bayes, random forest and LMT) were trained using the 10-fold cross-validation on the data of the first 40 participants (see section 3.4.3). Then, in order to evaluate the capacity of generalization of the models, we tested 10 new participants never seen by the classifiers (see paragraph 4.4.1). Accuracies obtained by the classifiers during training and testing are reported in Table 5. All the classifiers reached at least 90% of accuracy in 10-fold cross-validation and 80% in test set.

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	90% (12.9)	0.944	0.850	80%	0.714	1.000
SVM	95% (10.5)	0.909	1.000	80%	0.714	1.000
Naïve Bayes	90% (12.9)	0.900	0.900	90%	0.833	1.000
Random forest	90% (12.9)	0.864	0.950	80%	0.714	1.000
LMT	95% (10.5)	0.900	1.000	80%	0.714	1.000

Table 4.44: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation and test set. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

About the rate of false positive and false negative, the confusion matrix showed that the number of liars and truth-tellers misclassified in cross-validation is not equal for all the algorithms, while in the test set all the classifiers failed in recognizing liars. Logistic regression failed in detecting three truth-tellers and one liar in the training set, whereas in the test set it misclassified two liars. SVM and LMT were unbalanced toward the false negative, as they misclassified two liars both in training and in test set. Naïve Bayes misclassified two truth-tellers and two liars in cross-validation and one liar in the test set. Random forest failed the classification of one liar and three truth-tellers in training set and two liars in the test set.

4.4.8 Analysis on normalized predictors

Similarly to section 4.3.9, here we face the issue of using normalized predictors. In fact, there are a large a number of variables, such as age and cultural level, which may influence the subject's task performance. The analysis reported above were conducted on raw data using two groups of subjects (liars and truth-tellers) that were similar in age and cultural level, and that were tested in the same environment. In order to render the results generalizable, it would be interesting to see whether similar

results hold not only for raw data but also for normalized predictors. For this reason, we run again the classification models using only normalized indexes, less influenced by inter-individual and environmental variables. For example, raw RT for “yes” responses could be substituted by the ratio of the same data with the average RT of all subject responses. This ratio calibrates the result with the average speed of the participant, which, in turn, could depend on a number of factors. The complete list of the normalized predictors is provided in Annex 3.

Using the same feature selection logic, we extracted the subset of predictors more correlated with the dependent variable and less intercorrelated. The predictors are the following: Control Yes RT/Total RT ($r_{pb} = 0.59$, 2), Complex Tot RT/Total RT ($r_{pb} = 0.56$, 3), Complex No RT/Total RT ($r_{pb} = 0.66$), (Complex Yes RT – Complex No RT)/Total RT ($r_{pb} = 0.39$), Raw Simple Tot errors/Raw Control Tot errors ($r_{pb} = 0.60$).

Running the five ML algorithms on the new set of normalized predictors, we obtained similar accuracies to those highlighted using raw predictors. Results are in Table 4.45.

ML classifier	Training set (10-fold cross-validation)			Test set		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	85% (12.9)	0.850	0.850	90%	0.833	1.000
SVM	87.5% (17.7)	0.800	1.000	80%	0.714	1.000
Naïve Bayes	90% (12.9)	0.900	0.900	80%	0.714	1.000
Random forest	90% (12.9)	0.900	0.900	90%	0.833	1.000
LMT	85% (12.9)	0.850	0.850	90%	0.833	1.000

Table 4.45: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation and test set using normalized predictors. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

In short, when using normalized predictors, we observed similar results to those observed on raw data. The adoption of such normalized predictors instead of raw data renders, in theory, generalization more robust and less affected by the effects on reaction times of age, skill level, etc.

4.4.9 Analysis by stimuli

The results reported above were obtained with an analysis by subjects, and therefore, the accuracy of classifiers refers to the accuracy in classifying individual responders as liars or truth tellers. An interesting issue is whether the subject may be classified based on his individual responses through a majority vote. To investigate this issue we run an analysis by stimuli.

Given that the responses to complex sentences which require a “no” response are those that showed a higher correlation with the experimental condition (see paragraphs 4.4.1 and 4.4.2), we carried out the classification by stimuli using only the responses to these questions. A total of 400 responses were collected (40 subjects who responded each to 10 sentences which required a “no” response, for a total of 400 sentences). The predictors were the reaction time to the presented sentence and a categorical variable indexing whether the response was correct or wrong. Classifications results are reported in Table 4.46.

ML classifier	Training set	Test set				
	(10-fold cross-validation)	Precision	Recall	Accuracy	Precision	Recall
	Average accuracy (SD)					
Logistic	75.5% (3.9)	0.787	0.700	74%	0.853	0.580
SVM	73% (4.5)	0.865	0.545	70%	0.885	0.460
Naïve Bayes	73.5% (5.0)	0.856	0.565	69%	0.880	0.440
Random forest	69.25% (6.9)	0.695	0.685	70%	0.750	0.600
LMT	74.5% (4.7)	0.782	0.680	72%	0.844	0.540

Table 4.46: the table reports the accuracy obtained by five classifiers in correctly identify a response as true or false, in 10-fold cross-validation and test set. The models were developed taking into account only responses to complex “no” questions. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

This analysis showed that complex “no” sentences may be identified as originating from a truth-teller or a liar with an accuracy of 70%. Using the majority rule to classify a participant as a truth-teller or liar, the 10 participants of the validation sample were classified correctly 80% of the time (all truth-tellers were correctly classified, whereas two liars were misclassified).

To have an intuition on how responses are efficiently classified as belonging to a liar or a truth-teller, we run a J48 tree classification model (see section 3.4.3). This algorithm yielded an accuracy of 73% (DS = 5.6, precision = 0.719, recall = 0.755) in the 10-fold cross-validation and an accuracy of 72% (precision = 0.806, recall = 0.580) in the test.

```

PresentParola.RT <= 1870
|  errori <= 0: sincero (208.0/50.0)
|  errori > 0: mentitore (13.0/4.0)
PresentParola.RT > 1870: mentitore (179.0/38.0)

Number of Leaves :    3
Size of the tree :    5

```

Figure 4.15: output of the J48 tree from the 10-fold cross-validation. The output describes the model on which is based the decision about the classification of the response to complex “no” questions as belonging to a liar or a truth-teller.

The J48 works as follows (see Figure 4.15):

- if $RT \leq 1870$ ms and errors = 0, then the responder is a truth-teller (with accuracy = 75%).
- if $RT \leq 1870$ ms and errors = 1, the responder is a truth-teller (with accuracy = 70%).
- if $RT > 1870$ ms, then the responder is a liar (with accuracy = 78%).

In short, if RT is fast (below 1870 ms) and the response correct, then the responder is a truth-teller. If RT is slow (above 1870 ms), then the responder is a liar. Finally, if RT is fast but the response is an error, the truth-teller is classified with a slightly reduced accuracy.

4.4.10 Discussion

Through this experiment we demonstrated that people who declare a false identity can be detect asking complex questions and recording RT. The accuracy of this technique is slightly smaller compared to that obtained applying the technique of asking unexpected questions (see the experiment in section 4.2.). In fact, although the two experiments showed similar accuracies in cross-validation, in the test

set the results obtained from the complex questions technique are approximately 10% lower than results from the unexpected questions technique. However, it should be noticed that the results obtained by the application of complex questions are freer from errors and do not require the examiner to produce new information, as in the case of unexpected questions. For these reasons, as anticipated above, complex questions are more suitable for the application in those cases where details are unknown and the investigators have no elements to build unexpected questions.

4.5 The Detection of Faked Identity with Complex Questions and Mouse Dynamics

The general aim of this experiment is to replicate the paradigm of complex questions to spot people who declare false identity information analyzing mouse dynamics instead of RT

Methods and results that are reported in this section are under review for publication.

4.5.1 Participants

Forty participants were recruited, 20 liars and 20 the truth-tellers. The sample consisted of 18 male and 22 female, with average age = 23 (SD= 2.1) and average education level = 17 (SD = 1.3).

4.5.2 Experimental procedure

The experimental procedure was the same described in section 4.1.2. Twenty participants were asked to perform the experimental task responding truthfully, while the other 20 were asked to lie about their identity according to the faked information learned from a false ID card.

4.5.3 Stimuli

Stimuli presented to participants were the same reported in section 4.4.3. To sum up, each participant responded to 60 questions, 20 control, 20 simple and 20 complex. Half questions required a “yes” response, and the other half required to respond “no”. The experimental questions were preceded by 10 training questions (5 requiring a “yes” response and 5 requiring a “no” response) to allow the subject to familiarize with the task (data from training questions were not included in the analysis). For the complete list of questions and a definition of control, simple and complex question, see paragraph 4.4.3. For more details about the modalities of presentation of the stimuli, see paragraph 3.3.2.

4.5.4 Collected measures

During the subjects’ response, mouse dynamics were recorded by MouseTracker software (see section 3.3.2 for more details about data collection). In addition to the errors and the spatial-temporal features extracted by default by the MouseTracker software (IT, RT, MD, AUC, MD-time, x-flip, y-flip, see paragraph 3.3.2), we also calculated the minimum, maximum and average velocity and acceleration along the x and y -axis during the response. Finally, for each feature, we calculated the average value of the stimuli, separately for control yes, control no, simple yes, simple no, complex yes and complex no sentences. In this way we obtained a final list of 120 predictors.

4.5.5 Analysis of trajectories

First, a comparison between liars and truth-tellers’ motor response has been made observing their averaged mouse trajectories in control, simple and complex questions (see figure 4.16). It can be noticed that the two experimental groups have mostly overlapping trajectories for control and simple questions, whereas they differ in complex questions. In such stimuli, truth-tellers show straight trajectories from the origin to the response box. On the contrary, liars show wider trajectories, characterized by a greater AUC and MD. This visual pattern is in line to that obtained by the observation of motor trajectories on unexpected questions.

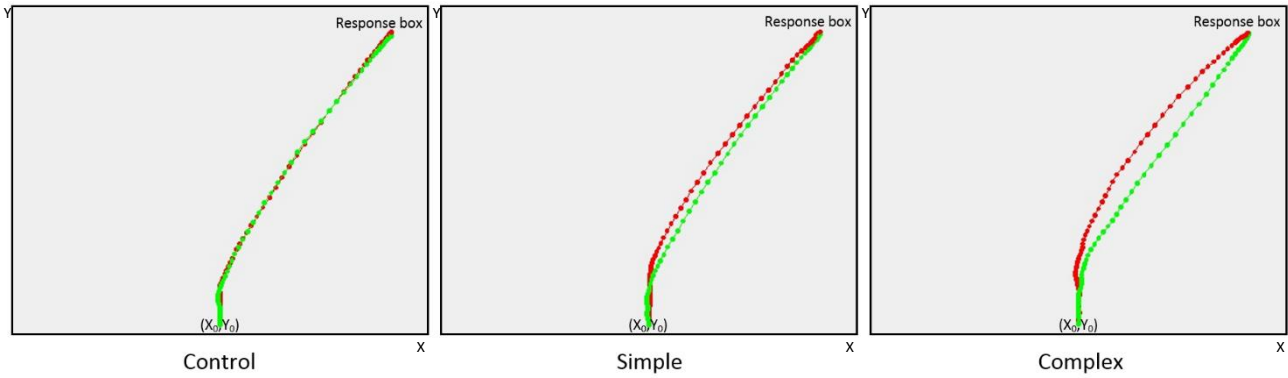


Figure 4.16: mouse trajectories for control questions (left figure), expected questions (central figure), and unexpected questions (right figure).

Focusing on complex stimuli, we split the trajectories of questions requiring a “yes” response from those requiring a “no” response (see Figure 4.17). The plot reveals that liars are more in trouble in responding complex questions requiring a “no” response compared to questions requiring a “yes” response. In fact, in the first response stage, they spend more time moving on y -axis, with a very erratic route. Then, they deviate toward the chosen response box with a wider curve respect to truth-tellers (AUC: liars $M=0.83$, $SD=0.71$ and truth-tellers $M=0.38$, $SD=0.73$; MD: liars $M=0.38$, $SD=0.28$ and truth-tellers $M=0.18$, $SD=0.29$). These observations are in line with the results obtained by the experiment 4.4, where we highlighted that liars showed greater RT in responding to complex “no” questions compared to complex “yes” questions.

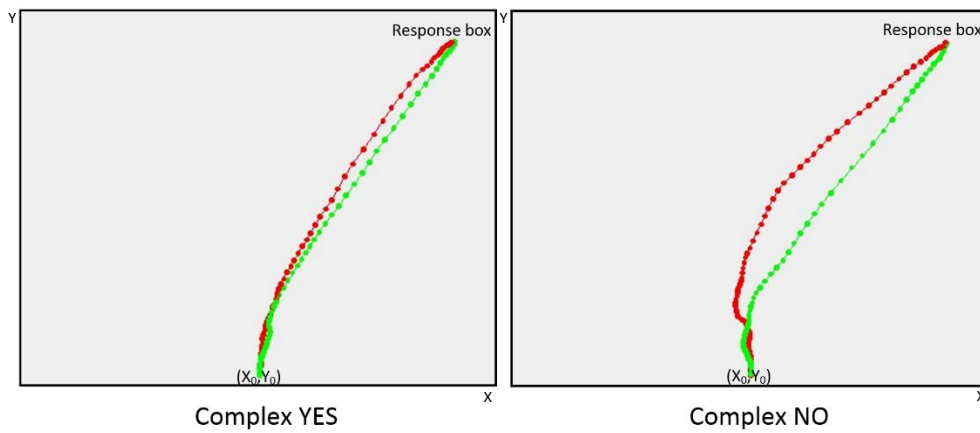


Figure 4.17: mouse trajectories for complex “yes” (left figure) and complex “no” questions (right figure).

Finally, taking into account only complex questions that required a “no” response, we analyzed the position of the mouse along the x and y -axis during the time, in search of time-points of maximum difference between truth-tellers and liars trajectories. As shown by Figure 4.18, the two groups had a maximum difference in the first half of the trajectory along the y -axis, and in the last part of the trajectory along the x -axis. We identified as representative points of maximum separation the time frames Y10, Y21 and X75, X80.

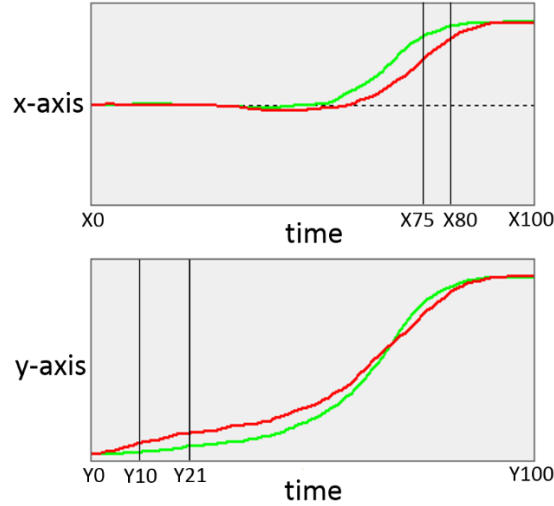


Figure 4.18: the figure reports the points of maximum difference between trajectories of truth-tellers and liars on x and y -axis over the time for responses to complex “no” questions.

In order to do not miss this information, we have considered the points of maximum deviation (Y10, Y21, X75, X80) as features for ML models.

4.5.6 Feature selection

According to the conclusions drawn from the analysis of trajectories, we decided to reduce the number of predictors to enter in the feature selection, considering only predictors related to complex “no” questions.

Then, in order to remove redundant and irrelevant features and to select those that improved the models accuracy and generalization, a correlation based feature selection was run (for an explanation about CFS algorithm see section 3.4.1). A total number of 24 predictors were entered into the correlation analysis. The following features were selected: errors ($r_{pb} = 0.43$), X75 ($r_{pb} = 0.69$), minimum velocity on x -axis ($r_{pb} = 0.36$), minimum acceleration on y -axis ($r_{pb} = 0.39$). In Table 4.47, the correlation matrix between features is reported, as well as the correlation value between the dependent and independent variables (r_{pb}).

	Errors	X75	Min $v_{(x)}$	Min $a_{(y)}$	Condition
Errors	1.00	0.48	0.58	0.47	0.43
X75	0.48	1.00	0.23	0.32	0.69
Min $v_{(x)}$	0.58	0.23	1.00	0.83	0.36
Min $a_{(y)}$	0.47	0.32	0.83	1.00	0.39
Condition	0.43	0.69	0.36	0.39	1.00

Table 4.47: the table reports the correlation matrix for the four features that were selected and their correlation value with the dependent variable. It should be remembered that these features are related to the responses to complex “no” questions.

4.5.7 Descriptive statistics

Feature selection isolated, from the original set of 24 predictors, four independent variables: errors, min $v_{(x)}$, min $a_{(y)}$, and X75. Table 4.48 reports the descriptive statistics for these features, as well as the analysis of the difference between truth-tellers and liars (t -test, Cohen’s d).

Feature	Group	M (SD)	<i>t</i> -test (<i>t</i> , <i>p</i> -value)	Cohen's <i>d</i>
Errors	Liars	0.08 (0.09)	2.98, < .01	0.94
	Truth-tellers	0.01 (0.03)		
Min $v(x)$	Liars	-0.07 (0.06)	3.02, < .01	0.95
	Truth-tellers	-0.04 (0.04)		
Min $a(y)$	Liars	-0.14 (0.06)	2.59, < .05	0.82
	Truth-tellers	-0.09 (0.03)		
X75	Liars	0.35 (0.16)	5.86, < .01	1.85
	Truth-tellers	0.63 (0.15)		

Table 4.48: descriptive statistics for errors, min $v(x)$, min $a(y)$, and X75. Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers). The last three columns report statistics about the difference between the two groups. In particular, the table shows the results of the independent *t*-test (*t*-value and *p*-value are reported) and effect-size (Cohen's *d*). For the interpretation of Cohen's *d*, see paragraph 3.4.2.

4.5.8 Machine learning models

Five ML classifiers (logistic, SVM, Naïve Bayes, random forest and LMT) were trained using a 10-fold cross-validation procedure (see section 3.4.3). Classification results are reported in Table 4.49. Accuracies range from 70% to 90%, with random forest reaching the maximum accuracy (92.5%) and Naïve Bayes the minimum performance (72.5%).

ML classifier	10-fold cross-validation		
	Average accuracy (SD)	Precision	Recall
Logistic	80% (19.7)	0.833	0.750
SVM	85% (12.9)	0.938	0.750
Naïve Bayes	72.5% (24.9)	0.765	0.650
Random forest	92.5% (12.1)	0.905	0.950
LMT	87.5% (13.2)	0.857	0.900

Table 4.49: the table reports the accuracy obtained by five classifiers in correctly identify fakers and truth-tellers, in 10-fold cross-validation. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

About the rate of false positive and false negative, the confusion matrix showed that the number of liars and truth-tellers misclassified in cross-validation is not equal for all the algorithms. Logistic regression failed in detecting three truth-tellers and five liars. SVM was unbalanced toward the false negative, as it misclassified five liars and only one truth-teller. Naïve Bayes misclassified four truth-tellers and seven liars. Random forest failed the classification of one liar and two truth-tellers, whereas LMT misclassified two liars and three truth-tellers.

4.5.9 Countermeasures and alternative efficient models

As reported in section 4.1.11, mouse tracking permits to collect numerous indices, which a human being cannot keep simultaneously under control. Moreover, the broad range of predictors allows building alternative classification models. In other words, even if the subject knows in advance which indices will be recorded during the test, he cannot know which of them will be used to predict the outcome.

We did not directly tested resistance to countermeasures of this technique, but we have developed alternative machine learning models entering subsets of predictors different from that used above. A

first new subset of predictors has been selected taking out the four features that are more correlated with the dependent variable. These are X75 ($r_{pb} = 0.69$), X80 ($r_{pb} = 0.54$), maximum acceleration on x -axis ($r_{pb} = 0.53$) and RT ($r_{pb} = 0.49$). A second subset of predictors has been chosen considering the features related to the amplitude of the trajectories: MD ($r_{pb} = 0.33$), AUC ($r_{pb} = 0.30$), MD-time ($r_{pb} = 0.41$). In the third set, we entered only features related to X and Y time frames: X75 ($r_{pb} = 0.69$), X80 ($r_{pb} = 0.54$), Y10 ($r_{pb} = 0.23$), Y21 ($r_{pb} = 0.27$).

Results obtained in the 10-fold cross-validation using this three new set of predictors are reported in Table 4.50.

Predictors	ML classifier	Training set (10-fold cross-validation)		
		Average accuracy (SD)	Precision	Recall
X75, X80, Max $a(x)$, RT	Logistic	82.5% (20.6)	0.842	0.800
	SVM	82.5% (20.6)	0.882	0.750
	Naïve Bayes	80% (19.7)	0.833	0.750
	Random forest	90% (12.9)	0.864	0.950
	LMT	85% (12.9)	0.850	0.850
MD, AUC, MD-time	Logistic	77.5% (21.9)	0.700	0.757
	SVM	72.5% (24.9)	0.800	0.600
	Naïve Bayes	77.5% (21.9)	0.824	0.700
	Random forest	72.5% (14.2)	0.714	0.750
	LMT	80% (15.8)	0.833	0.750
X75, X80, Y10, Y21	Logistic	82.5% (16.9)	0.882	0.750
	SVM	77.5% (18.4)	0.867	0.650
	Naïve Bayes	75% (20.4)	0.857	0.600
	Random forest	90% (12.9)	0.864	0.950
	LMT	80% (15.8)	0.833	0.750

Table 4.50: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in 10-fold cross-validation, using different set of predictors. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.5.10 Discussion

In this experiment, we replicated the experimental procedure reported in section 4.1, using complex questions instead of unexpected questions to increase liars' cognitive load. Results indicated that complex questions are efficient in discriminating liars and truth-tellers, with a slightly lower accuracy comparing to unexpected questions. In fact, using an equal number of participants as training sample and running the same classification algorithms, we have obtained accuracies ranging from 90% to 70% in the 10-fold cross-validation, whereas the accuracies that were reported in section 4.1.8 range from 90% to 95%. Although, an accuracy around 90% is not suitable for applications in the field of justice, it may be enough for screening applications (e.g., the detection of online deception).

An interesting result concerns the evidence that liars and truth-tellers differ in mouse dynamics parameters only for complex questions that required a "no" response, or rather for complex questions that contain at least one information that is incoherent with the lie they told. In other words, liars need

greater cognitive resources to identify one or more discrepancies with the lie they told, whereas they are skilled like truth-tellers in confirming their lie. Probably, it may be because the verification process [187] (the careful monitoring of the congruence between the various information provided during the production of the lie) is cognitively heavier for negative responses. This result has been confirmed also by the experiment reported in section 4.4, where liars showed higher RT than truth-tellers in complex “no” questions.

4.6 The Detection of False Autobiographical Events with Complex Questions and Mouse Dynamics

The aim of this experiment is to verify whether the technique of complex questions and mouse dynamics is generalizable to other topic, different from deception about identity. In particular, in this experiment we have applied complex questions and mouse dynamics recording to detect people who reported false autobiographical events. To test subjects on comparable autobiographical events, we have chosen the last holiday of each participant as target event [109]. For a person, recalling a holiday is not so much different from telling his alibi during a criminal investigation or reviewing on Yelp his experience at the restaurant.

When a subject decides to lie, he fabricates a completely invented experience or he inserts false details into a story with a kernel of truth. For this reason, we ran two different experiments collecting two different groups of liars. The first group was asked to learn a completely faked holiday, that was created by the experimenter for them and then, they were instructed to complete the experimental task pretending to have experienced this holiday. Liars in the second group were asked to narrate their last holiday, which was changed in some details (e.g., the name of the friends they were on vacation) by the experimenter and finally sent back to them. The subjects' task was to learn their modified autobiographical experience as they would have been tested about it during the experimental task.

4.6.1 Participants

A first sample of sixty participants was recruited and data were used as training set to build ML models. Twenty participants were assigned to the truth-teller condition, twenty were liars with a totally invented holiday (liars A) and the other twenty were liars telling an actual holiday with faked details (liars B). The demographic characteristics of the sample are reported in Table 4.51.

Then, a second sample of 30 participants (10 truth-tellers, 10 liars A and 10 liars B) was collected and used as test set to assess the models generalization. Demographic information about participants are in Table 4.51.

Sample	Condition	N	Gender	Age	Education
Training set	Truth-tellers	20	M = 8, F= 12	M = 22, SD = 2.3	M = 16, SD = 1.8
	Liars A	20	M = 7, F= 13	M = 22, SD = 1.6	M = 16, SD = 1.4
	Liars B	20	M = 10, F= 10	M = 23, SD = 2.4	M = 17, SD = 1.8
Test set	Truth-tellers	10	M = 6, F= 4	M = 22, SD = 2.6	M = 16, SD = 1.9
	Liars A	10	M = 2, F= 8	M = 20, SD = 2.1	M = 15, SD = 1.8
	Liars B	10	M = 4, F= 6	M = 22, SD = 3.1	M = 16, SD = 2.2

Table 4.31: demographic information about training and test set. In the second column (N) the number of participants for each sample is reported. The third column shows the number of male and female in each sample. The fourth and the fifth columns report mean (M) and standard deviation (SD) of participants' age and education.

4.6.2 Experimental procedure

Participants assigned to truth-tellers' group were contacted the day before the experiment and were asked to provide a holiday experience (min 2 days, maximum 7 days) happened in the last year, or a year and half at most. They received an outline that they had to fill with details regarding their vacation: where they went, how they travelled, when they left, how long they stayed there, with whom they went, how was the weather like and, like a travelogue, what they did on a daily basis. Figure

4.19 represents the outline that truth-tellers were asked to compile with the details of their vacation and provide to the experimenter. Participants were suggested to omit details if unable to remember them. Furthermore, they were suggested to help themselves with tools such as photographs, videos, real witnesses just to be sure that what they were saying was correct. According to the information that they provided, the experimenter created the stimuli for the computerized task. The day after, truth-teller participants were invited to the lab to complete the experiment. First, they were asked to recall twice their vacation. The examiner verified the correctness of the information and rectified any errors. All participants have recalled the holiday information correctly within the second recall. Between the two recalls, they were required to perform some mental arithmetic as distracting task. Finally, they were instructed to complete the experimental task, responding to any questions according to their vacation experience.

Participants assigned to liar A condition were contacted the day before the experiment and were asked to learn a faked holiday experience created by the experimenter. We told them that the day after they would be tested about what they had learnt. The faked holiday consisted of the same details mentioned above for truth-tellers: where, how, when, how long, with whom, how was the weather like, daily activities (see Figure 4.19). An example of holiday created for the experiment and sent to liars A for learning is reported in Figure 4.20. The day after, participants were invited to the lab to complete the experiment. First, they were asked to repeat twice the assigned holiday experience, explaining it in first person and pretending it had really happened. The examiner verified the correctness of the information and rectified any errors. All participants have recalled the holiday information correctly within the second recall. Between the two recalls, they were required to perform some mental arithmetic as distracting task. Finally, they were instructed to complete the experimental task, responding to any questions pretending to have done the vacation experience previously learned.

Participants assigned to liar B group were contacted the day before the experiment and, similar to truth-tellers, were asked to provide a holiday experience (min 2 days, maximum 7 days) happened in the last year, or a year and half at most. They received the same outline of truth-tellers (see Figure 4.19) that they had to fill with details regarding their vacation: where they went, how they travelled, when they left, how long they stayed there, with whom they went, how was the weather like and, like a travelogue, what they did on a daily basis. Once filled the form, all test takers had to send it back. According to the information that they provided, the experimenter created the stimuli for the computerized task. In particular, starting from the real vacation experience of each participant, some holiday details were modified, following this scheme:

- Keep the destination.
- Alter the length of the stay.
- Alter the place where they stayed.
- Alter the names of companion/companions.
- Alter the atmospheric conditions: main importance was given to the rain. In case it had really rained during the real holiday experience, the day it happened was switched. Otherwise, if it hadn't rained at all, a rainy day was inserted in the program.

- Alter the order of activities: they were inverted, so if something had happened on day 2, this was changed to day 4 and vice versa.
- Insert brand new activities and omit some that really happened.

GUIDELINE QUESTIONS	PLEASE, COMPLETE WITH YOUR EXPERIENCES OF JOURNEY
Where did I go?	
How long did I stay?	
How did I go?	
With whom did I go?	
Where did I sleep?	
How was the weather like?	
Day 1	
Day 2	
Day 3	
Day 4	
Day 5	
Day 6	
Day 7	

Figure 4.19: the Figure shows the outline table that truth-teller participants were asked to fill with the information related to their last holiday. It contains the following details about the holiday: where they went, how they travelled, when they left, how long they stayed there, with whom they went, how was the weather like and what they did on a daily basis.

After this procedure, the vacation outline with the modified information was sent back to the participant, who were requested to memorize the new faked details of his actual holiday. We told him that the day after he would be tested about what he had learnt. The day after, participants were invited to the lab to complete the experiment. First, they were asked to repeat twice the modified holiday experience, pretending that all details are true. The examiner verified the correctness of the information and rectified any errors. All participants have recalled the holiday information correctly within the second recall. Between the two recalls, they were required to perform some mental arithmetic as distracting task. Finally, they were instructed to complete the experimental task, responding to any questions pretending to have done the vacation experience as previously learned.

ASSIGNMENT: I ask you kindly if you can learn this short story. Learn it in the best and accurate way you can. If you have any doubts or misunderstandings about it, do not hesitate to contact me. You can also consult the internet or do some research for what you think it is necessary to know, in lack of personal experience.

ANTWERP BRUXELLES

“In November 2015 my friends and I decided to reach our friends in Antwerp. I took the plane in Bergamo and landed in Bruxelles, there I took the train to go to Antwerp. We left in four: Sabrina, Miriana, Alessandra and I to visit our friend Francesca who was there for the Erasmus programme. We stayed there two nights. Sabrina and I stayed in a hostel, far away from Francesca’s house. To get to her house we took the bus. Miriana and Alessandra instead stayed at Francesca’s house.

On the first day, we arrived in Bruxelles in the early afternoon; we had dinner and we walked around Bruxelles city-centre. Unfortunately, it was raining a lot, so we did not do anything special. In the evening, we met some Erasmus friends of Francesca and we went out altogether.

The day after (second day), we had a tour in Antwerp, the weather was better and we went to visit the Aan De Stroom museum, a 10 floor building with a beautiful panoramic terrace from where we could see the whole city. The museum is famous for its panoramic view. In the evening, we went to Francesca’s and we made a cocktail party with her friends. Being Italians, we cooked pasta. I actually can say that we were forced to do that! (That’s hilarious!).

Ah, I forgot! After the panoramic museum we went to visit a doorway of a underground street which links Antwerp to another city, called St. Anna’s tunnel, which was great and awesome because of the mosaic at the beginning!”

Figure 4.20: an example of faked holiday that liars were asked to learn. It contains the following details: where they went, how they travelled, when they left, how long they stayed there, with whom they went, how was the weather like and what they did on a daily basis.

The experimental task was computerized and consisted in responding 60 yes/no questions clicking with the mouse on one of the two alternative response labels. For more details about the modalities of presentation of the stimuli, see paragraph 3.3.2.

4.6.3 Stimuli

A total number of 60 questions were presented to each subject. Thirty sentences required a “yes” response, and 30 required to respond “no”. Questions were the same for both liars (A and B) and truth-tellers. The 60 experimental questions were preceded by 6 training questions (3 requiring a “yes” response and 3 requiring a “no” response). Questions included in the experimental task belonged to the following categories:

- 20 control questions. Control questions were sentences to which both truth-tellers and liars had to respond truthfully. These sentences were unrelated to holiday and referred to the experimental condition. Half of the control sentences required a “yes” response (e.g., “I am sitting in front of a computer”) and half a “no” response (e.g., “I am climbing a mountain”). Both liars and truth tellers were required to respond truthfully to all control questions.
- 20 simple questions. Simple questions were related to the holiday. Truth-tellers responded were asked to respond according to their real holiday, whereas liars A and B responded according to the faked holiday previously learned. Half of the simple sentences required to respond “yes” (e.g.,

“I was in France”) and the others required a “no” response (e.g., “I was in Japan”). For liars, a “yes” response corresponded to a lie.

- 20 complex questions. Complex questions were sentences that included two or three information about the holiday (e.g., “In October I was in France with my boyfriend”). Participants were instructed to respond “yes” when all of the information in the sentence was true according to the holiday, whereas they responded “no” when at least one of the information included in the sentence was false. In other words, participants had to respond “yes” when the entire sentence was true, and “no” when there was one or more pieces of false information in the sentence. Complex sentences could be composed by two or three holiday information. The false information was always in the last place, to avoid that subjects would not complete to read the phrase. An example of questions is reported in Table 4.52.

To sum up, both liars and truth-tellers responded to 30 control, simple and complex questions that required to respond “yes” and to 30 control, simple and complex questions that required “no” responses. Control, simple and complex questions were presented randomly and intermixed. For more details about the modalities of presentation of the stimuli, see paragraph 3.3.1.

It should be noted that liars told lies only in the simple “yes” and complex “yes” responses. In fact, for the liars, the simple and complex questions regarding their faked holiday were actually “no” responses that, because they were lying, required “yes” responses. In all of other questions (control “yes”, control “no”, simple “no”, complex “no”), both liars and truth-tellers responded truthfully.

4.6.4 Collected measures

During the subjects’ response, mouse dynamics were recorded by MouseTracker software. For more details about data collection, see section 3.3.2.

In addition to the errors and the spatial-temporal features extracted by default by the MouseTracker software (IT, RT, MD, AUC, MD-time, x-flip, y-flip, see paragraph 3.3.2), we calculated the minimum, maximum and average velocity and acceleration along the x and y -axis during the response.

For each feature, we calculated the average and the standard deviation value of all the stimuli and, then, separately for control, simple, complex sentences and control yes, control no, simple yes, simple no, complex yes and complex no sentences.

We also analyzed the position of the mouse along the x and y -axis in search of points of maximum difference between the trajectories of liars and truth-tellers. Comparing truth-tellers and liars A, we identified X75 and Y30 as the points of maximum deviation, respectively on x and y -axis. On the other hand, comparing truth-tellers and liars B, the point of maximum deviation resulted to be X75 and Y45. Finally, we computed the average value of these points for control, simple and complex questions.

In this way, we obtained a final list of 326 predictors. The complete list of predictors is reported in Annex 4.

Type of question	Question that requires “yes” response by both liars and truth-tellers	Question that requires “no” response by both liars and truth-tellers
Control	I’m in front of a computer I’m in front of a monitor I’m using a computer I’m answering with the mouse I’m sitting on a chair In this moment I’m sit I’m doing a test of psychology I’m doing a test I’m reading some sentences I’m answering to some sentences	I’m climbing a mountain I’m in the mountain to climb I’m on an airplane I’m sitting on a plane I’m on a beach I’m sunbathing I’m eating in a restaurant I’m having a lunch in a restaurant I’m playing football I’m having a football play
Simple	I went to Antwerp In November I did an holiday I went by airplane I stayed in an hostel I saw the Aan de Stroom museum I stayed there 3 days We left in four The public transport were closed for attacks The first day it rained the second one no I saw a really beautiful mosaic	I went to Frankfurt In December I did an holiday I went by train I stayed in an hotel I saw the A plan de corones museum I stayed there 5 days We left in six The public transport were closed for attacks The second day it rained the first one no I saw an art contemporary museum
Complex	I went to <u>Antwerp</u> with <u>three friends</u> in <u>November</u> At <u>Antwerp</u> I saw the <u>Aan de Stroom museum</u> The <u>first day it rained</u> and <u>after eating</u> we have a walk <u>Sabrina and I</u> went to sleep in an <u>hostel</u> far away from <u>Francesca’s house</u> The <u>second day in Antwerp</u> we saw the <u>Aan de Stroom museum</u> In <u>November</u> I visited less <u>Brussels</u> for terrorist attacks <u>With my friends</u> the <u>second day</u> we saw the <u>Saint Anne tunnel</u> <u>By plane</u> I stayed <u>three days</u> at <u>Antwerp-Brussels</u> The <u>first day we walked</u> the <u>second we saw a museum</u> At <u>Antwerp</u> the <u>second day</u> we saw the <u>Saint Anne tunnel</u>	I stayed at <u>Antwerp</u> with two friends in <u>November</u> In Brussels I saw the <u>Aan de Stroom museum</u> The first day we saw <u>the Aan de Stroom museum</u> Sabrina and Alessandra went to sleep to <u>Francesca’s house</u> in <u>Brussels</u> The first day in Brussels we saw <u>the Aan de Stroom museum</u> In <u>November</u> I visited a few Antwerp for attacks <u>With plane</u> we went 5 days to <u>Brussels</u> With a friend the <u>third day</u> we saw <u>the Saint Anne tunnel</u> The <u>first day we walked</u> and the third we visited a museum In Brussels the <u>first day</u> I saw <u>the Saint Anne tunnel</u>

Table 4.52: the table reports an example of the 60 control, simple and complex questions presented to participants and related to a truth or faked holiday. Information about holiday in complex questions are underlined. In bold the false holiday information according to which participants had to respond “no”.

4.6.5 Analysis of trajectories

A visual analysis was conducted to compare the truth-tellers’ mouse trajectories with liars A and liars B motor response. Trajectories were plotted separately for control, simple and complex questions (see figure 4.21). Comparing truth-tellers and liars A, it can be seen that the two experimental groups have mostly overlapping trajectories for control questions, whereas they differ in simple and complex questions. In these questions, truth-tellers show straight trajectories from the origin to the response box, whereas liars show wider trajectories, characterized by a greater AUC and MD. The same observations can be drawn matching truth-tellers and liars B trajectories, though a little difference in curves of the two groups can be noticed in control questions as well.

Finally, we plotted separately questions that required “yes” responses and questions requiring “no” responses. Figures are reported in Annex 5. In brief, liars A show the same response pattern that we found in the previous experiment about identity deception: they are more in trouble in responding to “no” questions respect to “yes” questions. This effect is evident both in simple and complex questions. On the other hand, liars B show very wide trajectories in simple and complex questions requiring both “yes” and “no” responses.

In conclusion, the visual pattern that we observed is in line to that obtained in the previous experiment (see sections 4.5), but the effect of deception on motor response seems to extend also to simple questions. Once again, it seems that liar participants had more difficulty in responding “no” to simple and complex questions respect to responding “yes”.

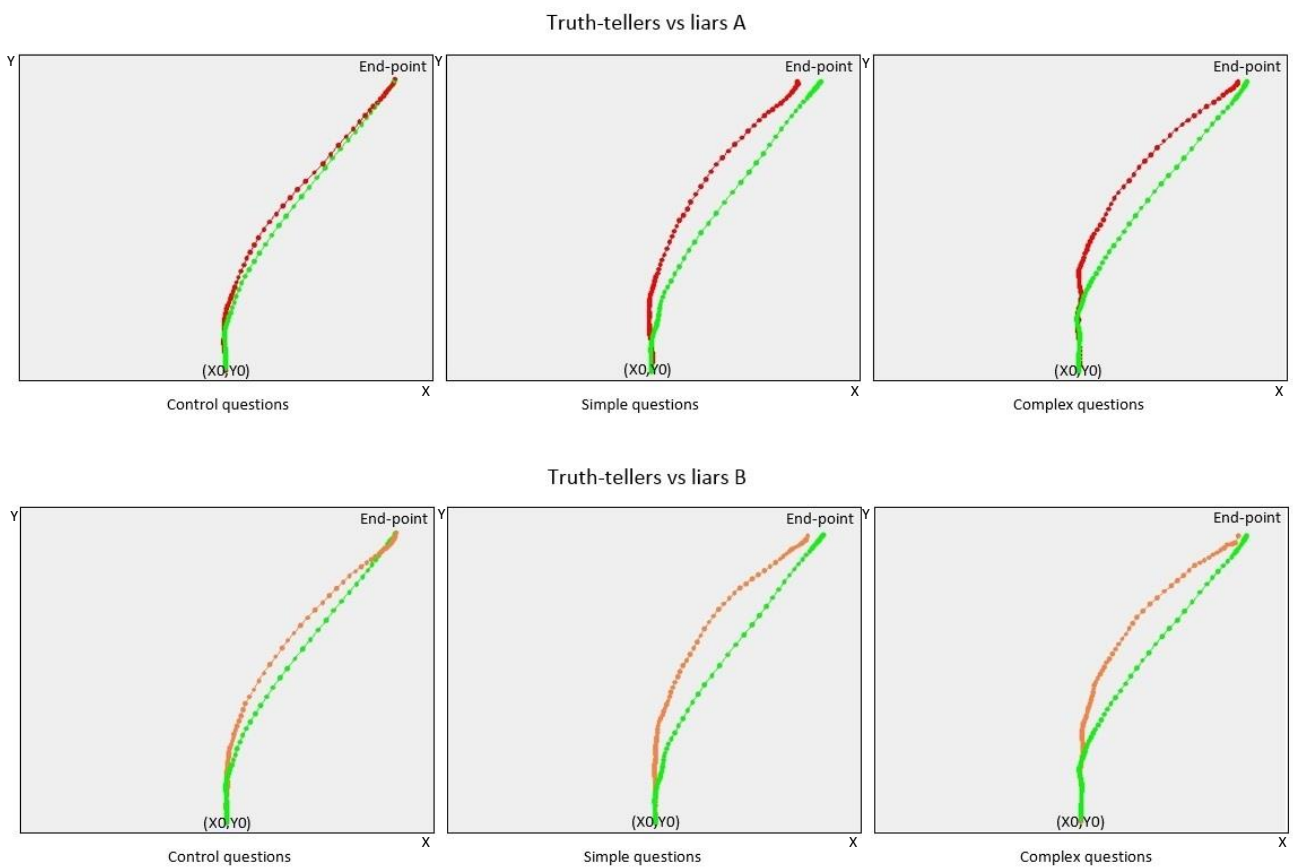


Figure 4.21: mouse trajectories for control questions (left figure), simple questions (central figure), and complex questions (right figure). The first figure compares trajectory of truth-tellers (in green) and liars A (in red). The second figure compares trajectories of truth-tellers and liars B (in orange).

4.6.6 Feature selection

In order to remove redundant and irrelevant features and to select those that improved the models accuracy and generalization, a correlation based feature selection was run (for an explanation about CFS algorithm see section 3.4.1).

We decided to enter in CFS algorithm all the predictors, as the visual analysis has highlighted differences between groups both in simple and complex questions, and also in control questions for truth-tellers compare with liars B.

Two separate features selection were performed, the first one pairing truth-tellers and liars A, and the second pairing truth-tellers and liars B. Tables 4.53 and 4.54 report the list of the features selected and the correlation value with the dependent variable (r_{pb}).

Feature	Group	M (SD)	r_{pb}
Simple_average_vel(x)	Liars A	0.007 (0.001)	0.70
	Truth-tellers	0.008 (0.0003)	
Simple_NO_average_vel(x)	Liars A	0.006 (0.001)	0.69
	Truth-tellers	0.008 (0.0005)	
All questions_average_vel(x)	Liars A	0.007 (0.0003)	0.65
	Truth-tellers	0.008 (0.0003)	
Simple_X75	Liars A	0.495 (0.151)	0.57
	Truth-tellers	0.676 (0.102)	
Simple_NO_min_vel(x)	Liars A	-0.046 (0.039)	0.49
	Truth-tellers	-0.013 (0.016)	
Complex_average_MD-time	Liars A	2285.50 (424.92)	0.46
	Truth-tellers	1898.47 (308.22)	
Simple_min_vel(x)	Liars A	-0.020 (0.019)	0.44
	Truth-tellers	-0.006 (0.006)	
Complex_NO_min_vel(x)	Liars A	-0.032 (0.026)	0.44
	Truth-tellers	-0.014 (0.009)	
Complex_NO_average_RT	Liars A	3486.20 (661.52)	0.43
	Truth-tellers	2888.57 (627.48)	
All questions_min_vel(x)	Liars A	-0.008 (0.006)	0.42
	Truth-tellers	-0.003 (0.002)	
Complex_X75	Liars A	0.264 (0.202)	0.42
	Truth-tellers	0.431 (0.152)	
Simple_NO_average_RT	Liars A	2696.61 (808.56)	0.41
	Truth-tellers	2107.32 (473.58)	
Simple_sd_acc(y)	Liars A	0.006 (0.003)	0.39
	Truth-tellers	0.004 (0.001)	
Simple_average_MD-time	Liars A	1457.95 (366.23)	0.33
	Truth-tellers	1240.20 (249.40)	

Table 4.53: the table reports the features that were selected and their correlation value (r_{pb}) with the dependent variable, taking into account truth-tellers and liars A. In addition, Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers).

4.6.7 Descriptive statistics

Tables 4.53 and 4.54 report average and the standard deviation of each selected feature, for the three experimental groups (truth-tellers, liars A and liars B). The correlation between each predictor and the dependent variable is reported as well.

Experiments

Feature	Group	M (SD)	r_{pb}
Simple_average_vel(x)	Liars B	0.007 (0.0008)	0.54
	Truth-tellers	0.008 (0.0003)	
Complex_NO_average_MD	Liars B	0.442 (0.228)	0.52
	Truth-tellers	0.208 (0.166)	
Simple_sd_vel(y)	Liars B	0.014 (0.003)	0.50
	Truth-tellers	0.018 (0.004)	
Complex_NO_average_y-flip	Liars B	7.51 (1.94)	0.50
	Truth-tellers	5.46 (1.65)	
All questions_average_vel(x)	Liars B	0.007 (0.0004)	0.49
	Truth-tellers	0.008 (0.0003)	
All questions_min_vel(x)	Liars B	-0.007 (0.004)	0.47
	Truth-tellers	-0.003 (0.002)	
All questions_sd_vel(x)	Liars B	0.011 (0.001)	0.46
	Truth-tellers	0.013 (0.002)	
Simple_X75	Liars B	0.527 (0.175)	0.46
	Truth-tellers	0.676 (0.103)	
Complex_sd_AUC	Liars B	1.070 (0.328)	0.45
	Truth-tellers	0.79 (0.453)	
Simple_NO_average_IT	Liars B	337.37 (204.46)	0.43
	Truth-tellers	551.22 (234.38)	
Complex_average_MD-time	Liars B	2267.04 (447.45)	0.43
	Truth-tellers	1902.25 (303.53)	
Complex_X75	Liars B	0.305 (0.126)	0.42
	Truth-tellers	0.431 (0.153)	
Control_YES_average_y-flip	Liars B	7.45 1.66)	0.41
	Truth-tellers	6.21 (1.06)	
Complex_average_x-flip	Liars B	8.67 (2.50)	0.40
	Truth-tellers	6.59 (2.28)	
Control_average_y-flip	Liars B	7.31 (1.42)	0.38
	Truth-tellers	6.26 (1.09)	
Control_NO_average_RT	Liars B	1947.31 (234.67)	0.38
	Truth-tellers	1739.57 (264.82)	
Simple_YES_average_y-flip	Liars B	7.34 (1.88)	0.38
	Truth-tellers	6.16 (0.86)	
All questions_max_acc(x)	Liars B	0.019 (0.013)	0.37
	Truth-tellers	0.011 (0.005)	
All questions_average_error	Liars B	0.05 (0.04)	0.37
	Truth-tellers	0.02 (0.03)	
Simple_YES_max_vel(y)	Liars B	0.06 (0.02)	0.37
	Truth-tellers	0.08 (0.01)	
Complex_YES_ds_AUC	Liars B	0.944 (0.523)	0.37
	Truth-tellers	0.547 (0.49)	
Simple_NO_sd_MD	Liars B	0.408 (0.14)	0.34
	Truth-tellers	0.289 (0.186)	
Simple_NO_sd_y-flip	Liars B	3.54 (0.89)	0.33
	Truth-tellers	2.84 (1.06)	
Control_sd_vel(y)	Liars B	0.016 (0.005)	0.31
	Truth-tellers	0.018 (0.003)	
All questions_sd_acc(y)	Liars B	0.005 (0.005)	0.24
	Truth-tellers	0.003 (0.001)	

Table 4.54: the table reports the features that were selected and their correlation value (r_{pb}) with the dependent variable, taking into account truth-tellers and liars B. In addition, Mean (M) and standard deviation (SD) are reported for the two experimental groups (liars and truth-tellers).

4.6.8 Machine learning models

The selected features were used to train different ML classifiers (logistic regression, SVM, LMT, random forest). We trained separate model for truth-tellers vs liars A and truth-tellers vs liars B. Models were evaluated following a 10-fold cross-validation procedure, as described in paragraph 3.4.3. Then, we tested the generalization of the models performance on the new set of 20 participants (see section 4.1.1). Results obtained are reported in Table 4.55 and Table 4.56. Comparing truth-tellers with liars A, accuracies ranged from 75% to 87.5% in 10-fold cross-validation and from 65% to 85% in the test set. As regards the classification of liars B, classification accuracies ranged from 80% to 87.5% in 10-fold cross-validation and from 75% to 80% in the test set..

ML classifier	Training set - truth-tellers vs liars A (10-fold cross-validation)			Test set - truth-tellers vs liars A		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	75% (15.8)	0.813	0.650	80%	0.750	0.900
SVM	77.5% (17.5)	0.923	0.600	65%	0.667	0.600
Naïve Bayes	87.5% (17.5)	0.895	0.850	75%	0.669	0.900
Random forest	82.5% (12.1)	0.842	0.800	85%	0.889	0.800
LMT	77.5% (12.1)	0.789	0.750	70%	0.833	0.500

Table 4.55: the table reports the accuracy obtained by five classifiers in correctly identify liars A and truth-tellers, in training and test set. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

ML classifier	Training set - truth-tellers vs liars B (10-fold cross-validation)			Test set - truth-tellers vs liars B		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	82.5% (15.8)	0.783	0.900	75%	0.778	0.700
SVM	85% (17.5)	0.850	0.850	80%	0.800	0.800
Naïve Bayes	87.5% (17.5)	0.941	0.800	75%	0.778	0.700
Random forest	87.5% (12.1)	0.857	0.900	80%	0.800	0.800
LMT	80% (12.1)	0.800	0.800	75%	0.692	0.900

Table 4.56: the table reports the accuracy obtained by five classifiers in correctly identify liars B and truth-tellers, in training and test set. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

About the rate of false positive and false negative, the confusion matrix showed that the number of liars and truth-tellers misclassified in cross-validation and test set are not equal for all the algorithms. So, there is not a general unbalancement of the classification error to one or the other class, but each classifier has peculiar displacement.

4.6.9 Discussion

This experiment demonstrated that the technique of complex questions along with mouse dynamic recording is efficient in detecting liars not only on identity but also on autobiographical events. In other words, it has been shown that the use of this technique may be extended to different topic. Accuracies are very similar to those obtained in the previous experiment (see paragraph 4.5.8), which was aimed to detect deception about identity. This is true especially for the classification of truth-tellers B, whereas liars A were identify with lower accuracy.

It is important to notice that liars B are better detectable than liars A. It means that it is simpler to recall a completely invented lie than a partial lie. I may be assumed that in the second case (liars B) the true elements of the memory interpose with the fabricated details and the subject need more cognitive resources to monitor the coherence of the recall.

4.7 Can mouse dynamics and complex questions generalize from a topic to another?

The final step of this work is to verify whether it is possible to build lie detection model based on mouse dynamics, which is independent from the topic of the deception. In other words, the aim of the following experiment is to prove that the models that we built, using data from complex questions and mouse dynamics, can generalize their performance from a topic to another.

To argue this point, we re-run fine classification algorithms (logistic regression, SVM, Naïve Bayes, random forest and LMT) using the data of the experiment in section 4.5 as training set and the data of the experiment in section 4.6 as test set, and vice versa. In other words, we classified subjects lying on holidays using the model built by subjects lying on identity and, then we sorted subjects lying on identity using the model built by subjects lying on holidays.

It is worth recalling that in the experiment that was described in section 4.5, twenty participants lied about their identity and twenty participants responded to questions according to their true identity. In the experiment that was reported in section 4.6, the training set consisted of twenty participant that told an actual holiday, twenty participants that responded to questions according to a totally faked holiday (liars A) and other 30 participants that lied only on some vacation details (liars B). We chose to focus our attention on liars B, which are the most detectable.

To sum up, in the following experiment training and test sets consisted of 40 participants that lied about identity and 40 participants that lied about vacation details. Participants were asked to respond questions about their identity or about their last vacation clicking with the mouse on the response button. Questions were of three different type: control, simple and complex, and required to respond “yes” or “no”. Mouse dynamics were recorded.

The key point of this step should be to identify the transverse indices of falsehood that remain constant in the different experiments. Summarizing the observation that emerged from the previous experiments, we can state that:

- 1) Generally, liars were more in trouble in responding to complex questions that required a “no” response. It was confirmed by experiments in section 4.4, 4.5 and 4.6.
- 2) In both the experiments collecting mouse dynamics (section 4.5 and 4.6), visual analysis revealed that in complex “no” questions liars draw with the mouse wider trajectories (greater MD and AUC) than truth-tellers.
- 3) In both the experiments recording mouse dynamics (section 4.5 and 4.6), X75 was the point of maximum separation between liars and truth-tellers on x -axis during the response.
- 4) Finally, experiment reported in section 4.4, as well as the experiments in sections 4.5 and 4.6, highlighted the importance of RT and errors in distinguishing the two groups. In fact, liars seems to be slower in their response and make more errors, especially in complex “no” questions.

In view of the above, we have identified a set of features that characterize the liar’s general trend of response: number of errors, RT, MD-time, average velocity on x -axis, MD, AUC, X75. Then, these features were entered in ML classifiers.

4.7.1 Detection of liars about autobiographical events using a model about identity

Here we predict the deception about an autobiographical event (holiday) using a model trained on participants who lied about identity.

The models were trained feeding the features mentioned above as input to classifiers: number of errors ($r_{pb} = 0.44$), RT ($r_{pb} = 0.49$), MD-time ($r_{pb} = 0.41$), average velocity on x -axis ($r_{pb} = 0.44$), MD ($r_{pb} = 0.34$), AUC ($r_{pb} = 0.30$), X75 ($r_{pb} = 0.69$). Then, models were tested on the same features extracted from the sample of participants lying about holidays. Results are reported in Table 4.57.

ML classifier	Training set – deception about identity (10-fold cross-validation)			Test set – deception about holiday		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	75% (16.7)	0.727	0.800	62.5%	0.692	0.450
SVM	77.5% (21.9)	0.730	0.850	70%	1.000	0.400
Naïve Bayes	67.5% (12.1)	0.652	0.750	62.5%	0.7271	0.400
Random forest	92.5% (16.9)	0.947	0.900	65%	0.875	0.350
LMT	80% (15.9)	0.800	0.800	67.5%	0.818	0.450

Table 4.57: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in 10-fold cross-validation and test set using. The training set consisted of 40 participants lying about identity and the test set consisted of 40 participants that lied about holiday. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10 folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.7.2 Detection of liars about identity using a model about autobiographical events

Here we predict the deception about identity using a model trained on participants who lied about an autobiographical event (holiday).

The models were trained feeding the features mentioned above as input to classifiers: number of errors ($r_{pb} = 0.30$), RT ($r_{pb} = 0.43$), MD-time ($r_{pb} = 0.43$), average velocity on x -axis ($r_{pb} = 0.31$), MD ($r_{pb} = 0.52$), AUC ($r_{pb} = 0.46$), X75 ($r_{pb} = 0.30$). Then, models were tested on the same features extracted from the sample of participants lying about identity. Results are reported in Table 4.58.

ML classifier	Training set – deception about holiday (10-fold cross-validation)			Test set – deception about identity		
	Average accuracy (SD)	Precision	Recall	Accuracy	Precision	Recall
Logistic	72.5% (27.5)	0.737	0.700	65%	0.600	0.900
SVM	75% (23.6)	0.708	0.850	67.5%	0.621	0.900
Naïve Bayes	70% (23.0)	0.667	0.800	75%	0.692	0.900
Random forest	57.5% (26.5)	0.588	0.500	70%	0.643	0.900
LMT	77.5% (18.4)	0.762	0.800	70%	0.643	0.900

Table 4.58: the table reports the accuracy obtained by five classifiers in correctly identify liars and truth-tellers, in 10-fold cross-validation and test set using. The training set consisted of 40 participants lying about holiday and the test set consisted of 40 participants that lied about identity. The accuracy in training set, using a 10-fold cross-validation, is the average accuracy resulting from the 10 folds. The standard deviation (SD) of the 10

folds is also reported. Recall is the percentage of liars who are correctly identified and precision represents the fraction of correct liars among those identified as liars.

4.7.3 Discussion

This experiment was aimed to demonstrate whether a classification model, which was trained on a sample of liars who responded to complex questions using mouse, could be generalized from a deception topic to another. Classification accuracies on test sets were higher than chance level (see paragraph 4.7.1 and 4.7.2), even though they did not exceed the 75%. The models that we trained on participants who lied about the holiday seem to be slightly more accurate in generalization than the models that were trained on participants lying on identity. To conclude, results are not amazing but demonstrated that more effort should be placed on the study of the subject's response when lying about different topics. One future direction should be the creation of a model that could be applied in a large number of cheating contexts, ranging from lies about identity in the real or online environment to security, forensic, or commercial applications.

Chapter 5 Conclusion

In this work, a series of proof of concept studies about lie detection via human-computer interaction have been presented. The experiments differ from each other for two main elements: the strategy that has been used to increase liars' cognitive load and the tool through which the participant interacted with the computer during the task. In particular, we applied two cognitive load strategies to bring out the liars' response distinctive features: the unexpected questions [146] and the complex questions [188]. Unexpected questions are questions to which participants were not explicitly trained to lie, whereas complex questions put together in the same sentence many information, each of which could be true or false. As regards the tools of user-pc interaction, three kind of measures were collected: RT on keyboard [151], keystroke dynamics [189] and mouse dynamics [121]. The experimental design was the same for all the experiment: we tested two independent groups, a group of liars and a group of truth-tellers, with the same experimental paradigm. Truth-tellers were asked to respond truthfully to all questions, whereas liars were instructed to lie on questions related to a specific topic according to some fake information previously learned. We have chosen to focus on identity as topic of deception for different reasons. First, the deception about identity is one of the most popular issue in the current historical and cultural frame, especially in the online environment [15]. Secondly, the use of one single topic has allowed to easily comparing the results obtained by different experimental paradigms. Then, to investigate the generalization of the techniques, a second topic, that is the deception about an autobiographical events, has been tested. Finally, for each experiment, we analyzed the data collected and, through machine learning algorithms, we developed classification models, which were trained to distinguished liars from truth-tellers. Experiments led to interesting results, which are summarized in the following section.

5.1 Main Results

The achieved results from the experiments can be summarized as follows:

- on the same topic (identity) the three techniques (RT, keystroke dynamics, mouse dynamics) give similar results in terms of accuracy. In fact, they classified subjects as liars or truth-tellers with maximum accuracies ranging from 92.5% to 97.5% (see Table 4.59). Mouse dynamics is the technique that reached the lower accuracy (92.5%-95%), followed by RT (95%-97.5%) and keystroke dynamics (97%).
- On the same topic (identity), unexpected questions are slightly more effective than complex questions to increase liars' cognitive load and stand out the distinctive features of the deceptive response (see Table 4.59). In fact, the maximum accuracies reached in the experiments using unexpected questions ranged from 97.5% to 95%, whereas using complex questions we reached maximum accuracies from 92.5% to 95%.

- The paradigm using complex questions together with mouse dynamics can be successfully generalize to other topic of deception, different from identity (e.g., autobiographical events, such as holidays), even if with a slightly weaker accuracy. In fact, maximum accuracy obtained by mouse dynamics and unexpected questions in spotting faked identities was of 92.5%, whereas faked holidays were detect with a maximum accuracy of 87.5%.

In Table 4.59, an overview of the best results obtained in each experiment is reported. Further considerations are the following:

- there are not classifiers better than others. In fact, all the classifiers gave very similar results within each experiment, and the classifier leading the best accuracy changed from an experiment to another. In other words, on these data all the algorithms are equivalent in terms of performance.
- This means that there are different classification logic that may be apply to distinguish liars from truth-tellers, and our results are independent from the algorithm that we used to build the classification model. In other words, results are very stable.

Deception topic			Identity		Holidays
Cognitive strategy			Unexpected	Complex	Complex
Collected measures	Mouse dynamics	Best accuracy in 10-fold cross-validation	95.00%	92.50%	87.50%
		Best classifier	Logistic	Random forest	Naïve Bayes
	RT	Best accuracy in 10-fold cross-validation	97.50%	95.00%	
		Best classifier	Random forest	SVM / LMT	
	Keystroke dynamics	Best accuracy in 10-fold cross-validation	97.50%		
		Best classifier	LMT		

Table 4.59: the table summarizes the best results obtained by the experiments in Chapter 4. Experiments are divided according to the topic of the deception (identity or holidays) and the cognitive strategy that we used to increase liars' cognitive load (expected questions or complex questions). Moreover, they are grouped based on the measures that were recorded during the task (RT, mouse dynamics and keystroke dynamics). For each experiment, the table reports the best result in terms of accuracy obtained by in the 10-fold cross-validation. The classifier that reached the best accuracy is reported as well.

5.2 General Conclusions

In Chapter 2, we stated that the proposed paradigms fit in the spectrum of behavioral lie detection techniques, such as the aIAT and the CIT. However, the aim of our work was to build lie detection models that overcomes the limitations of the traditional RT-based techniques for the web application. With regard to the average classification accuracy CIT [114] and aIAT [111] have similar accuracies to those of the experiments reported here (around 90%). However, we have shown that deception can be spotted in the absence of any ground truth information as well. In other words, differently from

CIT and aIAT, the paradigms that we have proposed do not require any true memory as alternative to build the test. Finally, while in the RT-based traditional test for lie detection the topic of the falsehood is predetermined, keystroke analysis could offer the possibility to investigate the deception in free text contents, or rather with open questions or analyzing spontaneous texts.

Our experiments confirmed the literature results according to which the cognitive complexity of the task increase the possibility to detect deception. The strategy of asking unexpected questions seems to be more efficient to increase liars' cognitive load than complex questions. However, unexpected questions are difficult to apply in some contexts (e.g., lie of omission), whereas complex questions are more adaptable to different situations, even when there are no elements to fabricate unexpected information.

Another major step that we reached and that was anticipated by the works of Duran, Dale and McNamara [116] and Banerjee et al. [133], is to demonstrate that the cognitive processes that underlie the deception production can be described ongoing. In fact, a crucial advantage of mouse dynamics and keystroke dynamics compared to RT is that they provide more information about the cognitive process of deception production. RT is a static index that describe the cognitive process only through a final behavioral outcome. Conversely, keystroke dynamics and, above all, mouse dynamics are online measures, which trace the constant evolution of the cognitive process and describe it through a complex set of dynamic behaviors [118].

Also concerning the resistance to countermeasures, the proposed paradigms seem to be encouraging. In general, mouse tracking is more resistant to countermeasures than keystroke dynamics and RT. In fact, mouse tracking collects a large number of dynamics indices, both spatial and temporal, which can be assemble in different way to produce alternative classification models. Moreover, the different nature of the indices makes it impossible for a human being to control the response features all together. On the contrary, RT and keystroke dynamics are based exclusively on temporal features, thus they are more easily beatable with countermeasures.

Other advantages of the paradigms that we have proposed in this work are related to their usability. The collection of RT, mouse dynamic and keystroke dynamics is inexpensive and does not require any equipment in addition to that the subject is already using during the his interaction with the computer. Again, these indices are very well adapted to the detection of lies in the context of web. They do not require the presence of an examiner and are suitable for large-scale applications. They can be collected automatically, quickly and anywhere.

5.3 Covert Lie Detection

In our opinion, the most innovative and important result of this work concern the use of keystroke dynamics as a tool for covert lie detection. This is the first time that a covert tool to detect liars has been proposed. In section 4.3, we demonstrated that keystroke dynamics is very promising in terms of covert lie detection and integration with the existing applications. Covert lie detection refers to the conditions under which the examinee is unaware that he is under scrutiny of a scientifically based lie detection technique. This condition permit to avoid that he puts in act countermeasure to beat the lie detector. Moreover, keystroke dynamics lie detectors may be implemented maintaining the current

user interfaces and without altering the user experience. User experience is a very important aspect in the use of a products, systems or services [190]. It includes the users' perceptions of system aspects such as utility, ease of use and efficiency. The high risk to integrate lie detection systems in online services is that users start to experience the service itself as a limitation of freedom and as a mean of close control of their activities more than a mean of free expression. For this reason, future directions should be address to design covert lie detection tools, such as deception detection via keystroke dynamics. By covert lie detection, deception is spotted without the risk that the examinee can alter the data collection, and the user experience is not affected by the purpose of the instrument.

5.4 Limitations

A first limitation of this work concern the experimental environment. Although the main goal of the research was to propose new lie detection paradigms for the online application, all the experiment were run in the lab under the supervision of an experimenter. In other words, data were collected in the laboratory and not online. Our choice to run the experiment in the laboratory was guided by previous experiences. In a preceding experiment [189] we collected online data, recruiting 244 subjects. 204 subjects completed the experiment, 31 had quit it before the end and nine had stop after registration form. After an observational data analysis, we discovered that too many participants showed a clear intention to compromise the test (e.g., to a question about the name the answer was "Goofy"). Examining the participants' responses one by one, we left out who seemed joking the test. The final sample counted 190 participants. However we cannot be sure that all the responses given by these last 190 subjects are genuine and, as consequence, that the data are closed to reality. Moreover, it was very difficult to instruct participants to lie with online instruction, risking that they did not understand the task.

Another issue that is often raised in lie detection research is the use of the instructed lies. Lie detection experimental paradigms explicitly asked participant to lie, whereas in the real situations the deception is spontaneous. In lab studies, the motivation to lie is generally low, as it does not lead to any gain or loss. However, there are good reasons to think that the cognitive mechanisms of the deception production are the same both in the experimental and real environment [71]. Rather, what is different from the lab to the daily reality is the amount of external stimuli, which may interfere with the task and which can lead to wrong conclusions. For example, a user could create a new Facebook account while he cooking and speaking with friends. For this reason, it should be appropriate to test the accuracy of the lie detectors in real situations.

Finally, we mention the need to align the lie detection paradigms to the continuous technological development. For example, a fake review or a false social network profile may be created using touch screen devices, which are without mouse and keyboard. Thus, the human-computer interaction during the act of lie should be investigate considering touch screen technologies as well.

5.5 Future Directions

As anticipated above, the main future direction consists in the integration of the lie detection paradigms within the existing online services, such as the social networks. This work has laid the groundwork for the development of covert lie detection tools and preserving the user experience. In these terms, the deception detection via keystroke dynamics seems to be the more promising technique.

Future directions should be oriented to solve the current limitations (see paragraph 5.3) and to test the generalization of the techniques to a broad range of deception topics. The main objective should be the creation of a unique model that could be apply in a large number of cheating context, ranging from lies about identity in the real or online environment to security, forensic, or commercial applications. In fact, in the introduction we have seen how the problem of the online deception is a current and very urgent issue. A large number of social network profiles are faked. Anybody may post fake information, fake news or fake reviews. Online financial services are every day at risk as well. User identification is clearly one of the most important point to deepen, in order to have a solid banking systems and genuine social interaction between real people.

Another context in which the techniques that we have proposed may fit well is the immigration office. Often, officers have to validate the immigrants' identities without any warranty about their real biographical data. Even in this situation, it is crucial to provide a powerful tool for detecting whether a person is telling the truth or a lie about himself.

Finally, our lie detection systems may be apply in the forensic context as well. In most cases, the defendant is called to confirm his version of the facts. In this situation, may be useful to have a tool that make an evaluation about his credibility. In addition, leaving aside homicides and similar situations, another possible application may be found in insurance issues, such as cases of malingering.

References

- [1] “ITU Statistics,” *International Telecommunications Union*, 2017. [Online]. Available: <http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>. [Accessed: 10-Oct-2017].
- [2] “Internet usage statistics,” *Internet World Status*, 2017. [Online]. Available: <http://www.internetworldstats.com/stats.htm>. [Accessed: 03-Oct-2017].
- [3] “Company info,” *Facebook newsroom*, 2017. [Online]. Available: <https://newsroom.fb.com/company-info/>. [Accessed: 03-Oct-2017].
- [4] “The Top 20 valuable Facebook statistics – Updated September 2017,” *Zephoria*, 2017. [Online]. Available: <https://zephoria.com/top-15-valuable-facebook-statistics/>. [Accessed: 03-Oct-2017].
- [5] J. Constine, “Facebook now has 2 billion monthly users... and responsibility,” *Techcrunch*, 2017. [Online]. Available: <https://techcrunch.com/2017/06/27/facebook-2-billion-users/>. [Accessed: 10-Oct-2017].
- [6] “Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017 (in millions),” *Statista*, 2017. [Online]. Available: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>. [Accessed: 03-Oct-2017].
- [7] “Quarterly report pursuant to section 13 or 15(d) of the Securities Exchange Act of 1934 for Facebook, INC.,” 2012.
- [8] K. Heather, “83 million Facebook accounts are fakes and dupes,” *CNN*, 2012. [Online]. Available: <http://edition.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/index.html>. [Accessed: 02-Oct-2017].
- [9] D. Wakabayashi and M. Isaac, “In race against fake news, Google and Facebook stroll to the starting line,” *The New York Times*, 2017. [Online]. Available: <https://www.nytimes.com/2017/01/25/technology/google-facebook-fake-news.html>. [Accessed: 04-Oct-2017].
- [10] E. Woollacott, “Amazon’s fake review problem is now worse than ever, study suggests,” *Forbes*, 2017. [Online]. Available: <https://www.forbes.com/sites/emmawoollacott/2017/09/09/exclusive-amazons-fake-review-problem-is-now-worse-than-ever/#42d129ca7c0f>. [Accessed: 04-Oct-2017].
- [11] M. Brignall, “So you think you’re safe doing internet banking?,” *The Guardian*, 2015. [Online]. Available: <https://www.theguardian.com/money/2015/nov/21/safe-internet-banking-cyber-security-online>. [Accessed: 04-Oct-2017].
- [12] S. Barber, “The direct link between identity theft and terrorism, and ways to stop it,” *The University of Texas at Austin*, 2015. [Online]. Available: <https://news.utexas.edu/2015/12/07/the-direct-link-between-identity-theft-and-terrorism>. [Accessed: 11-Oct-2017].
- [13] “Bruxelles: kamikaze usò identità ex giocatore dell’Inter,” *Agenzia Giornalistica Italia (AGI)*, 2016. [Online]. Available: https://www.agi.it/estero/bruxelles_kamikaze_uso_identita_ex_giocatore_dellinter-

- 650281/news/2016-03-28/. [Accessed: 11-Oct-2017].
- [14] “Belgium sentences fake-ID gang used by Brussels and Paris attackers,” *BBC news*, 2017. [Online]. Available: <http://www.bbc.com/news/world-europe-38683199>. [Accessed: 11-Oct-2017].
- [15] R. Gross and A. Acquisti, “Information revelation and privacy in online social networks,” in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, 2005, pp. 71–80.
- [16] A. Siibak, “Casanova’s of the virtual world. How boys present themselves on dating websites,” in *Proceeding of the 5th International Conference on Youth Research*, 2007, pp. 83–91.
- [17] A. E. Cano, M. Fernandez, and H. Alani, “Detecting child grooming behaviour patterns on social media,” in *Social Informatics. SocInfo 2014. Lecture Notes in Computer Science*, L. M. Aiello and D. McFarland, Eds. Springer, Cham, 2014, pp. 412–427.
- [18] J. Suler, “The online disinhibition effect,” *CyberPsychology Behav.*, vol. 7, no. 3, pp. 321–326, 2004.
- [19] “Fake name generator.” [Online]. Available: <http://it.fakenamegenerator.com/gen-female-us-us.php>. [Accessed: 11-Oct-2017].
- [20] C. J. Hoofnagle, “Identity theft: making the known unknowns known,” *Harvard J. Law Technol.*, vol. 21, 2007.
- [21] “An act to add Section 528.5 to the Penal Code, relating to impersonation,” *California Penal Code*, 2010. [Online]. Available: http://leginfo.ca.gov/pub/09-10/bill/sen/sb_1401-1450/sb_1411_bill_20100927_chaptered.html. [Accessed: 11-Oct-2017].
- [22] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, “Aiding the detection of fake accounts in large scale social online services,” in *9th Symposium on Networked Systems Design and Implementation*, 2012, pp. 197–210.
- [23] H. N. Pontell, “Identity theft: Bounded rationality, research, and policy,” *Criminol. Public Policy*, vol. 8, no. 2, pp. 263–270, 2009.
- [24] “How many of the Internet’s users are fake,” *Ghost Proxies*, 2015. .
- [25] S. Wachs, K. D. Wolf, and C.-C. Pan, “Cybergrooming: risk factors, coping strategies and associations with cyberbullying,” *Psicothema*, vol. 24, no. 4, pp. 628–633, 2012.
- [26] H. C. Whittle, C. Hamilton-Giachritsis, and A. R. Beech, “Victims’ voices: the impact of online grooming and sexual abuse,” *Univers. J. Psychol.*, vol. 1, no. 2, pp. 59–71, 2013.
- [27] J. Nye, “Tragic online love triangle built on LIES: two middle-aged lovers who started affair by BOTH posing as teenagers... before torrid romance drove Sunday school teacher to murder ‘rival’ over woman who didn’t EXIST,” *Mail Online News*, 2012. [Online]. Available: <http://www.dailymail.co.uk/news/article-2163757/Tall-hot-blonde-internet-love-triangle-left-man-shot-dead-jealous-rival.html>. [Accessed: 23-Oct-2017].
- [28] A. A. Gillespie, “Adolescents, sexting and human rights,” *Hum. Rights Law Rev.*, vol. 13, no. 4, pp. 623–643, 2013.
- [29] A. Eleuteri, “Il fenomeno del sexting tra i giovani,” *Istituto di Sessuologia Clinica*. [Online].

- Available: <http://www.sessuologiaclinicaroma.it/il-fenomeno-del-sexting-tra-i-giovani/>. [Accessed: 24-Oct-2017].
- [30] J. R. Temple, J. A. Paul, P. van den Berg, V. D. Le, A. McElhany, and B. W. Temple, “Teen sexting and its association with sexual behaviors,” *Arch. Pediatr. Adolesc. Med.*, vol. 166, no. 9, pp. 828–833, 2012.
- [31] M. Dal Negro, “Sexting in Italia: i dati aggiornati Telefono azzurro - Eurispes (16/04/2014),” *Mybestlife*, 2015. [Online]. Available: <http://www.mybestlife.com/sexuality/news-2014-apr/16042014-sexting-telefono-azzurro-Eurispes.html>. [Accessed: 24-Oct-2017].
- [32] “Snapchat,” *Snapchat*. [Online]. Available: <https://www.snapchat.com/l/it-it/>. [Accessed: 24-Oct-2017].
- [33] “#snapchat,” *Instagram*. [Online]. Available: <https://www.instagram.com/explore/tags/snapchat/?hl=it>. [Accessed: 24-Oct-2017].
- [34] S. Kleeman, “Instagram finally revealed the reason it banned nipples — it’s Apple,” *Mic*, 2015. [Online]. Available: <https://mic.com/articles/126137/instagram-banned-nipples-because-of-apple#.gZd2ReelZ>. [Accessed: 24-Oct-2017].
- [35] “ESTA,” *U.S. Customs and Border Protection*. [Online]. Available: <https://esta.cbp.dhs.gov/esta/>. [Accessed: 24-Oct-2017].
- [36] J. G. Hickey, “Report: 25 percent jump in illegals stopped at border for fake IDs,” *Newsmax*, 2015. [Online]. Available: <http://www.newsmax.com/Newsfront/illegals-border-fake-id/2015/01/08/id/617256/>. [Accessed: 24-Oct-2017].
- [37] R. Scarcella, “In Europa la fabbrica della cittadinanza: così funziona il business dei passaporti comprati,” *La Stampa*, 2016. [Online]. Available: <http://www.lastampa.it/2016/08/21/italia/cronache/in-europa-la-fabbrica-della-cittadinanza-cos-funziona-il-business-dei-passaporti-comprati-e0dWQ28J4sGXYEcJ0jWLOI/pagina.html>. [Accessed: 24-Oct-2017].
- [38] “Belgium sentences fake-ID gang used by Brussels and Paris attackers,” *BBC news*, 2017. [Online]. Available: <http://www.bbc.com/news/world-europe-38683199>. [Accessed: 24-Oct-2017].
- [39] P. Oltermann, “Fingerprints in Berlin truck match those of suspect Anis Amri,” *The Guardian*, 2016. .
- [40] “Perché i terroristi si lasciano dietro i documenti?,” *Il Post*, 2016. [Online]. Available: <http://www.ilpost.it/2016/12/22/perche-i-terroristi-si-lasciano-dietro-i-documenti/>. [Accessed: 24-Oct-2017].
- [41] “Crossing borders: how terrorists use fake passports, visas, and other identity documents,” *Frontline*, 2014. [Online]. Available: <http://www.pbs.org/wgbh/pages/frontline/shows/trail/etc/fake.html>. [Accessed: 24-Oct-2017].
- [42] “H.R.158 - Visa Waiver Program improvement and Terrorist Travel Prevention Act of 2015,” *Congress.Gov*, 2015. [Online]. Available: <https://www.congress.gov/bill/114th-congress/house-bill/158/text>. [Accessed: 24-Oct-2017].
- [43] S. D’Alfonso, “Synthetic identity theft: three ways synthetic identities are created,” *Security*

- Intelligence*, 2014. [Online]. Available: <https://securityintelligence.com/synthetic-identity-theft-three-ways-synthetic-identities-are-created/>. [Accessed: 24-Oct-2017].
- [44] “Identifying synthetic identity fraud,” *Trulioo*, 2017. [Online]. Available: <https://www.trulioo.com/blog/synthetic-identity-fraud/>. [Accessed: 24-Oct-2017].
- [45] “The new reality of synthetic ID fraud,” *Equifax*, 2015. [Online]. Available: https://www.equifax.com/assets/IFS/syntheticID-fraud_wp.pdf. [Accessed: 24-Oct-2017].
- [46] A. Charles and K. Stuart, “PlayStation Network users fear identity theft after major data leak,” 2011. [Online]. Available: <https://www.theguardian.com/technology/2011/apr/27/playstation-users-identity-theft-data-leak>. [Accessed: 27-Oct-2017].
- [47] M. Rose, “Canadian law firm files \$1B lawsuit against Sony over PSN data breach,” *Gamasutra*, 2011. [Online]. Available: https://www.gamasutra.com/view/news/34499/Canadian_Law_Firm_Files_1_Billion_Class_Action_Lawsuit_Against_Sony_Over_PSN_Data_Breach.php.
- [48] “Will the real Eriberito stand up,” *UEFA.com*, 2002. [Online]. Available: <http://www.uefa.com/news/newsid=34451.html>. [Accessed: 24-Oct-2017].
- [49] A. Levin, “The invisible victims of identity theft: our kids,” *Huffpost*, 2016. [Online]. Available: https://www.huffingtonpost.com/adam-levin/the-invisible-victims-of_b_8539352.html. [Accessed: 24-Oct-2017].
- [50] E. Hunt, “What is fake news? How to spot it and what you can do to stop it,” *The Guardian*, 2016. [Online]. Available: <https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate>. [Accessed: 24-Oct-2017].
- [51] “2016 Edelman trust barometer,” 2016.
- [52] H. Ritchie, “Read all about it: the biggest fake news stories of 2016,” *CNBC.com*, 2016. [Online]. Available: <https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html>. [Accessed: 03-Oct-2017].
- [53] “Facebook Under Fire for Fake News Stories,” *abc NEWS*, 2016. [Online]. Available: <https://www.youtube.com/watch?v=ox7YBHJUm6Q>. [Accessed: 04-Oct-2017].
- [54] C. Silverman, “This analysis shows how viral fake election news stories outperformed real news On Facebook,” *BuzzFeed*, 2016. [Online]. Available: https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.tg7MPwAVR#.vobYJGezq. [Accessed: 03-Oct-2017].
- [55] J. Malbon, “Taking fake online consumer reviews seriously,” *J. Consum. Policy*, vol. 36, no. 2, pp. 139–157, 2013.
- [56] “Online consumer reviews. The case of misleading or fake reviews,” 2015.
- [57] D. Streitfeld, “Give yourself 5 stars? Online, it might cost you,” *The New York Times*, 2013. [Online]. Available: <http://www.nytimes.com/2013/09/23/technology/give-yourself-4-stars-online-it-might-cost-you.html>. [Accessed: 24-Oct-2017].
- [58] M. Luca and G. Zervas, “Fake it till you make it: reputation, competition, and Yelp review fraud,” *Manage. Sci.*, vol. 62, no. 12, pp. 3412–3427, 2016.

- [59] E. Ford, "Lie detection: historical, neuropsychiatric and legal dimensions," *Int. J. Law Psychiatry*, vol. 29, no. 3, pp. 159–177, 2006.
- [60] N. Abe, "How the brain shapes deception: An integrated review of the literature," *Neurosci.*, vol. 17, no. 5, pp. 560–574, 2011.
- [61] V. Benussi, "Die atmungssymptome der lüge," *Arch. Gesamte Psychol.*, vol. 31, pp. 244–273, 1914.
- [62] P. A. Granhag, A. Vrij, and B. Verschuere, *Deception detection: current challenges and new approaches*. John Wiley & Sons, Ltd., 2015.
- [63] J. P. Rosenfeld, "Alternative views of Bashore and Rapp's (1993) alternatives to traditional polygraphy: a critique," *Psychol. Bull.*, vol. 117, no. 1, pp. 159–166, 1995.
- [64] J. E. Reid, "A revised questioning technique in lie detection tests," *J. Crim. Law, Criminol. Am. Police Sci.*, vol. 37, pp. 542–547, 1947.
- [65] V. V. MacLaren, "A quantitative review of the guilty knowledge test," *J. Appl. Psychol.*, vol. 86, no. 4, pp. 674–683, 2001.
- [66] S. Kugelmass and I. Lieblich, "Effects of realistic stress and procedural interference in experimental lie detection," *J. Appl. Psychol.*, vol. 50, no. 3, pp. 211–216, 1966.
- [67] J. J. Walczyk, D. A. Griffith, R. Yates, S. R. Visconte, B. Simoneaux, and L. L. Harris, "Lie detection by inducing cognitive load. Eye movements and other cues to the false answers of 'witnesses' to crimes," *Crim. Justice Behav.*, vol. 39, no. 7, pp. 887–909, 2012.
- [68] L. Warmelink, A. Vrij, S. Mann, S. Leal, D. Forrester, and R. Fisher, "Thermal imaging as a lie detection tool at airports," *Law Hum. Behav.*, vol. 35, no. 1, pp. 40–48, 2011.
- [69] F. Horvath, "Detecting deception: the promise and the reality of voice stress analysis," *J. Forensic Sci.*, vol. 27, no. 2, pp. 340–351, 1982.
- [70] P. Ekman and W. V. Friesen, "Detecting deception from body or face," *Journal Personal. Soc. Psychol.*, vol. 29, pp. 288–298, 1974.
- [71] G. Ganis and J. P. Keenan, "The cognitive neuroscience of deception," *Soc. Neurosci.*, vol. 4, no. 6, pp. 465–472, 2009.
- [72] J. Meixner and J. P. Rosenfeld, "A mock terrorism application of the P300-based Concealed Information Test," *Psychophysiology*, vol. 48, pp. 149–154, 2011.
- [73] C. Davatzikos *et al.*, "Classifying spatial patterns of brain activity with machine learning methods: application to lie detection," *Neuroimage*, vol. 28, no. 3, pp. 663–668, 2005.
- [74] J. K. Burgoon, J. P. Blair, J. F. Tiantian Qin, and J. Nunamaker, "Detecting deception through linguistic analysis," in *Intelligence and Security Informatics. ISI 2003. Lecture Notes in Computer Science*, 2003, vol. 2665, pp. 91–101.
- [75] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying Words: predicting deception from linguistic styles," *Personal. Soc. Psychol. Bull.*, vol. 29, no. 5, 2003.
- [76] G. Bogaard, E. H. Meijer, A. Vrij, N. J. Broers, and H. Merckelbach, "Contextual bias in verbal

- credibility assessment: criteria-based content analysis, reality monitoring and scientific content analysis,” *Appl. Cogn. Psychol.*, vol. 28, pp. 79–90, 2014.
- [77] N. Smith, “Reading between the lines: an evaluation of the scientific content analysis technique (SCAN),” *Police Res. Ser. Pap.*, vol. 135, pp. 1–42, 2001.
- [78] A. Vrij, *Detecting lies and deceit: Pitfalls and opportunities.*, 2nd Editio. Chichester: Wiley, 2008.
- [79] A. Sapir, “The LSI course on scientific content analysis (SCAN),” *LSI Laboratory for Scientific Interrogation, Inc.*, 2005. [Online]. Available: <http://www.lsiscan.com/>. [Accessed: 26-Oct-2017].
- [80] G. Nahari, A. Vrij, and R. P. Fisher, “Does the truth come out in the writing? SCAN as a lie detection tool,” *Law Hum. Behav.*, vol. 36, pp. 68–76, 2012.
- [81] S. H. Adams, “Statement analysis: what do suspects’ words really reveal?,” *Polygraph*, vol. 25, no. 4, pp. 266–278, 1996.
- [82] B. G. Amado, R. Arce, and F. Fariña, “Undeutsch hypothesis and Criteria Based Content Analysis: a meta-analytic review,” *Eur. J. Psychol. Appl. to Leg. Context*, vol. 7, no. 1, pp. 3–12, 2015.
- [83] M. Steller and G. Köhnken, “Criteria Based Statement Analysis,” in *Psychological methods in criminal investigation and evidence*, D. C. Raskin, Ed. New York: Springer, 1989, pp. 217–245.
- [84] A. Vrij, “Criteria Based Content Analysis: a qualitative review of the first 37 studies,” *Psychol. Public Policy, Law*, vol. 11, pp. 3–41, 2005.
- [85] A. Pittarello, D. Motro, E. Rubaltelli, and P. Pluchino, “The relationship between attention allocation and cheating,” *Psychon. Bull. Rev.*, no. February 2016, pp. 609–616, 2015.
- [86] M. K. Johnson and C. L. Raye, “Reality monitoring,” *Psychol. Rev.*, vol. 88, pp. 67–85, 1981.
- [87] S. L. Sporer, “The less travelled road to truth: verbal cues in deception detection in accounts of fabricated and self-experienced events,” *Appl. Cogn. Psychol.*, vol. 11, pp. 373–397, 1997.
- [88] “They Say,” 2015. [Online]. Available: <http://www.theysay.io/>. [Accessed: 26-Oct-2017].
- [89] R. Mihalcea and C. Strapparava, “The lie detector: explorations in the automatic recognition of deceptive language,” in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (ACLShort ’09)*, 2009, pp. 309–312.
- [90] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011, vol. 1, pp. 309–319.
- [91] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, “What Yelp fake review filter might be doing?,” in *ICWSM, the AAAI Press*, 2013.
- [92] A. Vrij, K. Edward, K. P. Roberts, and R. Bull, “Detecting deceit via analysis of verbal and nonverbal behavior,” *J. Nonverbal Behav.*, vol. 24, no. 4, pp. 239–263, 2000.

- [93] L. Zhou, "An empirical investigation of deception behavior in instant messaging," in *IEEE Transactions on Professional Communication*, 2005, vol. 48, no. 2, pp. 147–160.
- [94] D. C. Derrick, T. O. Meservy, J. L. Jenkins, J. K. Burgoon, and J. F. Nunamaker, "Detecting deceptive chat-based communication using typing behavior and message cues," *ACM Trans. Manag. Inf. Syst.*, vol. 4, no. 2, 2013.
- [95] "Why Yelp has a review filter," *Yelp official blog*, 2009. [Online]. Available: <https://www.yelpblog.com/2009/10/why-yelp-has-a-review-filter>.
- [96] K. David, "Understading the Yelp filter: an exploratory study," *First Monday*, vol. 19, no. 9, 2014.
- [97] L. Held, "Behind the curtain of Yelp's powerful reviews," *Entrepreneur*, 2014. [Online]. Available: <https://www.entrepreneur.com/article/235271>. [Accessed: 25-Oct-2017].
- [98] K. Suchotzki, B. Verschuere, B. Van Bockstaele, G. Ben-Shakhar, and G. Crombez, "Lying takes time: A meta-analysis on reaction time measures of deception.," *Psychol. Bull.*, vol. 143, no. 4, pp. 428–453, 2017.
- [99] A. Vrij, R. Fisher, S. Mann, and S. Leal, "A cognitive load approach to lie detection," *Investig. Psychol. Offender Profiling*, vol. 5, pp. 39–43, 2008.
- [100] A. Vrij, R. P. Fisher, and H. Blank, "A cognitive approach to lie detection: a meta-analysis," *Leg. Criminol. Psychol.*, vol. 22, no. 1, pp. 1–21, 2017.
- [101] J. J. Walczyk, F. P. Igou, A. P. Dixon, and T. Tcholakian, "Advancing lie detection by inducing cognitive load on liars: a review of relevant theories and techniques guided by lessons from polygraph-Based approaches," *Front. Psychol.*, vol. 4, p. 14, 2013.
- [102] V. A. Gombos, "The cognition of deception: the role of executive processes in producing lies," *Genet. Soc. Gen. Psychol. Monogr.*, vol. 132, no. 3, pp. 197–214, 2006.
- [103] I. Blandon-Gitlin, E. Fenn, J. Masip, and A. H. Yoo, "Cognitive-load approaches to detect deception: searching for cognitive mechanisms," *Trends Cogn. Sci.*, vol. 18, no. 9, pp. 441–444, 2014.
- [104] M. R. Sheridan and K. A. Flowers, "Reaction Times and deception - the lying constant," *Int. J. Psychol. Stud.*, vol. 2, no. 2, pp. 41–51, 2010.
- [105] E. Debey, J. De Houwer, and B. Verschuere, "Lying relies on the truth," *Cognition*, vol. 132, pp. 324–334, 2014.
- [106] D. T. Lykken, "The GSR in the detection of guilt," *J. Appl. Psychol.*, vol. 43, no. 6, pp. 385–388, 1959.
- [107] A. P. Gregg, "When vying reveals lying: the timed antagonistic response alethiometer," *Appl. Cogn. Psychol.*, vol. 21, no. 5, pp. 621–647, 2007.
- [108] S. A. Spence, T. F. Farrow, A. E. Herford, I. D. Wikinson, Y. Zheng, and P. W. R. (2001) Woodruff, "Behavioral and functional anatomical correlates deception in humans," *Neuroreport*, vol. 12, pp. 2849–2853, 2001.
- [109] G. Sartori, S. Agosta, C. Zogmaister, S. D. Ferrara, and U. Castiello, "How to accurately detect

- autobiographical events,” *Psychol. Sci.*, vol. 19, no. 8, pp. 772–780, 2008.
- [110] G. A. Greenwald, E. D. McGhee, and K. L. J. Schwartz, “Measuring individual differences in implicit cognition: the implicit association test,” *J. Pers. Soc. Psychol.*, vol. 74, no. 6, pp. 1464–1480, 1998.
- [111] S. Agosta and G. Sartori, “The autobiographical IAT: a review,” *Front. Psychol.*, vol. 4, no. 519, 2013.
- [112] B. Van Bockstaele, B. Verschuere, T. Moens, K. Suchotzki, E. Debey, and A. Spruyt, “Learning to lie: effects of practice on the cognitive cost of lying,” *Front. Psychol.*, vol. 3, p. 526, 2012.
- [113] B. Verschuere, G. Ben-Shakhar, and E. Meijer, *Memory detection: theory and application of the concealed information test*. Cambridge University Press, 2011.
- [114] B. Kleinberg and B. Verschuere, “Memory detection 2.0: the first web-based memory detection test,” *PLoS One*, vol. 10, no. 4:e0118715, 2015.
- [115] B. Verschuere and B. Kleinberg, “ID-Check: Online Concealed Information Test Reveals True Identity,” *J. Forensic Sci.*, vol. 61 Suppl 1, pp. S237-40, 2016.
- [116] N. D. Duran, R. Dale, and D. S. McNamara, “The action dynamics of overcoming the truth,” *Psychon. Bull. Rev.*, vol. 17, no. 4, pp. 486–491, 2010.
- [117] G. . Grimes, J. L. Jenkins, and J. S. Valacich, “Assessing credibility by monitoring changes in typing behavior: the keystrokes dynamics deception detection model,” in *Hawaii International Conference on System Sciences. Deception Detection Symposium*, 2013.
- [118] J. B. Freeman, R. Dale, and T. A. Farmer, “Hand in motion reveals mind in motion,” *Front. Psychol.*, vol. 2, no. 59, 2011.
- [119] J. H. Song and K. Nakayama, “Hidden cognitive states revealed in choice reaching tasks,” *Trends Cogn. Sci.*, vol. 13, no. 8, pp. 360–366, 2009.
- [120] A. Calcagni and L. Lombardi, “Dynamic Fuzzy Rating Tracker (DYFRAT): a novel methodology for modeling real-time dynamic cognitive processes in rating scales,” *Appl. Soft Comput.*, vol. 24, pp. 948–961, 2014.
- [121] J. B. Freeman and N. Ambady, “MouseTracker: software for studying real-time mouse-tracking method,” *Behav. Res. Methods*, vol. 42, no. 1, pp. 226–241, 2010.
- [122] R. Dale and N. D. Duran, “The cognitive dynamics of negated sentence verification,” *Cogn. Sci.*, vol. 35, no. 5, pp. 983–996, 2011.
- [123] J. B. Freeman, K. Pauker, and D. T. Sanchez, “A perceptual pathway to bias: interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias,” *Psychol. Sci.*, vol. 27, pp. 502–517, 2016.
- [124] B. Quétard *et al.*, “Combined effects of expectations and visual uncertainty upon detection and identification of a target in the fog,” *Cogn. Process.*, vol. 16 Suppl 1, pp. 343–348, 2015.
- [125] D. H. Abney, D. M. McBride, A. M. Conte, and D. W. Vinson, “Response dynamics in prospective memory,” *Psychon. Bull. Rev.*, vol. 22, no. 4, pp. 1020–1028, 2015.

- [126] L. Barca and G. Pezzulo, “Unfolding visual lexical decision in time,” *PLoS One*, vol. 7, no. 4, p. e35932, 2012.
- [127] M. Hibbeln, J. Jenkins, C. Schneider, J. Valacich, and M. Weinmann, “Investigating the effect of insurance fraud on mouse usage in human-computer interactions,” in *International Conference on Information Systems (ICIS 2014)*, 2014.
- [128] J. S. Valacich, J. L. Jenkins, S. Hariri, and J. Howie, “Identifying insider threats through monitoring mouse movements in concealed information tests,” in *Hawaii International Conference on System Sciences iHawaii International Conference on Computer and Systems Sciences, Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment*, 2013, p. 114.
- [129] R. Moskovitch *et al.*, “Identity theft, computers and behavioral biometrics,” in *IEEE International Conference on Intelligence and Security Informatics, 2009 (InISI '09)*, 2009, vol. 16.
- [130] N. Ahmad, A. Szymkowiak, and P. a Campbell, “Keystroke dynamics in the pre-touchscreen era,” *Front. Hum. Neurosci.*, vol. 7, p. 835, 2013.
- [131] P. S. Teh, A. B. Teoh, and S. Yue, “A survey of keystroke dynamics biometrics,” *Sci. World J.*, 2013.
- [132] F. Monroe and A. D. Rubin, “Keystroke dynamics as a biometric for authentication,” *Futur. Gener. Comput. Syst.*, vol. 16, pp. 351–359, 2000.
- [133] R. Banerjee, S. Feng, J. S. Kang, and Y. Choi, “Keystroke Patterns as Prosody in Digital Writings: A Case Study with Deceptive Reviews and Essays,” *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP 2014)*, pp. 1469–1473, 2014.
- [134] M. Monaro, L. Gamberini, and G. Sartori, “The detection of faked identity using unexpected questions and mouse dynamics,” *PLoS One*, vol. 12, no. 5: e0177851, pp. 1–19, 2017.
- [135] X. Hu, H. Chen, and G. Fu, “A repeated lie becomes a truth? The effect of intentional control and training on deception,” *Front. Psychol.*, vol. 3, p. 488, 2012.
- [136] B. Van Bockstaele, C. Wilhelm, E. Meijer, E. Debey, and B. Verschuere, “When deception becomes easy: The effects of task switching and goal neglect on the truth proportion effect,” *Front. Psychol.*, vol. 6, p. 1666, 2015.
- [137] A. Vrij, R. Fisher, S. Mann, and L. S., “Detecting deception by manipulating cognitive load,” *Trends Cogn. Sci.*, vol. 10, no. 4, pp. 141–142, 2006.
- [138] E. Debey, B. Liefoghe, J. De Houwer, and B. Verschuere, “Lie, truth, lie: the role of task switching in a deception context,” *Psychol. Res.*, vol. 79, pp. 478–488, 2015.
- [139] J. J. Walczyk, J. P. Schwartz, R. Clifton, B. Adams, M. Wei, and P. Zha, “Lying person-to-person about life events: A cognitive framework for lie detection,” *Pers. Psychol.*, vol. 58, pp. 141–170, 2005.
- [140] A. Vrij, S. Mann, S. Leal, and R. Fisher, “Is anyone there? Drawings as a tool to detect deceit in occupation interviews,” *Psychol. Crime Law*, vol. 18, pp. 377–388, 2012.
- [141] A. Vrij, S. A. Mann, R. P. Fisher, S. Leal, R. Milne, and R. Bull, “Increasing cognitive load to

- facilitate lie detection: the benefit of recalling an event in reverse order,” *Law Hum. Behav.*, vol. 32, no. 3, pp. 253–265, 2008.
- [142] M. G. Frank and E. Svetieva, “Lies worth catching involve both emotion and cognition,” *J. Appl. Res. Mem. Cogn.*, vol. 1, no. 2, pp. 131–133, 2012.
- [143] A. Vrij *et al.*, “Outsmarting the liars: The benefit of asking unanticipated questions,” *Law Hum. Behav.*, vol. 33, no. 2, pp. 159–166, 2009.
- [144] L. Warmelink, A. Vrij, S. Mann, S. Leal, and F. H. Poletiek, “The effects of unexpected questions on detecting familiar and unfamiliar lies,” *Psychiatry, Psychol. Law*, vol. 20, no. 1, pp. 29–35, 2013.
- [145] M. Hartwig, P. A. Granhag, and L. Strömwall, “Guilty and innocent suspects’ strategies during interrogations,” *Psychol. Crime Law & Law*, vol. 13, pp. 213–227, 2007.
- [146] G. L. J. Lancaster, A. Vrij, L. Hope, and B. Waller, “Sorting the Liars from the Truth Tellers: The Benefits of Asking Unanticipated Questions on Lie Detection,” *Appl. Cogn. Psychol.*, vol. 27, pp. 107–114, 2013.
- [147] L. M. Swol and M. T. Braun, “Communicating deception: differences in language use, justifications, and questions for lies, omissions, and truths,” *Gr. Decis. Negot.*, vol. 23, no. 6, pp. 1343–1367, 2014.
- [148] E. J. Williams, L. A. Bott, J. Patrick, and M. B. Lewis, “Telling Lies: The Irrepressible Truth?,” *PLoS One*, vol. 8, no. 4, p. e60713, 2013.
- [149] “Ethical principles of psychologists and code of conduct,” *American Psychological Association*, 2017. [Online]. Available: <http://www.apa.org/ethics/code/index.aspx>. [Accessed: 04-Oct-2017].
- [150] K. Magnusson, “Understanding statistical power and significance testing.” [Online]. Available: <http://rpsychologist.com/d3/NHST/>. [Accessed: 08-Oct-2017].
- [151] W. Schneider, A. Eschman, and A. Zuccolotto, *E-Prime getting started guide*. Psychology Software Tools, Inc., 2007.
- [152] R. Bruyer and M. Brysbaert, “Combining speed and accuracy in cognitive psychology: is the Inverse Efficiency Score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)?,” *Psychol. Belg.*, vol. 51, no. 1, pp. 5–13, 2011.
- [153] P. Lucisano and M. E. Piemontese, “GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana,” *Sc. e città*, vol. 3, pp. 110–124, 1988.
- [154] M. A. Hall, “Correlation-based Feature Selection for Machine Learning,” The University of Waikato, Hamilton, 1999.
- [155] M. L. Bermingham *et al.*, “Application of high-dimensional feature selection: evaluation for genomic prediction in man,” *Sci. Rep.*, vol. 5, pp. 1–12, 2015.
- [156] M. Hall *et al.*, “The WEKA data mining software: an update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.
- [157] “The R Project for statistical computing,” 2015. [Online]. Available: <https://www.r->

- project.org/. [Accessed: 10-Oct-2017].
- [158] D. Navarro, *Learning statistics with R: a tutorial for psychology students and other beginners*. University of Adelaide, Adelaide, 2015.
- [159] “Package ‘ez,’” 2016. [Online]. Available: <https://cran.r-project.org/web/packages/ez/ez.pdf>. [Accessed: 10-Oct-2017].
- [160] “effsize: efficient Effect Size computation,” 2017. [Online]. Available: <https://cran.r-project.org/web/packages/effsize/index.html>. [Accessed: 13-Oct-2017].
- [161] J. Cohen, “A power primer,” *Psychol. Bull.*, vol. 112, no. 1, pp. 155–159, 1992.
- [162] “BayesFactor: computation of Bayes Factors for common designs,” 2015. [Online]. Available: <https://cran.r-project.org/web/packages/BayesFactor/index.html>. [Accessed: 13-Oct-2017].
- [163] R. E. Kass and A. E. Raftery, “Bayes Factors,” *J. Am. Stat. Assoc.*, vol. 90, no. 430, pp. 773–795, 1995.
- [164] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, vol. 2, no. 12, pp. 1137–1143.
- [165] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, “The reusable holdout: preserving validity in adaptive data analysis,” *Science (80-.)*, vol. 349, no. 6248, pp. 3–6, 2015.
- [166] O. Nelles, *Nonlinear system identification. From classical approaches to neural networks and fuzzy models*. Springer-Verlag Berlin Heidelberg, 2001.
- [167] S. le Cessie and J. C. van Houwelingen, “Ridge estimators in logistic regression,” *Appl. Stat.*, vol. 41, no. 1, pp. 191–201, 1992.
- [168] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, “Improvements to platt’s SMO algorithm for SVM classifier design,” *Neural Comput.*, vol. 13, no. 3, pp. 637–649, 2001.
- [169] J. C. Platt, “Fast training of support vector machines using sequential minimal optimization,” in *Advances in Kernel Methods*, C. J. C. Burges, B. Schölkopf, and A. J. Smola, Eds. Cambridge: MIT Press, 1999.
- [170] G. H. John and P. Langley, “Estimating continuous distributions in Bayesian classifiers,” in *Proceeding of the 11th Conference on Uncertainty in Artificial Intelligence*, 1995, pp. 338–345.
- [171] L. Breiman, “Random forest,” *Mach. Learn.*, vol. 45, no. 1, p. 5–32., 2001.
- [172] N. Landwehr, M. Hall, and E. Frank, “Logistic model trees,” *Mach. Learn.*, vol. 95, no. 1–2, pp. 161–205, 2005.
- [173] J. S. Quinlan, *C4.5: programs for machine learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
- [174] T. Mitchell, “Decision tree learning,” in *Machine Learning*, T. Mitchell, Ed. McGraw Hill, 1997.

- [175] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," in *Wadsworth International Group*, 1984.
- [176] M. Monaro, L. Gamberini, and G. Sartori, "Identity verification using a kinematic memory detection technique," in *Advances in Neuroergonomics and Cognitive Engineering. Advances in Intelligent Systems and Computing, vol 488*, K. Hale and K. Stanney, Eds. Springer, Cham, 2017, pp. 123–132.
- [177] M. Monaro, F. I. Fugazza, L. Gamberini, and G. Sartori, "How human-nouse interaction can accurately detect faked responses about identity," in *Symbiotic Interaction. Symbiotic 2016. Lecture Notes in Computer Science, vol 9961*, L. Gamberini, A. Spagnolli, G. Jacucci, B. Blankertz, and J. Freeman, Eds. Springer, Cham, 2017, pp. 115–124.
- [178] J. Peth, K. Suchotzki, and G. Matthias, "Influence of countermeasures on the validity of the Concealed Information Test," *Psychophysiology*, vol. 53, no. 9, pp. 1429–1440, 2016.
- [179] G. Ganis, J. P. Rosenfeld, J. Meixner, R. A. Kievit, and H. E. Schendan, "Lying in the scanner: covert countermeasures disrupt deception detection by functional magnetic resonance imaging.," *Neuroimage*, vol. 55, no. 1, pp. 312–319, 2011.
- [180] S. Agosta, V. Ghirardi, C. Zogmaister, U. Castiello, and G. Sartori, "Detecting fakers of the autobiographical IAT," *Appl. Cogn. Psychol.*, vol. 25, no. 2, pp. 299–306, 2010.
- [181] S. Agosta, A. Mega, and G. Sartori, "Detrimental effects of using negative sentences in the autobiographical IAT," *Acta Psychol. (Amst.)*, vol. 136, no. 3, pp. 269–275, 2011.
- [182] B. Verschuere, V. Prati, and J. D. Houwer, "Cheating the lie detector: faking in the autobiographical Implicit Association Test," *Psychol. Sci.*, vol. 20, no. 4, pp. 410–413, 2009.
- [183] M. Tettamanti, R. Manenti, A. Della Rosa, Pasquale A. Falini, D. Perani, S. F. Cappa, and A. Moro, "Negation in the brain: modulating action representations," *Neuroimage*, vol. 43, no. 2, pp. 358–367, 2008.
- [184] K. R. Christensen, "Negative and affirmative sentences increase activation in different areas in the brain," *J. Neurolinguistics*, vol. 22, no. 1, pp. 1–17, 2009.
- [185] F. T. Liu, M. K. Ting, and Z. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. from Data*, vol. 6, no. 1, 2012.
- [186] K. Hempstalk, E. Frank, and I. . Witten, "One-Class classification by combining density and class probability estimation," in *Proceedings of the 12th European Conference on Principles and Practice of Knowledge Discovery in Databases and 19th European Conference on Machine Learning, ECMLPKDD2008*, 2008, pp. 505–519.
- [187] G. Nahari, A. Vrij, and R. P. Fisher, "Exploiting liars' verbal strategies by examining the verifiability of details," *Leg. Criminol. Psychol.*, vol. 19, no. 2, pp. 227–239, 2014.
- [188] E. J. Williams, "Lies and Cognition: How do we tell lies and can we detect them?," p. 254, 2012.
- [189] M. Monaro, R. Spolaor, L. QianQian, M. Conti, L. Gamberini, and G. Sartori, "Type me the truth!: Detecting deceitful users via keystroke dynamics," in *Proceedings of the 12th International Conference on Availability, Reliability and Security, ARES '17*, 2017, vol. 60.

- [190] D. Norman, J. Miller, and A. Henderson, “What you see, some of what’s in the future, and how we go about doing it: HI at Apple Computer,” in *Proceeding of Conference Companion on Human Factors in Computing Systems, CHI '95*, 1995, p. 155.

Annex 1

Complete list of the 62 attributes extracted from keystroke dynamics tasks.

Number of errors
Prompted-firstdigit
Prompted-firstdigit adjusted GULPEASE
Prompted-enter
Firstdigit-enter
Time before enter key down
Time before enter key flight
Answer length
Writing time
Di-graph down time average
Di-graph down time maximum
Di-graph down time minimum
Di-graph down time median
Di-graph down time standard deviation
Di-graph down time variance
Di-graph up time average
Di-graph up time maximum
Di-graph up time minimum
Di-graph up time median
Di-graph up time standard deviation
Di-graph up time variance
Di-graph up and down time average
Di-graph up and down time maximum
Di-graph up and down time minimum
Di-graph up and down time median
Di-graph up and down time standard deviation
Di-graph up and down time variance
Di-graph press time average
Di-graph press time maximum
Di-graph press time minimum
Di-graph press time median
Di-graph press time standard deviation
Di-graph press time variance
Di-graph flight time average
Di-graph flight time maximum
Di-graph flight time minimum
Di-graph flight time median
Di-graph flight time standard deviation
Di-graph flight time variance
Tri-graph down time average

Tri -graph down time maximum
Tri -graph down time minimum
Tri -graph down time median
Tri -graph down time standard deviation
Tri -graph down time variance
Tri -graph up time average
Tri -graph up time maximum
Tri -graph up time minimum
Tri -graph up time median
Tri -graph up time standard deviation
Tri -graph up time variance
Tri -graph up and down time average
Tri -graph up and down time maximum
Tri -graph up and down time minimum
Tri -graph up and down time median
Tri -graph up and down time standard deviation
Tri -graph up and down time variance
Number of Shift
Number of Del
Number of Canc
Number of Space
Number of Arrows

Annex 2

Details on ML classifiers parameters

Random Forest:

weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

- seed -- The random number seed to be used. = 1
- storeOutOfBagPredictions -- Whether to store the out-of-bag predictions. = FALSE
- numExecutionSlots -- The number of execution slots (threads) to use for constructing the ensemble. = 1
- bagSizePercent -- Size of each bag, as a percentage of the training set size. = 100
- numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model. = 2
- batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size. = 100
- printClassifiers -- Print the individual classifiers in the output. = FALSE
- numIterations -- The number of iterations to be performed. = 100
- debug -- If set to true, classifier may output additional info to the console. = FALSE
- outputOutOfBagComplexityStatistics -- Whether to output complexity-based statistics when out-of-bag evaluation is performed. = FALSE
- breakTiesRandomly -- Break ties randomly when several attributes look equally good. = FALSE
- doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime). = FALSE
- maxDepth -- The maximum depth of the tree, 0 for unlimited. = 0
- calcOutOfBag -- Whether the out-of-bag error is calculated. = FALSE
- numFeatures -- Sets the number of randomly chosen attributes. If 0, $\text{int}(\log_2(\#\text{predictors}) + 1)$ is used. = 0

Logistic:

weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4

- numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model. = 4
- batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size. = 100
- debug -- Output debug information to the console. = FALSE
- ridge -- Set the Ridge value in the log-likelihood. = 1.0E-8
- useConjugateGradientDescent -- Use conjugate gradient descent rather than BFGS updates; faster for problems with many parameters. = FALSE
- maxIts -- Maximum number of iterations to perform. = -1
- doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime). = FALSE

SMO:

weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

- buildCalibrationModels -- Whether to fit calibration models to the SVM's outputs (for proper probability estimates). = FALSE
- numFolds -- The number of folds for cross-validation used to generate training data for calibration models (-1 means use training data). = -1
- randomSeed -- Random number seed for the cross-validation. = 1
- c -- The complexity parameter C. = 1.0
- numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model. = 2
- batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size. = 100
- kernel -- The kernel to use. Polykernel -C 250007 -E 1.0
- checksTurnedOff -- Turns time-consuming checks off - use with caution. = FALSE
- debug -- If set to true, classifier may output additional info to the console. = FALSE
- filterType -- Determines how/if the data will be transformed. = Normalized training data
- toleranceParameter -- The tolerance parameter (shouldn't be changed). = 0.001
- calibrator -- The calibration method to use. = Logistic
- doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime). = FALSE
- epsilon -- The epsilon for round-off error (shouldn't be changed). 1.0E-12

LMT:

weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0

- splitOnResiduals -- Set splitting criterion based on the residuals of LogitBoost. There are two possible splitting criteria for LMT: the default is to use the C4.5 splitting criterion that uses information gain on the class variable. The other splitting criterion tries to improve the purity in the residuals produces when fitting the logistic regression functions. The choice of the splitting criterion does not usually affect classification accuracy much, but can produce different trees. = FALSE
- useAIC -- The AIC is used to determine when to stop LogitBoost iterations. The default is not to use AIC. = FALSE
- numDecimalPlaces -- The number of decimal places to be used for the output of coefficients. = 2
- batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size. = 100
- weightTrimBeta -- Set the beta value used for weight trimming in LogitBoost. Only instances carrying (1 - beta)% of the weight from previous iteration are used in the next iteration. Set to 0 for no weight trimming. The default value is 0. = 0.0

- doNotMakeSplitPointActualValue -- If true, the split point is not relocated to an actual data value. This can yield substantial speed-ups for large datasets with numeric attributes. = FALSE
- debug -- If set to true, classifier may output additional info to the console. = FALSE
- numBoostingIterations -- Set a fixed number of iterations for LogitBoost. If ≥ 0 , this sets a fixed number of LogitBoost iterations that is used everywhere in the tree. If < 0 , the number is cross-validated. = -1
- fastRegression -- Use heuristic that avoids cross-validating the number of Logit-Boost iterations at every node. When fitting the logistic regression functions at a node, LMT has to determine the number of LogitBoost iterations to run. Originally, this number was cross-validated at every node in the tree. To save time, this heuristic cross-validates the number only once and then uses that number at every node in the tree. Usually this does not decrease accuracy but improves runtime considerably. = TRUE
- minNumInstances -- Set the minimum number of instances at which a node is considered for splitting. The default value is 15. = 15
- doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime). = FALSE
- errorOnProbabilities -- Minimize error on probabilities instead of misclassification error when cross-validating the number of LogitBoost iterations. When set, the number of LogitBoost iterations is chosen that minimizes the root mean squared error instead of the misclassification error. = FALSE
- convertNominal -- Convert all nominal attributes to binary ones before building the tree. This means that all splits in the final tree will be binary. = FALSE

Naïve Bayes:**weka.classifiers.bayes.NaiveBayes**

- useKernelEstimator -- Use a kernel estimator for numeric attributes rather than a normal distribution. = FALSE
- numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model. = 2
- batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size. = 100
- debug -- If set to true, classifier may output additional info to the console. = FALSE
- displayModelInOldFormat -- Use old format for model output. The old format is better when there are many class values. The new format is better when there are fewer classes and many attributes. = FALSE
- doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime). = FALSE
- useSupervisedDiscretization -- Use supervised discretization to convert numeric attributes to nominal ones. = FALSE

J48:**weka.classifiers.trees.J48**

- batchSize -- The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size. = 100
- binarySplits -- Whether to use binary splits on nominal attributes when building the trees. = FALSE
- collapseTree -- Whether parts are removed that do not reduce training error. = TRUE
- confidenceFactor -- The confidence factor used for pruning (smaller values incur more pruning). = 0.25
- debug -- If set to true, classifier may output additional info to the console. = FALSE
- doNotCheckCapabilities -- If set, classifier capabilities are not checked before classifier is built (Use with caution to reduce runtime). = FALSE
- doNotMakeSplitPointActualValue -- If true, the split point is not relocated to an actual data value. This can yield substantial speed-ups for large datasets with numeric attributes. = FALSE
- minNumObj -- The minimum number of instances per leaf. = 2
- numDecimalPlaces -- The number of decimal places to be used for the output of numbers in the model. = 2
- numFolds -- Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the tree. = 3
- reducedErrorPruning -- Whether reduced-error pruning is used instead of C.4.5 pruning. = FALSE
- saveInstanceData -- Whether to save the training data for visualization. = FALSE
- seed -- The seed used for randomizing the data when reduced-error pruning is used. = 1
- subtreeRaising -- Whether to consider the subtree raising operation when pruning. = TRUE
- unpruned -- Whether pruning is performed. = FALSE
- useLaplace -- Whether counts at leaves are smoothed based on Laplace. = FALSE
- useMDLcorrection -- Whether MDL correction is used when finding splits on numeric attributes. = TRUE

Annex 3

Complete list of the 23 raw attributes and 18 normalized predictors extracted from RT and complex questions task.

Raw predictors	
Total RT	Mean Reaction Time on all test questions
Total Yes RT	Mean Reaction Time on all test questions that required a YES response
Total No RT	Mean Reaction Time on all test questions that required a NO response
Control Tot RT	Mean Reaction Time on all control questions
Control Yes RT	Mean Reaction Time on control questions that required a YES response
Control No RT	Mean Reaction Time on control questions that required a NO response
Simple Tot RT	Mean Reaction Time on all simple questions
Simple Yes RT	Mean Reaction Time on simple questions that required a YES response
Simple No RT	Mean Reaction Time on simple questions that required a NO response
Complex Tot RT	Mean Reaction Time on all complex questions
Complex Yes RT	Mean Reaction Time on complex questions that required a YES response
Complex No RT	Mean Reaction Time on complex questions that required a NO response
Mean Total errors	Mean number of errors on all test questions (number of errors on all test questions divided by total number of stimuli on entire test)
Mean Control Tot errors	Mean number of errors on all control questions (number of errors on all control questions divided by total number of control stimuli)
Mean Simple Tot errors	Mean number of errors on all simple questions (number of errors on all simple questions divided by total number of simple stimuli)

Mean Complex Tot errors	Mean number of errors on all complex questions (number of errors on all complex questions divided by total number of complex stimuli)
Raw Total errors	Total number of errors on all test questions
Raw Control Tot errors	Number of errors on all control questions
Raw Simple Tot errors	Number of errors on all simple questions
Raw Complex Tot errors	Number of errors on all complex questions
IES Control	Inverse Efficiency Score calculated for RT and errors of control questions
IES Simple	Inverse Efficiency Score calculated for RT and errors of simple questions
IES Complex	Inverse Efficiency Score calculated for RT and errors of complex questions
Normalized predictors	
Total Yes RT/Total RT	Ratio between RT on all test questions that required a YES response and RT on all test questions
Total No RT/Total RT	Ratio between RT on all test questions that required a NO response and RT on all test questions
Control Tot RT/Total RT	Ratio between RT on all control questions and RT on all test questions
Control Yes RT/Total RT	Ratio between RT on control questions that required a YES response and RT on all test questions
Control No RT/Total RT	Ratio between RT on control questions that required a NO response and RT on all test questions
Simple Tot RT/Total RT	Ratio between RT on all simple questions and RT on all test questions
Simple Yes RT/Total RT	Ratio between RT on simple questions that required a YES response and RT on all test questions
Simple No RT/Total RT	Ratio between RT on simple questions that required a NO response and RT on all test questions

Complex Tot RT/Total RT	Ratio between RT on all complex questions and RT on all test questions
Complex Yes RT/Total RT	Ratio between RT on complex questions that required a YES response and RT on all test questions
Complex No RT/Total RT	Ratio between RT on complex questions that required a NO response and RT on all test questions
(Total Yes RT - Total No RT)/Total RT	Ratio of the difference between the RT on all test questions that required a YES response and the RT on all test questions that required a NO response with the RT on all test questions
(Control Yes RT - Control No RT)/Total RT	Ratio of the difference between the RT on control questions that required a YES response and the RT on control questions that required a NO response with the RT on all test questions
(Simple Yes RT - Simple No RT)/Total RT	Ratio of the difference between the RT on simple questions that required a YES response and the RT on simple questions that required a NO response with the RT on all test questions
(Complex Yes RT - Complex No RT)/Total RT	Ratio of the difference between the RT on complex questions that required a YES response and the RT on complex questions that required a NO response with the RT on all test questions
Raw Simple Tot errors/Raw Control Tot errors	Ratio between the number of errors on all simple questions and the number of errors on all control questions
Raw Complex Tot errors/Raw Control Tot errors	Ratio between the number of errors on all complex questions and the number of errors on all control questions
Raw Complex Tot errors/Raw Simple Tot errors	Ratio between the number of errors on all complex questions and the number of errors on all simple questions

Annex 4

Complete list of the 326 raw attributes predictors extracted from mouse dynamics and complex questions task related to deception about holidays.

control_average_error	simple_sd_MD-time	complex_average_vel(y)
control_average_IT	simple_sd_x-flip	complex_average_acc(y)
control_average_RT	simple_sd_y-flip	complex_sd_vel(x)
control_average_MD	simple_min_vel(x)	complex_sd_acc(x)
control_average_AUC	simple_max_vel(x)	complex_sd_vel(y)
control_average_MD-time	simple_min_vel(y)	complexsd_acc(x)(y)
control_average_x-flip	simple_max_vel(y)	complex_YES_average_error
control_average_y-flip	simple_min_acc(x)	complex_YES_average_IT
control_sd_error	simple_max_acc(x)	complex_YES_average_RT
control_sd_IT	simple_min_acc(y)	complex_YES_average_MD
control_sd_RT	simple_max_acc(y)	complex_YES_average_AUC
control_sd_MD	simple_average_vel(x)	complex_YES_average_MD-time
control_sd_AUC	simple_average_acc(x)	complex_YES_average_x-flip
control_sd_MD-time	simple_average_vel(y)	complex_YES_average_y-flip
control_sd_x-flip	simple_average_acc(y)	complex_YES_sd_error
control_sd_y-flip	simple_sd_vel(x)	complex_YES_sd_IT
control_min_vel(x)	simple_sd_acc(x)	complex_YES_sd_RT
control_max_vel(x)	simple_sd_vel(y)	complex_YES_sd_MD
control_min_vel(y)	simple_sd_acc(y)	complex_YES_sd_AUC
control_max_vel(y)	simple_YES_average_error	complex_YES_sd_MD-time
control_min_acc(x)	simple_YES_average_IT	complex_YES_sd_x-flip
control_max_acc(x)	simple_YES_average_RT	complex_YES_sd_y-flip
control_min_acc(y)	simple_YES_average_MD	complex_YES_min_vel(x)
control_max_acc(y)	simple_YES_average_AUC	complex_YES_max_vel(x)
control_average_vel(x)	simple_YES_average_MD-time	complex_YES_min_vel(y)
control_average_acc(x)	simple_YES_average_x-flip	complex_YES_max_vel(y)
control_average_vel(y)	simple_YES_average_y-flip	complex_YES_min_acc(x)
control_average_acc(y)	simple_YES_sd_error	complex_YES_max_acc(x)
control_sd_vel(x)	simple_YES_sd_IT	complex_YES_min_acc(y)
control_sd_acc(x)	simple_YES_sd_RT	complex_YES_max_acc(y)
control_sd_vel(y)	simple_YES_sd_MD	complex_YES_average_vel(x)
control_sd_acc(y)	simple_YES_sd_AUC	complex_YES_average_acc(x)
control_YES_average_error	simple_YES_sd_MD-time	complex_YES_average_vel(y)
control_YES_average_IT	simple_YES_sd_x-flip	complex_YES_average_acc(y)
control_YES_average_RT	simple_YES_sd_y-flip	complex_YES_sd_vel(x)
control_YES_average_MD	simple_YES_min_vel(x)	complex_YES_sd_acc(x)
control_YES_average_AUC	simple_YES_max_vel(x)	complex_YES_sd_vel(y)
control_YES_average_MD-time	simple_YES_min_vel(y)	complex_YES_sd_acc(y)
control_YES_average_x-flip	simple_YES_max_vel(y)	complex_NO_average_error
control_YES_average_y-flip	simple_YES_min_acc(x)	complex_NO_average_IT
control_YES_sd_error	simple_YES_max_acc(x)	complex_NO_average_RT
control_YES_sd_IT	simple_YES_min_acc(y)	complex_NO_average_MD
control_YES_sd_RT	simple_YES_max_acc(y)	complex_NO_average_AUC

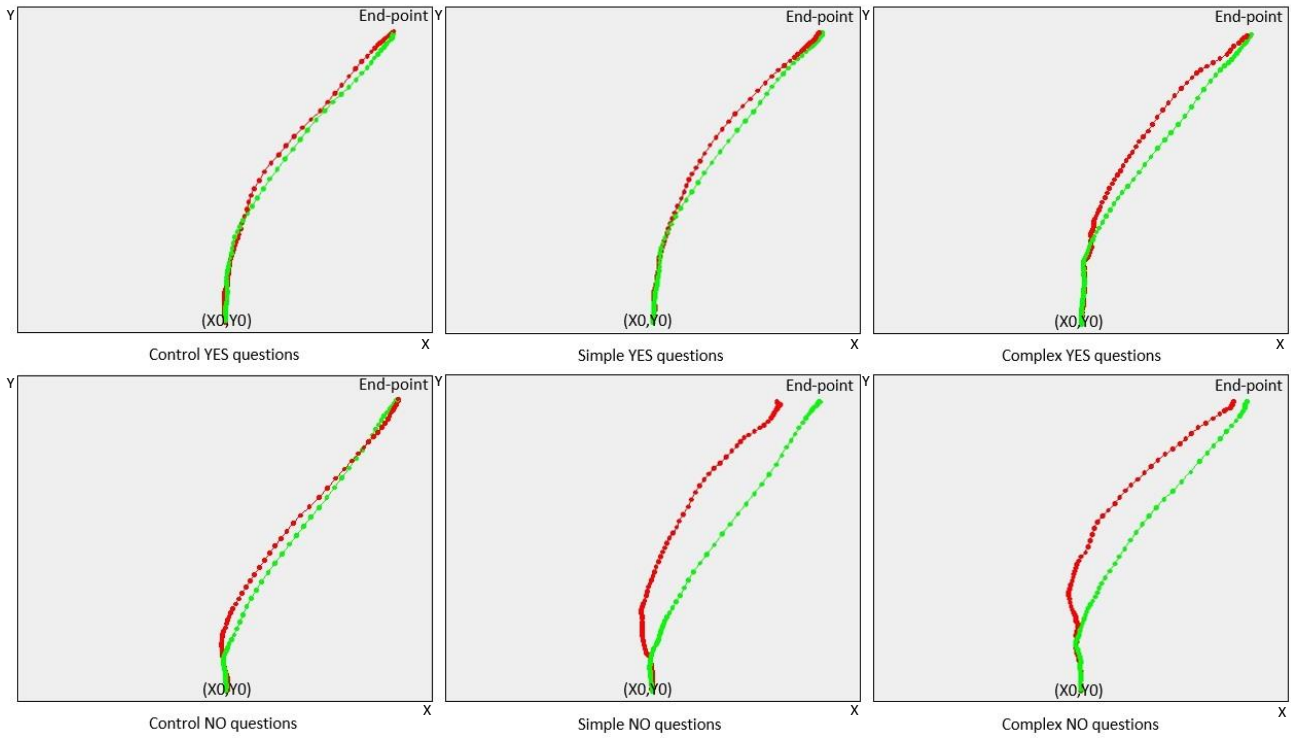
control_YES_sd_MD	simple_YES_average_vel(x)	complex_NO_average_MD-time
control_YES_sd_AUC	simple_YES_average_acc(x)	complex_NO_average_x-flip
control_YES_sd_MD-time	simple_YES_average_vel(y)	complex_NO_average_y-flip
control_YES_sd_x-flip	simple_YES_average_acc(y)	complex_NO_sd_error
control_YES_sd_y-flip	simple_YES_sd_vel(x)	complex_NO_sd_IT
control_YES_min_vel(x)	simple_YES_sd_acc(x)	complex_NO_sd_RT
control_YES_max_vel(x)	simple_YES_sd_vel(y)	complex_NO_sd_MD
control_YES_min_vel(y)	simple_YES_sd_acc(y)	complex_NO_sd_AUC
control_YES_max_vel(y)	simple_NO_average_error	complex_NO_sd_MD-time
control_YES_min_acc(x)	simple_NO_average_IT	complex_NO_sd_x-flip
control_YES_max_acc(x)	simple_NO_average_RT	complex_NO_sd_y-flip
control_YES_min_acc(y)	simple_NO_average_MD	complex_NO_min_vel(x)
control_YES_max_acc(y)	simple_NO_average_AUC	complex_NO_max_vel(x)
control_YES_average_vel(x)	simple_NO_average_MD-time	complex_NO_min_vel(y)
control_YES_average_acc(x)	simple_NO_average_x-flip	complex_NO_max_vel(y)
control_YES_average_vel(y)	simple_NO_average_y-flip	complex_NO_min_acc(x)
control_YES_average_acc(y)	simple_NO_sd_error	complex_NO_max_acc(x)
control_YES_sd_vel(x)	simple_NO_sd_IT	complex_NO_min_acc(y)
control_YES_sd_acc(x)	simple_NO_sd_RT	complex_NO_max_acc(y)
control_YES_sd_vel(y)	simple_NO_sd_MD	complex_NO_average_vel(x)
control_YES_sd_acc(y)	simple_NO_sd_AUC	complex_NO_average_acc(x)
control_NO_average_error	simple_NO_sd_MD-time	complex_NO_average_vel(y)
control_NO_average_IT	simple_NO_sd_x-flip	complex_NO_average_acc(y)
control_NO_average_RT	simple_NO_sd_y-flip	complex_NO_sd_vel(x)
control_NO_average_MD	simple_NO_min_vel(x)	complex_NO_sd_acc(x)
control_NO_average_AUC	simple_NO_max_vel(x)	complex_NO_sd_vel(y)
control_NO_average_MD-time	simple_NO_min_vel(y)	complex_NO_sd_acc(y)
control_NO_average_x-flip	simple_NO_max_vel(y)	all questions_average_error
control_NO_average_y-flip	simple_NO_min_acc(x)	all questions_average_IT
control_NO_sd_error	simple_NO_max_acc(x)	all questions_average_RT
control_NO_sd_IT	simple_NO_min_acc(y)	all questions_average_MD
control_NO_sd_RT	simple_NO_max_acc(y)	all questions_average_AUC
control_NO_sd_MD	simple_NO_average_vel(x)	all questions_average_MD-time
control_NO_sd_AUC	simple_NO_average_acc(x)	all questions_average_x-flip
control_NO_sd_MD-time	simple_NO_average_vel(y)	all questions_average_y-flip
control_NO_sd_x-flip	simple_NO_average_acc(y)	all questions_sd_error
control_NO_sd_y-flip	simple_NO_sd_vel(x)	all questions_sd_IT
control_NO_min_vel(x)	simple_NO_sd_acc(x)	all questions_sd_RT
control_NO_max_vel(x)	simple_NO_sd_vel(y)	all questions_sd_MD
control_NO_min_vel(y)	simple_NO_sd_acc(y)	all questions_sd_AUC
control_NO_max_vel(y)	complex_average_error	all questions_sd_MD-time
control_NO_min_acc(x)	complex_average_IT	all questions_sd_x-flip
control_NO_max_acc(x)	complex_average_RT	all questions_sd_y-flip
control_NO_min_acc(y)	complex_average_MD	all questions_min_vel(x)
control_NO_max_acc(y)	complex_average_AUC	all questions_max_vel(x)
control_NO_average_vel(x)	complex_average_MD-time	all questions_min_vel(y)
control_NO_average_acc(x)	complex_average_x-flip	all questions_max_vel(y)
control_NO_average_vel(y)	complex_average_y-flip	all questions_min_acc(x)

control_NO_average_acc(y)	complex_sd_error	all questions_max_acc(x)
control_NO_sd_vel(x)	complex_sd_IT	all questions_min_acc(y)
control_NO_sd_acc(x)	complex_sd_RT	all questions_max_acc(y)
control_NO_sd_vel(y)	complex_sd_MD	all questions_average_vel(x)
control_NO_sd_acc(y)	complex_sd_AUC	all questions_average_acc(x)
simple_average_error	complex_sd_MD-time	all questions_average_vel(y)
simple_average_IT	complex_sd_x-flip	all questions_average_acc(y)
simple_average_RT	complex_sd_y-flip	all questions_sd_vel(x)
simple_average_MD	complex_min_vel(x)	all questions_sd_acc(x)
simple_average_AUC	complex_max_vel(x)	all questions_sd_vel(y)
simple_average_MD-time	complex_min_vel(y)	all questions_sd_acc(y)
simple_average_x-flip	complex_max_vel(y)	control_X75
simple_average_y-flip	complex_min_acc(x)	control_Y30
simple_sd_error	complex_max_acc(x)	simple_X75
simple_sd_IT	complex_min_acc(y)	simple_Y30
simple_sd_RT	complex_max_acc(y)	complex_X75
simple_sd_MD	complex_average_vel(x)	complex_Y30
simple_sd_AUC	complex_average_acc(x)	

Annex 5

Average trajectories of truth-tellers (in green), liars A (in red) and liars B (in orange), separately for control, simple and complex questions requiring a “yes” or “no” response. Data refer to mouse dynamics and complex questions task related to deception about holidays.

Truth-tellers vs liars A



Truth-tellers vs liars B

