The University of Nottingham
Faculty of Science
School of Computer Science

Exploring Deception Using Automatically Detected Facial Action Units

Doratha Vinkemeier, Diplom Informatikerin

Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science of The University of Nottingham. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

August 2020

Abstract

Lie detection has always gripped mankind. Today, applications range from individual employee screening to mass terror scenarios. Yet, signs of deception are still not well understood and there is general agreement that humans are bad at detecting them. We suffer from bias and subjectivity as well as a lack of stamina and observational acuity. For this reason, hope is now being placed on automatic methods such as action unit (AU) detectors, which detect facial muscle movements that can reveal affective states.

Automatic AU detectors are still in their developmental stage; they are proving, however, to be useful for both detecting and learning about deception. This thesis used CNN-BLSTM and OpenFace AU detectors and decision trees in two different deception scenarios. In one of them, the game of poker, deception is integral and desirable. Videos obtained from the University of Southern California showed pairs of players who communicated over a network and behaved spontaneously in a laboratory setting. I ascertained that players who were folding, as opposed to calling or raising, displayed significantly more AU12 and AU5, action units associated with smiling and other emotions, whereby CNN-BLSTM and OpenFace showed only limited overlap.

The study of deceit is hindered also by the lack of relevant datasets that simultaneously have a ground truth. For that reason, the second part of my thesis was dedicated to building and researching such a dataset - the dice rolling experiment - where participants roll a virtual die and decide themselves whether or not to lie to increase their earnings. This dataset consists of over 1.7 million frames of good quality video along with concurrent mouse tracking information and timestamps of events covering 373 different subjects. It has a defined ground truth and also investigates the effects of cold water stress on deceptive behaviour. This experiment revealed that males lied more than females and that stress reduced lying. Low detection levels and distinct patterns of false positive facial AUs lead me to use head pose estimators, which showed that under stress, deceptive participants moved their heads significantly more than honest ones.

In summary, this study automatically detected scenario-specific clues of deception, explored the limitations of current AU detectors, and generated a large, novel data set uniquely suitable for studying deception and its automatic detection.

Acknowledgements

I would like to thank my supervisor, Dr. Michel Valstar, and my co-supervisor, Professor Christian Wagner, for their expertise and support. I am grateful to Professor Bob John and Dr. Per Kristian Lehre for encouraging me to begin a PhD and supporting me in my successful application for a Vice Chancellor's Scholarship for Research Excellence at the University of Nottingham. My special thanks go out to Professor Roberto Hernán-González and Professor Thorsten Chmura for sharing expert advice in the art of behavioural economics and their mentorship. They made my PhD exciting and working with them has been a great experience. I would also like to gratefully acknowledge Dr. Jose Guinot Sapporta for showing me the ropes of running behavioural economic experiments in CEDEX and always making sure I had a helping hand when needed. Kudos to Victor Huddleston and Joseph Best for making sure I had plenty of computing power throughout my PhD. Many thanks to Aaron Jackson for advice on Matlab, especially at the beginning when it was new to me. My office mates, Tom Smith and Dimitrios Bellos, earned my praise for being great company as well as frequently giving good advice on various matters, Dr. Joy Egede for critically reading this thesis, and Dr. Shashank Jaiswal for providing me with his excellent AU detectors and for also giving me helpful advice.

Last but not least, I would like to express my love and gratitude to my family, Uwe, Isabel and Bertie. Their affection and encouragement got me through the toughest times including the writing up of my thesis under Coronavirus lockdown. I am indebted to my husband for reading my thesis and exchanging thoughts on a subject that is not his own. Uwe, the first million will be yours alone. He learned more computer science than he ever intended to and I think he likes it. I would also like to express my love and gratitude to my mother, Nancy, who has unconditionally supported and loved me throughout my life, and to my father, Dee, whose love and unfaltering encouragement have been a comfort and inspiration to me.

"Überall lernt man nur von dem, den man liebt."

– Johann Wolfgang von Goethe

Contents

Chapter 1

Introduction - Exploring Deception Using Automatically Detected Facial Action Units

Human behaviour is rife with deception. If one takes deception to be an intentional attempt to get others to believe something which is not true, then even the most upright are often guilty of telling white lies. Accordingly, humans have made a huge effort to detect when others are being deceitful in order to get to the truth. Examples range from a person trying to find out if their spouse has been unfaithful, a judge or juror punishing crime to commercial negotiators or diplomats needing to establish if their counterpart is trustworthy. Detecting deception is a much sought after ability and whoever can do it is believed to be at an advantage.

Attempts to detect deception are not new. In the Middle Ages, in criminal cases where it was otherwise impossible to establish the facts, the judgement of God was frequently called upon and $Trial\ by\ Ordeal$ used. An accused person could agree to such a trial to prove their innocence and risk serious injury, such as having their "hand boiled to rags", or worse (Leeson, 2012) . Nowadays, lie detection techniques have turned to seemingly more scientific methods, although it is hotly disputed whether even the most common modern method of lie detection, the polygraph, is effective or scientifically sound (Bell, 2012). The ambivalence concerning the polygraph can be seen by the fact that even though the US Supreme Court has declared polygraph evidence to be hearsay, polygraphs are still frequently used in legal as well as many other settings such as the government, industry and the private sector, turning the polygraph into a booming multibillion dollar

industry (Harris, 2018).

Nowadays, people come together in huge masses - at concerts, at malls and shopping areas, in sports stadiums and in airports - to name a few examples. This makes them vulnerable to terrorist attacks and they are aware of this. To deal with these situations, it has become an urgent matter to come up with lightweight methods of deception detection that can be deployed on a large scale, for which the polygraph, which is very time intensive and expensive, is not suitable. In the US, in the case of airports, the Department of Homeland Security's Transportation Security Administration has been training specialists called Behavioural Detection Officers (BDOs) whose job it is to visually "identify passenger behaviors indicative of stress, fear or deception" (U.S. Government Accountability Office, 2013). These are considered to be indicators of mal-intent, which is intent to harm. The program is called *Screening of Passengers by Observation Techniques* (SPOT) and has an annual budget of $200 million (Weinberger, 2010). SPOT has been heavily criticised by many, including the United States government's own accountability agency, the Government Accountability Office (GAO), which published a congressional report recommending the program be discontinued until there is proof of its effectiveness (U.S. Government Accountability Office, 2013). The GAO argued that human observation unaided by technology is not a reliable means for detecting deceit. One of the main criticisms of programs like SPOT, is that humans detect deceit no better than chance, that they are subjective and suffer from other shortcomings. The GAO did suggest that automated technologies might overcome many human limitations regarding detecting telltale signs of deception and mal-intent. They might solve problems of fatigue, bias and subjectivity and they could perhaps notice things the naked human eye might miss. While this approach shows promise, a RAND report (Davis et al., 2013) found that these technologies were currently only in their infancy.

Computer vision and machine learning offer the promise of providing objective, repeatable, ubiquitous and inexpensive tools to study and detect human behaviour such as deceit. Machine learning techniques are capable of detecting patterns in data that humans cannot discern, making it possible that automatic methods can discover unknown behavioural markers, provided that the data they learn from is representative of the problem at hand (Kirkpatrick, 2017).

There are several supposed manifestations of deception, including body pose, voice tone changes (DePaulo et al., 2003) and certain physiological markers like blood pressure

(ten Brinke et al., 2015). One of the most commonly studied sources of clues to deceit is the face (Stel and van Dijk, 2018). As yet, there is no strong scientific link between facial cues and a person's intentions, though, and this needs to be explored. Systems using automatic detection of facial expressions are a field of rapidly advancing research with many applications in human computer interactions, gaming and advertising (Martinez et al., 2019; Cohn and De La Torre, 2015; Pantic and Bartlett, 2007), including a burgeoning field of research into automatic detection of affective behaviour in medicine (Valstar, 2014). It has long been recognized that human affect plays a decisive role in human interactions, whether they be human-human or human-computer, and as the human face is one of the richest sources of affective information, efforts to capture and automatically understand this will grow (Picard, 2000). In addition, facial expression recognition systems are already being deployed at places like border controls and airports to screen people to detect risks such as potential terrorist attacks (Rothwell et al., 2006). Given these high stakes, it is thus important to establish whether these technologies are truly effective particularly concerning deceptive behaviours. The research presented in this thesis aims to further our knowledge in this scientific area.

1.1 Motivation

This thesis is a study in automatic detection of deception based on visual cues of the face. Up until now, there have only been a few such studies. This is primarily because of a lack of real-life deception scenarios in a setting that is also appropriate for a computer vision study. Also, due to ethical reasons, data involving deception is often not made public and so not easily accessible to researchers. The motivation of this work is to investigate the feasibility of using automatic detection of facial expressions to determine if a person is being deceptive or not. Machine learning can potentially revolutionize our understanding of human facial expressions and of human behaviour. Computers and webcams are ubiquitous and we are already being analysed by them, although the behavioural science behind it is still developing. Therefore, the purpose of this study is to investigate the scientific basis and feasibility of automatic detection of deception.

1.2 Contributions

This thesis

1. extends earlier works on automatic detection of deceit by investigating deceit in more spontaneous settings than previous studies, namely, in poker game play and dice rolling.

2. analyses automatically detected facial features and, to a lesser extent, head movements with decision trees. Decision trees have a very clear structure and can provide an understanding of what is being classified and why.

3. presents a new dice rolling database in which the deceit displayed is not posed and is solely the decision of the study subject. This is different from most other databases where the subject displaying deceit has been instructed to do so. This database was made in conjunction with behavioural economists to also be a genuine investigation into human decision making and human-computer interactions. It rigorously follows the methods of behavioural economics and to my knowledge is the first such database.

4. presents a behavioural economics analysis of the dice rolling study.

5. presents an initial computer vision and machine learning analysis of the dice rolling database.

6. provides metadata from the new dice rolling database for future investigations.

1.3 Publication

Doratha Vinkemeier, Michel Valstar and Jonathan Gratch (2018). "Predicting Folds in Poker Using Action Unit Detectors and Decision Trees." Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, pages 504-511.

1.4 Thesis structure

This thesis is structured as follows:

Chapter 2 expounds basic concepts and tools that will be used in the thesis: a short history of deception detection, three models of facial expression frequently used in affective computing and the tools for detecting facial muscle activity that are used in this thesis.

Chapter 3 reviews previously published related works on automatic detection of deceit.

Chapter 4 presents my research on detecting deceit in poker game play.

Chapter 5 introduces a new dice rolling database that is a behavioural economics study of decision making and deceit involving human-computer interactions.

Chapter 6 presents an automatic analysis of the aforementioned dice rolling database.

Chapter 7 concludes the thesis with a discussion and interpretation of results and suggestions for future research.

Chapter 2

Background - Detection of Deceit

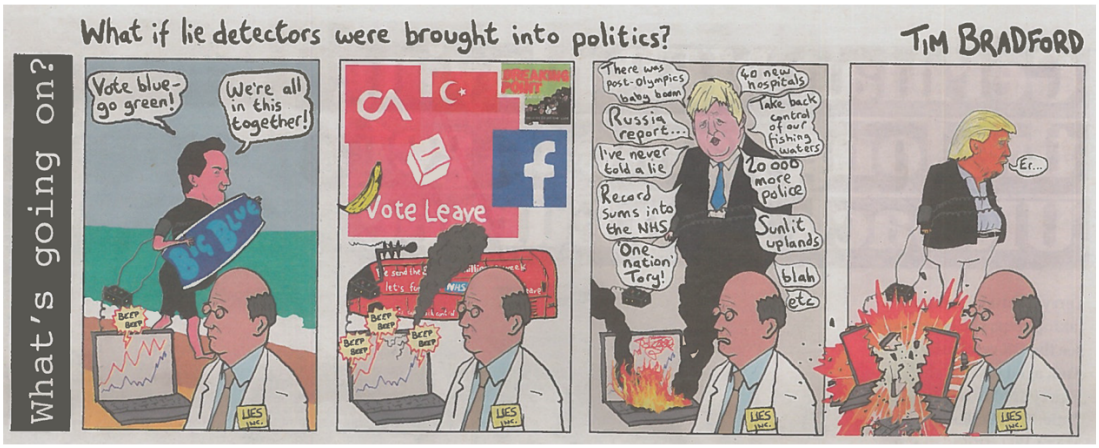

Figure 2.1: Cartoon from the *New European* (Bradford, 2020).

2.1 Description of deceit

Deceit is often defined to be an intentional attempt to cause someone to take as true something which is false, usually for some sort of personal gain. There is a large spectrum of types of deceit from lies that swindle a person out of their life and possessions to untrue flattery that people tell every day in social settings in order to be liked. Though completely condemned by Aristotle (384-322 BC), deception was considered a legitimate tool of the state by Plato (ca. 424-344 BC) in his *Republic* (Roochnik, 2005). He believed it could be justified in some cases. For instance, he posited that creating a myth for people to believe in would make them live together more harmoniously. Machiavelli (1469-1527) advised in *The Prince* (Machiavelli, 2010, reprinted August 2018)

that a ruler should be good at the art of deception and know when to apply it in order to maintain power among immoral and vulgar people. On the other hand, a ruler must simultaneously appear to his subjects to be virtuous and honest to inspire their esteem as nothing could be more detrimental to his authority than to have his subjects despise him. If deceit can be unconscious, then animals and plants, who for instance use camouflage and mimicry, also practice it and it is most likely ingrained deep in our own biology. As ubiquitous as deception is and as advantageous as it may seem to be, it has been looked down upon by society going back as far as Aristotle, who condemned it as immoral and damaging to society and oneself (Zembaty, 1993). Deceit remains a complex subject. There are many forms of it and many opinions about it. While it might be condemned socially, it is also considered by many to be an indispensable, pervasive part of everyday life for everyone (DePaulo et al., 1996). This is based on the idea that as social animals we fashion a version of ourselves for the purpose of creating an impression of ourselves that we want the world to see (Goffman, 1990). This edited version of ourselves is not the same as our true selves and is in itself an act of deception, even self-deception. To keep our 'face' and to preserve the face of those we interact with, we have developed a complex form of politeness (Brown and Levinson, 1987). This politeness requires the occasional white lie, such as telling a friend a lie in order to spare their feelings or lying in order to preserve one's own social image, for instance as being a kind person. Thus, deception is likely necessary for society to function.Despite any ambiguity about it, deceit has been shown to have a corrosive effect on society and nowadays there are even ways to quantify this (Gächter and Schulz, 2016).

2.2 Early methods of deception detection

As much as deception has been instrumentalized to gain advantage, so have the attempts to crack it, among other things with lie detectors. Wouldn't it be great to have a device, like that in Figure 2.1, which simply tells you whether a person is lying or not? The history of lie detection goes back at least 3000 years ago. In China, to determine if a person was telling the truth or not, they had to fill their mouth with rice and then spit it out some time later. If the rice was wet, they were deemed to be telling the truth. If it was dry, they were deemed to be lying. This test was based on the physiological observation that people who are nervous and afraid usually have dry mouths (Vicianova, 2015). Since

this physiological state was not very well understood at the time, and since it does not follow from being nervous or afraid that one is being deceptive, it is likely that many innocent people were found guilty and executed. In Europe in the Middle Ages, Trial by Ordeal was used to extract the truth. In this case, the accused was put through some test, usually very injurious or even deadly. If they passed the test, they were innocent and whether or not they passed was viewed as God's decision. For instance, a person accused of a crime might be put through the water test in which they were dunked under water. If they floated they were found guilty, if they sank they were innocent. Trials by Ordeal continued until the 1700s (Vicianova, 2015).

In the 18th century the idea emerged that criminal behaviour, including lying and deception, is an inherited trait. An early proponent was Franz Joseph Gall (1758-1828), who examined the physiognomy of the human skull (Eberle, 2008). He correctly localized the brain region responsible for speech, and claimed that a person's character can be determined from their shape of the head. This area of research (Schädelkunde, engl. phrenology) blossomed in the 19th century, and reached a peak with the theory of anthropological criminology by the Italian physician Cesare Lombroso (1835-1909). It stated that criminals and liars are born as such and could be identified by the head shape and other physical traits (Tanner, 2019). Phrenology was popular into the 20th century, see Figure 2.2, when it was recognized as pseudoscience and fell into disrepute (The Editors of Encyclopaedia Britannica, 2018).

2.3 The polygraph and how to beat it

Today, the polygraph is the most common type of lie detector and its use is widespread. It was invented by John Larson and Leonard Keele in 1921 and it measures physiological changes in a person, such as their blood pressure, breathing rate and galvanic skin response, which indicate stress (Marsh, 2019). These changes are deemed to be associated with telling lies (APA Editorial, 2004). To measure these changes, a person is hooked up to the polygraph machine and then they are asked a series of questions. These consist of control questions and relevant questions. The purpose of the control questions is to determine the subject's normal respiratory and heart rates and galvanic skin pressure. The relevant questions are those related to the crime or issue at hand that one wants to find out the truth about. The assumption behind the efficacy of the polygraph is that a guilty

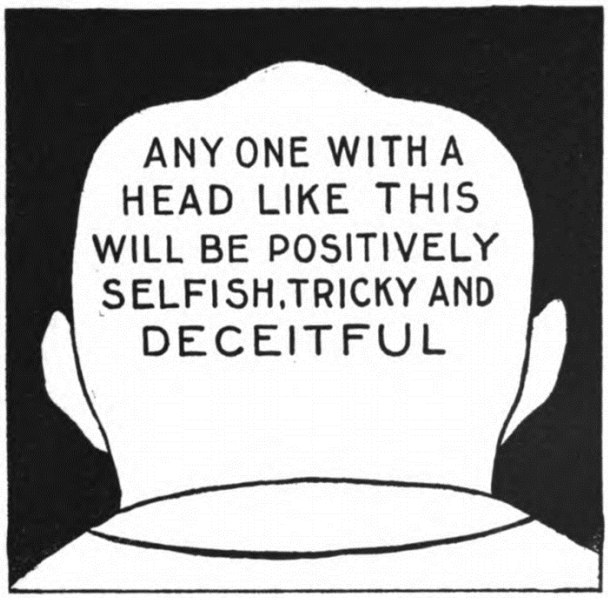

Figure 2.2: Illustration advising how to detect a liar from *Vaught's Practical Character Reader* (Vaught, 2010 (first published 1902).

person will find the relevant questions and lying about them stressful, and an innocent person will not, and the polygraph will detect this difference. However, the physiological changes measured by the polygraph are not unique to deception. It is possible that an innocent person fails the test, as Floyd Fay did in 1978, receiving a life sentence for murder and narrowly missing the death penalty (Kennedy, 2020). He was only acquitted after the real murderers confessed. It is also possible that a guilty person passes the test, as the infamous Russian spy Aldrich Ames did twice. Years later, Ames' flamboyant lifestyle aroused suspicion and he was put under FBI and CIA surveillance. This led to his finally getting caught in 1994, after causing the deaths or imprisonments of several CIA agents. When asked how he managed to pass the polygraph tests, he replied that his Russian handlers, who understood the flaws of the polygraph, had simply told him, "Just relax, don't worry, you have nothing to fear" (Hart, 2020).

According to the British Psychological Society, the accuracy of the polygraph in detecting guilty individuals is about 85 percent, that is, if you are guilty it will likely pick you out. The accuracy of polygraphs in correctly detecting innocent people, in contrast, is estimated to be as low as 50 percent, which means if you are innocent you have about a 50 percent chance of being found guilty (Bell, 2012). If one were to apply the polygraph to 1000 people of which one had actually committed the crime, it would single out around 500 as having failed the test. Perhaps it would also single out the one guilty

person, but even this is not guaranteed. Given that the polygraph is also time consuming and expensive to administer, at over £500 each test, even using its results as a rough guideline to narrow down a search from among a large number of suspects is infeasible (UK Test).

In most courts in the US and the UK polygraph evidence is either not allowed at all or is considered to be nothing more than hearsay (Ewaschuck, 1978; Rothwell et al., 2006). Yet, it is a multi-billion dollar industry (Bittle, 2020). It is used by the FBI, CIA, and police forces in the US and UK. Recently the UK government started a program for subjecting paroled sex offenders to lie detector tests (Bowcott, 2020). There are also plans to use them on people convicted of domestic violence (Grierson, 2019), or terrorist offences (Grierson, 2020) on release from prison. There is also a large industry around lie detector tests in the private sector, where they are used, for example, to screen potential employees. Still, in settings like concerts, sports stadiums and airports where thousands of people come together and are often pressed for time, lie detectors are too cumbersome to be suitable for detecting deception related to mal-intent.

2.4 Facial expressions and emotions

It is widely believed that a person's face can reveal many things about them, including what they are thinking or feeling (Stel and van Dijk, 2018). Facial expressions are considered more reliable than words; when a person tells us how they are feeling, we often look to their face for confirmation or even a clearer understanding of what they mean. When they assert something and we fear it may be a lie, we look to their face for proof, for while it is easy to manipulate words it is more difficult, so we believe at least, to manipulate one's facial expression. The human face has been an object of scientific study going back at least as far as the neurologist Duchenne de Boulogne (1806–1875), who methodically studied the role of facial muscles in forming facial expressions (Duchenne de Boulogne, 1990), and the biologist Charles Darwin (1809–1882), who studied human and animal facial expressions in the context of the theory of evolution, (Darwin, 2009). These two contemporaries collaborated with one another. Darwin believed that facial expressions developed as part of evolutionary adaptation and that they were universal, not only across cultures, but across species. He also believed that they served as visible signals of communication. Duchenne systematically mapped facial muscles to facial ex-

pressions to show that each emotion display corresponded to a specific combination of stimulated facial muscles. He did this by electrically stimulating the facial muscles of living human subjects to form facial expressions that could be recognized by an observer as emotions such as happiness. The works of Darwin and Duchenne laid the scientific basis for a theory of facial expressions and their link with emotions.

Darwin supported his idea of the universality of facial expressions by grounding it in the theory of evolution and making many observations of the expressions of humans and animals. To gather more evidence, he asked friends and acquaintances living in distant lands questions about the facial expressions they observed among the natives where they lived. He also used his five years travel on the Beagle to make observations. He is the founder of the $judgement\ study$, which consists of showing an observer photos of human facial expressions and asking them what emotion they think is being displayed (Ekman, 2003); Ekman (2009). Practitioners of the judgement study, such as Paul Ekman, claim that the idea of the universality of facial expressions has been strengthened by these studies. For references to some of these studies see Ekman and Friesen (2003), pages 32-33.

However, cross-cultural studies of the universality of expressions have drawn criticism, too. It has been argued that the pictures used for displaying facial expressions were not really valid as the facial displays were nearly always posed and exaggerated and taken out of context. The possible answers the study subjects could give were also limited. Furthermore, for some cultures emotional categories are foreign and the way that the test were organized prevented the subjects from inferring non-emotional categories from photos, such as social intent (Barrett et al., 2019). There are also theories that oppose Ekman's theories of basic emotions. One is the $behavioural\ ecology\ view$ (BECV) that facial displays do not reveal anything about a person's internal emotional state, but rather developed as tools for communicating social motives and intentions. They are learned through social interactions, hence they are not biologically determined and immutable, but flexible and can only be understood in the context in which they occur (Crivelli and Fridlund, 2018). Most researchers agree, though, that there are some muscle movements which are directly inherited, although they disagree over whether these are related to emotions or other affective states (Parkinson, 2005). Concerning the question of universality, the studies here are restricted to fairly homogeneous groups: the first, poker, takes place in the US. The second, dice rolling, takes place in the UK. Based on the literature,

it is reasonable to assume similar behaviour in culturally homogenous groups.

Separate from theories of emotion, facial expressions have also been viewed in the context of social signalling. Signalling is something done frequently throughout the animal world to convey messages between members of a species and also between different species. Examples of messages sent between members of a species are those conveyed during courtship rituals to advertise the fitness of a potential mate, warnings sent to the group about an outside threat of predators and messages indicating where food is to be found (Laidre and Johnstone, 2013). These messaging systems can be very complex and have been formed by the evolutionary process of adaptation. Deception is also a part of these messaging systems, as sending a false message can give the deceiver an advantage. This theory argues that important signals must, however, be fairly reliable, otherwise it would be detrimental for the recipient of the signals to pay heed to them and so they would simply be ignored. Different ideas have been put forward for how nature keeps signals reliable. Some signals, for instance, the male peacock's tail, require so much energy to make that they are themselves an honest indication of fitness and health that cannot be faked. Similarly, signals can be produced by specialized organs whose size or shape determine the important qualities of the signal. Signalling can be costly, so that they can be made by only those who can afford them (Searcy and Nowicki, 2005). Dishonesty can damage the signaller's reputation so that the deceptive party is penalized by having their signals ignored in the future. There can also be severe penalties for deception, such as being injured by other members of the group. These are some of the mechanisms that serve to keep signals honest, on average at least, and maintain an equilibrium of reliability. These also ensure that deception is difficult and costly (Laidre, 2009).

According to some, the signals humans send through various means, including that of facial expressions, are not products of conscious effort alone, but also happen automatically, or unconsciously. In *Honest Signals*, Sandy Pentland sets down a theory of two channels of communication that humans have: one is conscious and the other unconscious (Pentland, 2008). The two run parallel and largely independently of each other. We are rarely aware of the second channel but if we learn how to read it we can understand social situations better and use this understanding to our advantage. He terms the unconscious signals *honest* and defines them similarly to Zahavi, as signals that "are either so costly to make or so difficult to suppress that they are reliable in signalling

intention". Unconscious facial expressions belong to the honest signals. Humans are capable of suppressing or masking their facial expressions but only at the cost of great conscious effort (Pentland, 2008). This subconscious honest signal and the conscious effort to control it could provide the key to detecting deception. Pentland also proposes using technology to detect and study signalling to gain an understanding of people's true intentions in social interactions, and to use this understanding to gain an advantage.

Paul Ekman is one of the most steadfast and influential living proponents of Darwin's theory of facial expressions. His own theories of facial expressions and emotions developed directly out of those of Charles Darwin, who Ekman credited with being the founder of the field of psychology. As facial expressions are imbedded in our biology they are not really under our conscious control, without, as Pentland writes, making a great effort. Ekman and Friesen argue that we can learn to understand ourselves and others by understanding facial expressions (Ekman and Friesen, 2003). Even though we all interpret the facial expressions of others constantly, whether consciously or not, we do not really understand facial displays and the emotions they portray that well, even when it comes to understanding our own. This point is especially relevant to deception. Though a person might try and hide their true feelings by simulating an emotion or covering it up, Ekman and Friesen claim that careful analysis of how the muscles are used to make a faked facial expression can show that it is not genuine (Ekman and Friesen, 1974, 2003). In addition to these patterns, there are also facial expressions, called microexpressions, which are so brief that they are at "the threshold of recognition" (Friesen and Ekman, 1969). These occur when a genuinely felt emotion makes itself briefly visible before it can be suppressed and these, so they claim, can also be used to recognize deception.

If the appearance of the human face can reliably help us understand a person's behaviour as many claim, then using computer vision to automatically analyse and interpret facial expressions becomes feasible. This is currently a rapidly developing field with many possible applications in affective computing. For example, automatic facial expression analysis has been used to study depression (Scherer et al., 2014; Girard et al., 2013), automatically recognize attention deficit hyperactivity disorder and autism spectrum disorder (Jaiswal et al., 2017) and automatically measure pain (Egede et al., 2017). As described above, deception might also make itself visible and there have been a few works on automatically detecting this. These will be introduced in the next chapter.

Computer vision and machine learning have the potential to revolutionize the science of

detecting and interpreting facial expressions. Computers can potentially pick up facial movements that are too fast or too subtle for humans to notice. Also, they can process details and patterns that only become apparent over a long period of time, which humans cannot perceive so well. Computers can produce repeatable results and have the potential to remove bias and subjectivity from the process of interpreting facial expressions. They can process mass amounts of data, potentially very quickly, without suffering from fatigue, as humans do. Perhaps they can help distinguish between the honest and dishonest signals and help further study the scientific basis for facial expressions and emotions.

2.5 Three approaches to quantifying facial expressions and their use in computer vision

Before automatically measuring facial expressions, one first needs a descriptive model for quantifying them. There are three main models which stand out because they lead to a clear concept for quantifying facial expressions in ways that are useful for machine learning and computer vision. All three have been used as the underlying model of automatic detectors for a person's affective state. Two models, the *basic emotions* and the *circumplex model of core affect*, are message-based. This means that they interpret facial expressions according to the message they are meant to convey without being concerned with the physical means by which this is done. The basic emotions model categorizes facial expressions as 'happy', 'sad', and so on; the circumplex model categorizes facial expressions as coordinates in a two dimensional space of affect. The third model, the *facial action coding system* (FACS), is sign-based as opposed to message-based. Its intent is to provide an objective description of the state of a person's face in terms of which muscles are active, without assuming any underlying theory of emotions. Rather, FACS can be used as a tool for exploring theories of emotion and affect.

Machine learning techniques use specialized datasets for each of the three models to learn their detection algorithms. These datasets are annotated by humans in order to form a so-called *ground truth*. The way in which devices detect is heavily dependent upon the datasets that they learn from. This thesis focuses on the facial action coding system because of its descriptive power and because it allows one to explore unknown aspects of facial expression. However, as each of the three models can contribute to the

Figure 2.3: Ekman's six basic emotions as displayed by the author's children, Isabel and Bertrand, with their permission.

interpretation of facial expressions, I will now give a brief outline.

2.5.1 The basic emotions model

Paul Ekman developed a theory of basic emotions in the 1970s, which he based on the ideas of Darwin and Tomkins (Darwin, 2009; Tomkins, 2008). This theory consists of two main tenets. The first is that there exists a limited set of basic emotions which are distinct from one another in important ways. The second is that these emotions were developed through the process of evolution to enable us to deal with certain interpersonal tasks in life automatically and quickly, without having to make conscious decisions. Importantly, emotions entail communicative signals most strongly involving and conveyed through the face. Ekman also believed, like Darwin, that these signals were universal to humanity. Ekman has discovered at least six basic emotions, anger, fear, sadness, enjoyment, disgust and surprise, each of which has its own distinctive signal, or facial expression (Ekman, 1992), see Figure 2.3. While there might be more basic emotions, as yet there is not enough evidence for this.

While six basic emotions have been described, Ekman states that these are really six distinct families of emotions. Each of these has many variations, which have not been described thoroughly yet. The repertoire of facial expressions covers many more affective states than the six basic emotions. He also stated that, concerning the basic emotions, there might be some threshold at which the emotion must be present in order to elicit a display. Nonetheless, it is possible for an individual to consciously suppress an emotional display. Still, according to Ekman, the central nervous system should always exhibit a reliable pattern of activity for each individual basic emotion even if it is not able to af-

fect the face, and this pattern should be detectable with modern technology like magnetic resonance imaging (MRI) and electromyography (EMG). This was hypothesized but has not yet been demonstrated.

Automatic detectors for basic emotions are trained on images of faces that have been annotated by humans with the emotion they signal, similar to those shown in Figure 2.3. However, since the six basic emotions cover only a small subset of affective states, they are often not descriptive enough for real-life applications.

2.5.2 The circumplex model

The circumplex model had its origins in early studies on how accurately humans can judge what emotions a face is expressing (Woodworth, 1938). Woodworth had observers look at images of faces and then assign each image to a word that in their opinion best described the emotion the face in the image was displaying. For instance, an observer might assign an image of a smiling face to the word 'happy'. The results of this experiment suggested that human's are not good at assigning the correct emotion to a given photograph of a facial expression if one evaluates their performance only on the basis that the assignment is 'right' or 'wrong'. Woodworth discovered, however, that if he ordered the words in the following linear fashion - (1) Love, Hapiness, Mirth; (2) Suprise; (3) Fear; (4) Anger, Determination; (5) Disgust; (6) Contempt - the performance markedly improved. It was found that when a person assigned a photo to the wrong emotion-word they tended to assign it to an adjacent word, and hence got close to the correct answer. He concluded that humans are indeed capable of correctly judging human emotions from facial expressions given this linear arrangement.

Building on this linear set of words, which can be thought of as bins one to six, Schlosberg (Schlosberg, 1941) tested Woodworth's method on a different set of photos, the Frois-Wittmann pictures of facial expressions (Hulin and Katz, 1935). Schlosberg noticed that people often assigned photos that belonged to bin 6 (contempt) to bin 1 (love, happiness). An example of this is the image number 52 shown on the left of Figure 2.4. Here, the facial expression for love is being displayed. However, Woodworth's study subjects occasionally classified this as contempt. This observation and other experiments led to an important conclusion, namely that the scale was not linear but circular. It followed that it should be possible to describe all facial expressions more accurately with

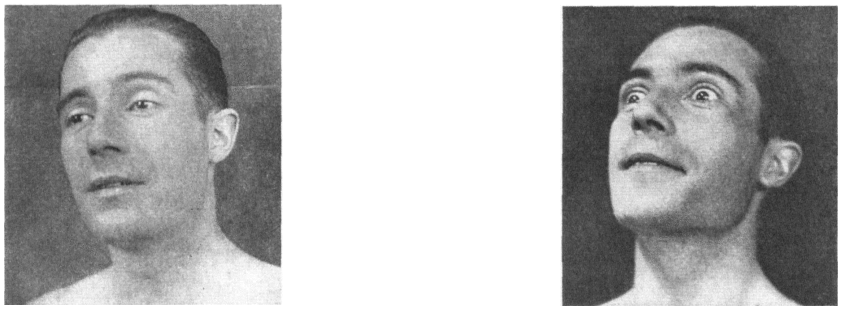

Figure 2.4: Frois-Wittmann images 52 (left) and 10 (right).

two dimensions. Schlosberg created the axes of these two dimensions, namely Pleasantness – Unpleasantness and Attention – Rejection, which were each split into nine numeric values as shown in Figure 2.5 (Schlosberg, 1952). Observers were asked to look at each picture in the Frois-Wittmann picture set and assign it a value of 1 to 9 on the Pleasantness – Unpleasantness scale and then the Attention – Rejection scale. Afterwards, he plotted each image according to the values it had been assigned onto the two dimensional graph. As an example, I show image 10 of the Frois-Wittmann set, see Figure 2.4 right side, plotted in the two dimensions in Figure 2.5, part A. The origin is located at (5,5) and image 10, which displays "Pleased Surprise", is located at (7,7). He then placed the linear Woodworth set of words around the circumference of the graph and projected each word onto the Woodworth scale by means of a ray centred at the origin. He discovered that this method usually gave the correct Woodworth value. In the case of image 10, this was 1.75, as shown in Figure 2.5 A. Using this method of projection, the two dimensional model was able to identify the correct Woodworth bin for the images in most cases. Schlosberg concluded that two dimensions were as descriptive as the linear arrangement of Woodworth's six categories and he proposed that all facial expressions could be most accurately described by two dimensions alone (Schlosberg, 1952).

Schlosberg's concept of a continuous spatial model for representing affective states was further developed by James Russell, who also coined the term *circumplex model* to describe it in 1980 (Russell, 1980). It is thus different from the basic emotions model as it is continuous and not discrete. In the circumplex model, Russell replaced the concept of emotions with the concept of *core affect* (Russell, 2003). The model describes some affective states that are not emotions, such as sleepiness. Conversely, there are some emotions such as fear, jealousy, anger and shame which are not distinct points in the continuous space of core affect of the circumplex model, see Figure 2.5 B.

The circumplex model is appealing in its simplicity and the ease with which it can be

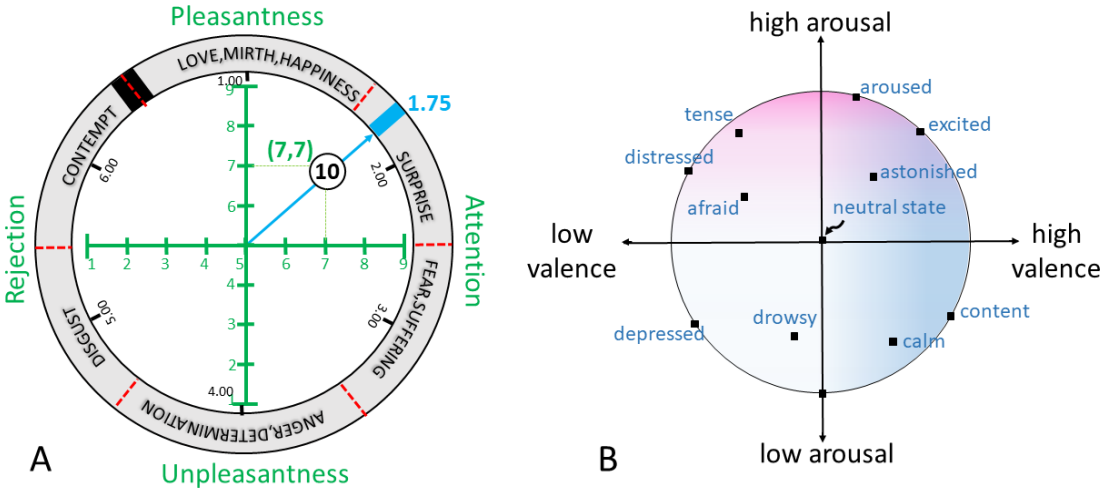

Figure 2.5: Circumplex diagrams with affective states according to A, Schlosberg (1952) and B, Russell (2003). In A, the encircled 10 indicates the position of Frois-Wittmann image 10, which is shown in Figure 2.4, right.

represented. It, unlike discrete classes of emotions, can represent continuous and subtle mental states, which encompass more affective states and facial expressions than the discrete emotions described by Ekman. However, it also has descriptive limitations and cannot, for instance, represent the two emotions 'fear' and 'anger' as being separate. The two dimensional model can be extended to arbitrarily many more dimensions in order to distinguish between more closely related emotions and other affective states. Fontaine and colleagues proposed a four dimensional model but suggested that researchers should design their dimensional model according to what they are researching as different models serve different purposes (Fontaine et al., 2007). The circumplex model is also challenging for observers to comprehend. Moreover, given two professional annotators, both of whom have been trained to judge circumplex values for images of facial expressions, the values they assign to the same images often don't agree, making the inter-rater reliability of the circumplex model lower. For this reason, it can be difficult to establish a reliable ground truth for images, whether they be posed in a laboratory or spontaneous in the wild. This, in turn can be problematic when designing datasets for machine learning algorithms to train on (Gunes and Pantic, 2010a).

2.5.3 Facial action coding system - the FACS model

The third model of facial expressions is the facial action coding system (FACS), as presented in the FACS Manual (Ekman et al., 2002). The purpose of the FACS model is

Action Unit	Description	Corresponding muscle group
AU1	Inner brow raiser	frontalis (middle part)
AU2	Outer brow raiser	frontalis (outer part)
AU4	Brow lowerer	procerus, depressor and corrugator supercilii
AU5	Upper lid raiser	levator palpebrae superioris, superior tarsal
AU6	Cheek raiser	orbicularis oculi
AU9	Nose wrinkler	levator labii superioris
AU12	Lip corner puller	zygomaticus major
AU15	Lip corner depressor	depressor anguli oris
AU20	Lip stretch	risorius, platysma
AU25	Lips part	depressor labii, mentalis, orbicularis oris
AU26	Jaw drop	masseter
AU45	Blink	levator palpebrae, orbicularis oculi

Table 2.1: The twelve Action Units used in this thesis and their associated muscle groups.

to create a complete descriptive language for the visual appearance of the face in order to provide a tool for studying facial expressions for many different disciplines. These descriptors should ideally be well-defined, be built of the most basic units possible and be capable of describing all possible facial expressions. They also should be repeatable, objective and easy to understand, and rely only on external appearance to make them accessible to a human or machine observer.

This 'objective' approach should describe facial expressions as a list of facial muscle contractions along with their magnitudes without recourse to interpreting the emotions being conveyed. This was the approach first systematically studied by Darwin and Duchenne and the facial action coding system can be traced back to them. In FACS, the most basic visual descriptor for facial expressions is the *Action Unit* (Hjortsjö, 1970; Ekman and Friesen, 1978). There are currently about 100 of these and they correspond roughly to the different facial muscles, see Figure 2.6, as well as some non-facial descriptors for things like head pose and gaze direction. For instance, AU1, also known as inner brow raiser, corresponds to the frontalis muscle. Some action units correspond to groups of muscles, like AU4, brow lowerer, which corresponds to three face muscles, namely procerus, depressor and corrugator supercilii. See Table 2.1 for descriptions of the twelve action units used in this thesis. The intensities of each of these action units was defined in FACS in the order from A - trace, B - slight, C - marked, D - severe to E - maximum.

To make the results easy to understand and repeatable, the FACS manual gives detailed

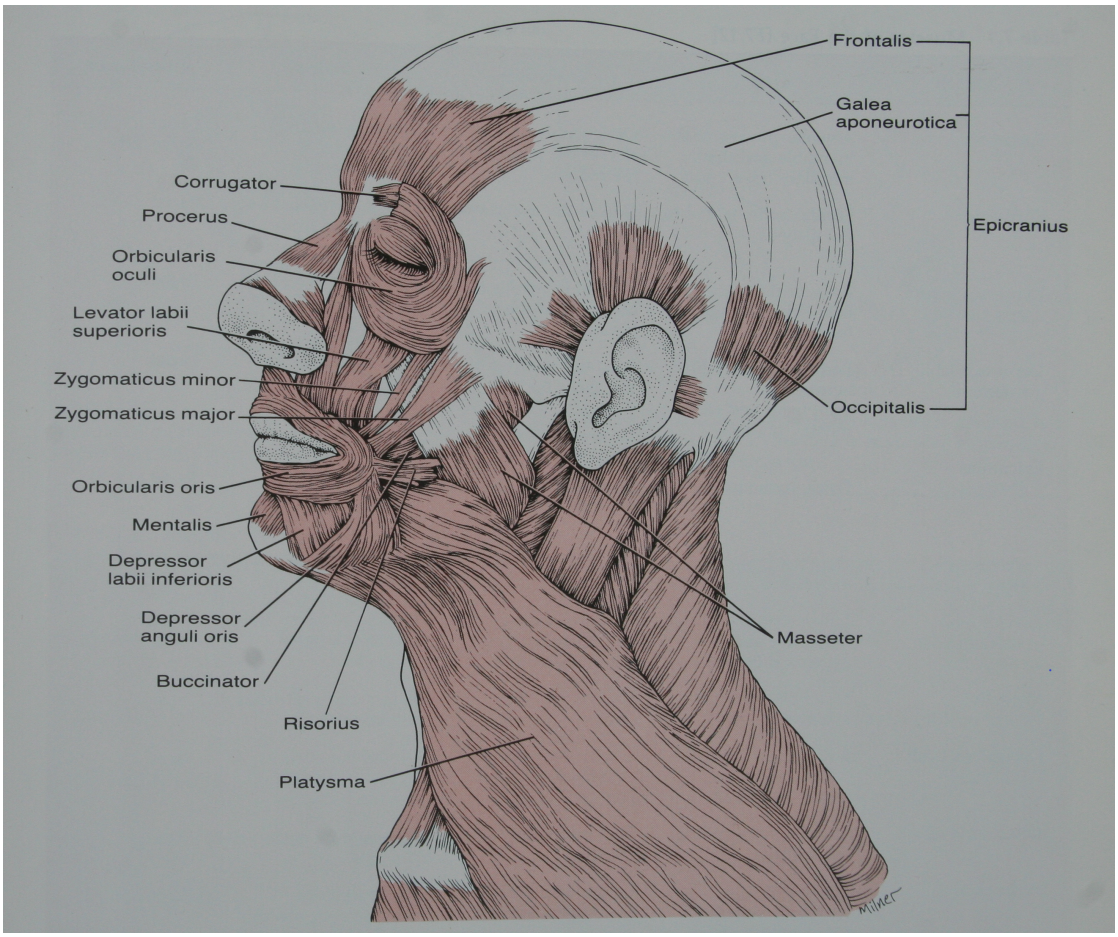

Figure 2.6: Anatomy of the muscles of the face. AU1 (inner brow raiser) corresponds to frontalis. AU4 (brow lowerer) corresponds to the procerus, depressor (not shown in image) and corrugator supercilii muscles. Taken from Spence (1990), page 193.

descriptions of how to evaluate the presence and intensity of each action unit (Ekman et al., 2002). Visible traits such as changes of the positions of facial features, changes of shape, the appearance or disappearance of bulges, furrows or wrinkles are used to evaluate when an AU is activated and what its intensity is. The FACS Manual is used to train FACS annotators, who annotate images of facial expressions with the action units displayed in them. In addition, FACS describes the consecutive phases of activation of Action Units in an intuitive way that reflects muscle functioning. For instance, in the case of AU1, inner brow raiser, an occurrence consists of a neutral phase before the corresponding muscle starts to contract, followed by the onset stage, when the muscle becomes increasingly contracted, followed by the apex stage, where it reaches its maximum contraction as indicated by the eyebrow being maximally raised, followed by the muscle gradually relaxing in the offset stage and then returning to the neutral stage, see Figure 2.7.

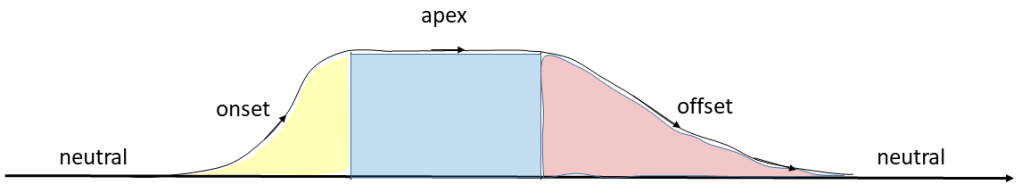

Figure 2.7: Model of (facial) muscle activation as sequence of neutral, onset, apex and offset stages.

Since FACS is theoretically capable of describing all facial configurations, including but not limited to those related to emotions, it should be able to cover both the prototypical facial expressions, see Figure 2.3, and the circumplex model of core affect. The prototypical facial expressions and some variations on these have been translated into action units, for example, in Du et al. (2014). 'Happiness' is described there by the following FACS formula: AU12 + AU25 + [possibly AU6] and 'anger' by: AU4 + AU7 + AU24 + [possibly AU10, AU17 or AU23]. The circumplex model, in contrast, has not been thoroughly translated into FACS formulas. Since there are currently nearly one hundred action units there are a huge number of combinations, each of which represents a different facial expression. Of these, over 4,000 configurations have been reported in the literature (Bartlett et al., 2014). Thus, FACS provides a highly effective tool for studying the human face based on visual appearance, and it is fairly complete in its ability to describe facial expressions. Datasets made for developing automatic detectors consist of images that have been annotated for their action units. These annotations follow the descriptions given in the FACS manual.

2.6 The two action unit detectors used in this research

There are numerous different action unit detector systems. In this section, I describe the two action unit detectors that were used in this thesis. One was recently developed here at the University of Nottingham, Convolutional and Bi-directional Long Short-Term Memory Neural Networks (CNN-BLSTM). The other, OpenFace, is available and widely used in the research community. I shall begin with a discussion of the three databases

they were trained on.

2.6.1 The databases from which they are constructed

FACS's comprehensiveness and descriptiveness makes it suited for computer vision techniques. There is a large and ongoing effort to create detectors that automatically annotate images of humans with their correct AU labels. Studying and detecting human behaviour such as deception requires annotating large amounts of images and it takes a human hours just to annotate a few minutes of video. Methods for automatically annotating video with AU labels are now being developed by several groups. In this thesis, two such action unit detectors are used and their performances compared. One, CNN-BLSTM, was developed in the School of Computer Science at the University of Nottingham called (Jaiswal and Valstar, 2016). The other one is a publicly available, open source tool, called OpenFace (Baltrušaitis et al., 2016), which has been used for several published studies. CNN-BLSTM attained the highest score on the FERA 2015 challenge (Valstar et al., 2015), while OpenFace was never tested on this exact benchmark as it was no longer available, making a direct comparison difficult. In addition, OpenFace was tested on the DISFA dataset, where CNN-BLSTM was tested on BP4D and they also used different metrics. A comparison of the results is shown in Table 2.2.

Both action unit detectors used here require no specialist equipment, only videos produced by webcams or similar devices. The two detectors were trained largely on the same datasets, but were otherwise developed independently of one another using different techniques. These detectors were chosen because of their high quality and because they were designed to capture spontaneous behaviour. This provides an interesting comparison as the deceptive behaviour studied here is spontaneous as opposed to posed and thus potentially difficult to detect. There are also commercial AU detection systems, such as CERT and Facet. These have been excluded because they do not provide publicly available details about how they were constructed and they are frequently not benchmarked (Baltrušaitis et al., 2016).

There are also numerous databases for training AU detectors, most consisting of posed behaviours. In contrast, the three databases used to train the CNN-BLSTM and OpenFace detectors contain more spontaneous behaviours and consist of videos of people responding to emotion-eliciting tasks. This is important, as spontaneous behaviour dif-

Action Unit	CNN-BLSTM	OpenFace
AU1	0.64	0.28
AU2	0.50	0.28
AU4	0.70	0.34
AU5	0.67	not given
AU6	0.59	0.70
AU9	0.54	not given
AU12	0.85	0.78
AU15	0.39	0.20
AU20	0.22	not given
AU25	0.85	not given
AU26	0.67	not given
AU45	not given	not given

Table 2.2: A comparison of CNN-BLSTM and OpenFace. Note that OpenFace was tested on DISFA while CNN-BLSTM was tested on BP4D. The two have never had their performance measures compared to each other in a uniform way. They also used different performance measures: OpenFace used Pearson Correlation Coefficient and CNN-BLSTM used F1 scores. Thus it is difficult to foresee how they might perform on unseen data. Also, the version of CNN-BLSTM I use has developed since its publication in Jaiswal and Valstar (2016).

fers from posed behaviour in important ways and is more subtle and harder to detect. These databases, called SEMAINE, BP4D and DISFA, were designed for researchers to use for training automatic action unit detection systems. They each consist of frontal videos of subjects' faces that have been annotated frame by frame by human FACS annotators. Each database, also known as a corpus, contains sequences of images along with their corresponding annotations. Some of these images have been annotated separately by more than one annotator to compare the accuracy of the annotations, usually summarized by a value known as the inter-rater reliability (Cohen, 1988). SEMAINE is annotated for six AUs, BP4D for 27 AUs and DISFA for 12 AUs, which partially overlap between the databases. These annotations serve as the ground truth for the systems that learn from them. In the following, I will briefly describe how these databases were designed, as this understanding also sheds light on how the detectors that have been trained on them work.

- The SEMAINE corpus (McKeown et al., 2010, 2012) consists of high-quality videos together with audio of 20 participants each interacting separately with an operator who is imitating four different stereotypical characters to produce four different conversational interactions. The participant and operator communicate

through cameras and screens. Videos are taken from a frontal view and, using partially reflective mirrors, allow direct eye contact between the participant and the operator. This corpus was designed to be used as a tool for studying human-computer interactions and for advancing natural language processing. It focuses on language and non-verbal social signals that accompany conversation. Although the conversational pair is human-human, the operator tried to behave in a machine-like manner in order to simulate computer-human interactions. The audio and video were annotated by up to four separate annotators for the affective dimensions valence, activation, power, anticipation and overall emotional intensity, similar to the circumplex model, for the six basic emotions anger, disgust, amusement, happiness, sadness and contempt, see Figure 2.3, and also for the occurrence (present/not present) of six action units: AU2, AU12, AU17, AU25, AU28 and AU45. For the second Facial Expression Recognition and Analysis challenge (FERA 2015), 130,695 frames of the SEMAINE database were used for training detectors (Valstar et al., 2015).

- The BP4D corpus (Zhang et al., 2014) was designed to provide spontaneous FACS-annotated 3D images of facial expressions. The corpus contains 41 subjects performing eight different emotion-eliciting tasks. Some of the tasks were conversational and involve communicating with a professional actor, others were not. The role of the actor was to guide the subject through the tasks and make the interaction seem as natural as possible. The eight tasks were designed to evoke happiness, sadness, surprise, embarrassment, nervousness, pain, anger and disgust, which roughly correspond to the six basic emotions. After the sessions had been run, a continuous 20 second segment was selected for each of the eight tasks and each subject. This segment was chosen by FACS annotators to be the one that had the highest level of expression. Therefore, this dataset focused on high expressivity as opposed to more subtle expressions. These 20 second segments were then annotated by two professional FACS annotators for 27 different AUs: 1, 2, 4 - 7, 9 - 20, 22 - 24, 27, 28, 30, 32, 38 and 39. Their results were then compared for accuracy. To ensure the correct emotions had been elicited, the subjects self-reported on their own feelings, and images were shown to uninvolved observers to get their impression of the emotion being shown on the subjects' faces. Finally, professional annotators checked to see if the AU annotations made for each segment corresponded to those associated with the targeted emotions. BP4D consists

of 368,036 annotated images.

- The DISFA corpus (Mavadati et al., 2013) was especially designed to capture spontaneous behaviour. Unlike SEMAINE and BP4D, it does not involve conversation. It contains videos of 27 adults viewing emotion-eliciting videos. The video sessions were made on an individual basis; a subject sat alone in a room and watched a video on a screen that had a stereo camera mounted on top which made a recording of their face. A certified human FACS annotator then annotated the images with 12 separate action units: AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU17, AU20, AU25 and AU26. These were annotated for occurrence (present/not present) as well as for intensity on a scale of A-E as mentioned in Section 2.5.3. These particular AUs were chosen by Mavadati et al. as they were deemed to be the most commonly occurring. Altogether, this dataset consists of 230,000 annotated images.

In this study, twelve action units are used: AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU20, AU25, AU26 and AU45, see Table 2.1. These were chosen because they are very relevant to deciphering frequently occurring facial expressions, they show a high degree of overlap with the action units used in these three databases and they are detected by both detectors studied in this thesis.

2.6.2 A comparison of different machine learning approaches for building the two AU detectors

This thesis investigates deception using two different ready-made action unit detectors. They are really each a set of detectors, one detector for each action unit. Here, they are collectively referred to as CNN-BLSTM detectors and OpenFace detectors, respectively.

The CNN-BLSTM detectors were designed for detecting AUs in spontaneous facial expressions in uncontrolled circumstances. They are made out of a combination of convolutional neural networks (CNNs) and bi-directional long short-term memory neural networks (BLSTM) (Jaiswal and Valstar, 2016). To carry out the difficult task of detecting spontaneous behaviour, the authors focused on geometrical facial features such as the shape of the mouth, appearance features such as wrinkles, bulges, furrows and the temporal dynamics of these features. For their machine learning model they used

convolutional neural networks (CNNs) to learn action units and they avoided intermediate handcrafted features as much as possible to allow these to be learned by the neural network. In a preprocessing step, they first automatically located facial points in each image and used these points to segment the face into rectangular regions, such as eyes and mouth regions (Sánchez-Lozano et al., 2018). To capture geometrical features in each of the regions, they used the facial points to create a binary mask of the target feature. This rendered the features into a simple geometrical shape without texture. To capture appearance, they simply used the same rectangular regions as is. They reasoned that dynamics are also important to determining action units, so the input into their neural network was the sequence of frames directly surrounding the frame that was to be automatically annotated – the two images directly preceding the current image and the two images directly following the current image. In addition to these dynamical features from a small window around the current frame to be annotated, they also used a BLSTM to capture dynamics spanning a longer time frame. A CNN-BLSTM was learned for each action unit using only those rectangular regions associated with the action unit. This was determined by the authors' expert opinion. These detectors were trained on the three datasets described in the previous section.

The open source toolkit OpenFace (Baltrušaitis et al., 2018) was also used in this thesis to gain a further understanding of action unit detectors. OpenFace was designed to provide researchers with a fast, easy to use and state of the art tool for studying facial expressions. It has been used several times in behavioural studies and was designed differently to Jaiswal and Valstar's CNN-BLSTM detectors and thus provides an important comparison. Like the CNN-BLSTM, it takes into account geometric features, appearance features and dynamic features of facial expressions, but in a much more different way and using a different machine learning model, support vector machines (SVMs) (Baltrušaitis et al., 2015). The decision to use SVMs was taken partly to increase speed allowing annotations to occur in real-time. This attention to speed is one characteristic that distinguishes OpenFace from the CNN-BLSTM detectors. OpenFace is about thirty times faster and uses less memory than CNN-BLSTM, possibly at the expense of accuracy, see Chapter 4 for details. Like CNN-BLSTM, in a preliminary step, facial points were first detected and the face aligned. However, facial points were not used to segment the face into geometrical regions, but they were kept intact along with information about their non-rigid motion. For appearance features, the aligned image was segmented into a regular grid and then histograms of oriented gradients (HOGs) were made for each

region and these added to the vector. To reduce the number of dimensions, principal component analysis (PCA) was used on these features.

To capture dynamic information and remove individual bias in the training samples, the authors of OpenFace wanted to create a reference to the neutral face (Baltrušaitis et al., 2015). Their assumption of a neutral face was based on the premise that in real life people show the neutral face the majority of the time. Accordingly, the authors calculated the statistical medians of all values in the feature vectors that were input to the SVM for a given individual, and assumed that this represents the neutral face for that person. For each individual, they then subtracted the median value from each feature to obtain a new dynamic feature vector normalized around this estimate of the neutral face. Subsequently, the authors applied the assumption that the neutral face is the most frequently occurring face a second time, namely to the final output of the AU detectors. They did this by subtracting the value of the nth percentile of the value of the action unit from all its values for that particular individual. Thus, normalization was done twice for each individual - once on the feature vector that was input to the SVM classifier and once on the output of the detector. The authors note that for some action units this dynamic representation works better while for others the static, unnormalized vector works better. They also note that the dynamic model might not work well on low-level expressions or in situations where the assumption does not hold that the most frequent expression displayed is the neutral face. Like the CNN-BLSTM detectors, OpenFace was trained on the three datasets described in the previous Section 2.6.1, plus some additional ones.

Chapter 3

Related Work - Automatic Detection of Deceit Using Facial Action Units

There is a large and rising demand for methods of automatically detecting deceit using facial action unit detectors, yet, there are relatively few studies so far in this emerging field. Here, I briefly present the most important and fundamental works in this area of research. In addition to exploring ways to automatically detect deceit, each of these studies had to first grapple with the issue of eliciting deceptive behaviour in their study subjects in such a way as to capture it on video for later experimentation and analysis.

3.1 Detecting deceit from the face

The first study presented, *Who Can Catch a Liar* (Ekman and O'Sullivan, 1991), did not employ automatic action unit detectors or any other form of automatic detection of facial expressions, as these had not yet been developed. Instead human FACS annotators were used. These human annotators played the same role as the automatic annotators would have. Otherwise, the study serves as a paradigm in many ways for later investigations of human deception using automation. Not least because the ideas of Paul Ekman and the research associated with him permeate most later works on automatic detection of deception.

The objective of Paul Ekman and Maureen O'Sullivan's 1991 groundbreaking study on detecting deceit from facial expressions was to discover how well human observers could detect deception from facial and postural clues, and to find out which clues people used in their attempt to uncover deceit. To do this, they needed a set of videos of people displaying both deceitful and honest behaviour that they could then show to their human observers to find out how well they performed at distinguishing between the two behaviours. In pursuing this aim, they were confronted with the same fundamental questions that would arise when trying to automatically detect deceit: How does one evoke deceitful behaviour in humans and what kind of deceitful behaviour should it be? How does one capture facial and body expressions of the study subjects? How can one be sure of the ground truth?

Ekman and Sullivan used a dataset that had already been created for an earlier study by Paul Ekman and Wallace Friesen (Ekman and Friesen, 1974). This dataset had been designed to capture deceit where the deceiver was intentionally deceiving their counterpart and where the deciever had a lot to lose if they were caught. The counterpart was also intentionally trying to uncover acts of deception. To achieve this, Ekman and Friesen created a scenario whereby nursing students were told they would be given an interview during which they would watch four videos and then describe them to the interviewer. Two of these videos were pleasant nature films and two were gruesome videos of burn victims and amputations. The nursing students were told they were to lie about one of the gruesome films and try and convince the interviewer, who had no knowledge of what they were watching, that they were really watching a pleasant film. To make the deception high stakes and motivate the nursing students to try and succeed in their deceit, they were told that their success in the nursing profession depended on their ability to successfully suppress expressions deemed negative such as those indicating anxiety and disgust. To create further antagonism between the student and the interviewer, the interviewer was told to actively try and discover when the nursing students were not being truthful. In order to capture the behaviour so that it could be used for their experiment, Ekman and Friesen secretly filmed the faces and bodies of the nursing students during these interviews and obtained permission to use the video recordings after the experiment. Two synchronized cameras were used, one to capture the face and one to capture the whole body of the subjects. Since the order of the films was known the ground truth was also known.

For their experiment, Ekman and O'Sullivan used ten short video segments from the videos of 31 nursing students, five of them representing honest behaviour and five of them dishonest behaviour, all of them containing both the face, body and audio. These ten segments were selected by the authors because they contained physical signs of deceit or honesty as measured by the FACS coding system. Ekman and O'Sullivan reasoned that they could not penalize lie catchers for not being able to recognize deceit, if deceitful and truthful behaviour was visually indistinguishable. Thus, the ten test videos were chosen to contain measurable "differences between honest and deceptive samples" so that their study could "focus on the question of how well observers can detect deception".

After their experiments were run, Ekman and O'Sullivan reached the conclusion that none of the professionals tested, except for specially trained Secret Service agents, could detect lies significantly better than chance. The Secret Service agents chose on average 64% correct when distinguishing between the two behaviours.

Ekman and O'Sullivan commented on weaknesses in their study. It is possible that the type of deceit in this study, the concealment of strong negative emotions, is not in itself very relevant and that it does not generalize to other deceit scenarios. The ground truth is also not so certain; the nursing students had to persuade the interviewer they were watching something pleasant even when they really were, therefore they might have even been deceitful for the pleasant films. For this reason, Ekman and O'Sullivan referred to the two behaviours as 'deceitful' and 'less deceitful'. Also, the sample of behaviours of size ten was perhaps also too small.

The rest of this chapter explores past efforts to see how well computers can detect deceptive behaviour, sometimes directly comparing their abilities with those of humans. As will be seen, the approaches they used and problems they encountered were similar to those of Ekman and O'Sullivan (1991).

3.2 Posed versus spontaneous smiles: The first work to automatically distinguish posed from spontaneous behaviour

Cohn and Schmidt (2004) were the first to automatically distinguish posed from spon-

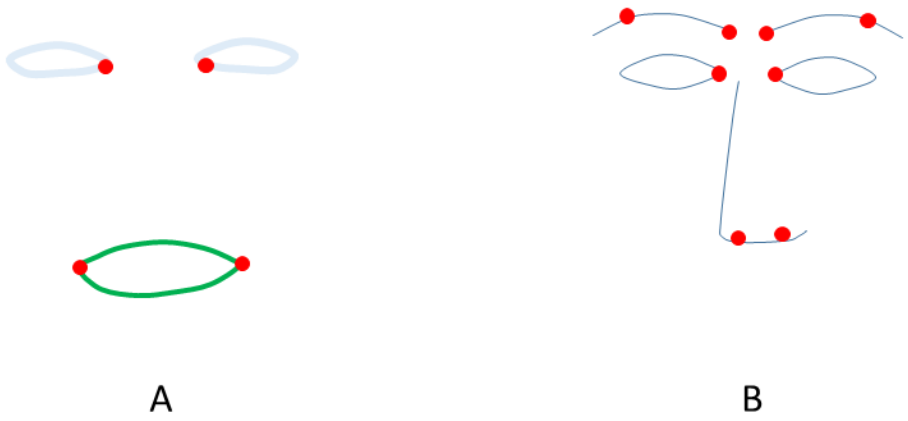

Figure 3.1: A, tracking of lips and four fiducial points in Cohn and Schmidt (2004). B, tracking of eight fiducial points in Valstar et al. (2006).

taneous behaviour. As such their work can be considered to be the first work to automatically distinguish deceptive from honest behaviour. The authors investigated whether dynamics and morphology of facial features together could be used to distinguish posed and genuine smiles. For examples of faked smiles, they used videos of 33 subjects who had been instructed to smile. For examples of genuine and spontaneous smiles, they used videos of 48 subjects watching comedy films. To capture shape and dynamics of smiles, they tracked the lips, the left and right mouth corners, and the inner eye corners, see Figure 3.1 part A. The authors found that in genuine smiles there is a consistent and deterministic relationship between the amplitude of the smile and its duration whereas in posed smiles this relationship is arbitrary. This characteristic, which captures dynamic features of the smile, could be used to detect posed versus spontaneous smiles with 93% accuracy. This work did not use action unit detectors, but the authors confirmed that the facial tracking of features was concurrent with the activities of the underlying muscles that correspond to AU6 (orbicularis oculi), AU12 (zygomaticus major), AU 15 (depressor anguli oris) and AU 17 (mentalis).

Spontaneous smiles	Posed smiles
slow onset	abrupt onset
multiple AUs involved	involves primarily only AU12
multiple apexes	a single apex
symmetrical	asymmetrical

Table 3.1: Rules for distinguishing spontaneous from posed smiles according to Valstar et al. (2006).

3.3 Automatically distinguishing genuine from posed behaviour with eyebrow dynamics

In 2006, Valstar and colleagues presented one of the first studies of automatic detection of deceit (Valstar et al., 2006). They investigated whether computers could automatically detect the difference between posed and genuine, spontaneous expressions by means of analysing the eyebrows alone. This was also an investigation into whether brow actions differ between posed and spontaneous behaviour with special emphasis on the role of facial dynamics in conveying meaning (Ambadar et al., 2005; Bassili, 1978). This approach was similar to that pursued with regard to smile dynamics (Cohn and Schmidt, 2004). This had, however, been done by humans, not automatically by computers. Valstar et al. postulated that rules similar to those governing smiles might also govern other facial actions such as brow actions. Spontaneous smiles tend to be slow to start (onset), have multiple peaks (apexes), usually involve multiple AUs, not just AU12 (lip corner puller), and are usually more symmetrical than posed smiles (Ekman and Friesen, 2003), see Table 3.1. Therefore, the authors investigated if brow dynamics could be similarly used to automatically discriminate between posed and spontaneous behaviour.

Their work centred around the concept of automatically detecting action units of the FACS system, as opposed to most studies in automatic detection of facial expressions up until then that sought to recognize prototypical expressions like the six basic emotions. Valstar et al. considered only the movements of the eyebrows, which can be described by the facial action units AU1 (inner brow lift), AU2 (outer brow lift) and AU4 (brow lowerer). To investigate whether the characterisation for posed versus spontaneous smiles could be applied to brow motion, they were particularly interested in the temporal dynamics of these AUs - the amount and speed at which positions of eyebrows changed over time, the order in which the different AUs occurred and how symmetrical they were.

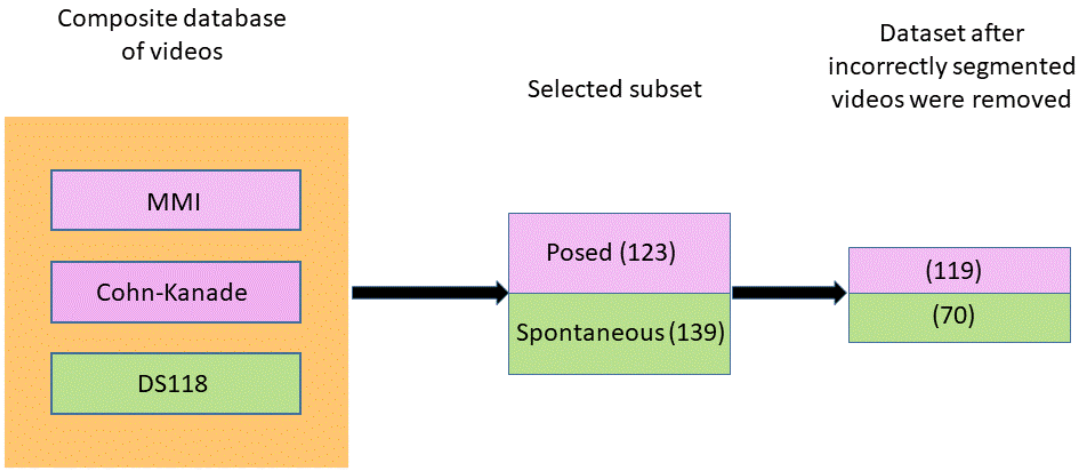

Figure 3.2: Process of selecting a database containing videos with posed (red) and spontaneous (green) behaviours.

To carry out their study, Valstar et al. needed a dataset of videos exhibiting the relevant behaviour. They composed a database out of three pre-existing databases: the MMI facial expression database, which contains over 4000 videos of 52 adults performing the six basic expressions on command (Pantic et al., 2005); the Cohn-Kanade facial expression database containing over 2000 videos of 210 adults, also producing the six basic expressions on command (Kanade et al., 2000); and the DS118 database consisting of videos of interviews of 85 people with heart disease (Rosenberg et al., 1998). To provide examples of posed behaviour, they extracted 60 sample videos from the MMI and 63 videos from the Cohn-Kanade facial expression databases, respectively. To provide examples of spontaneous behaviour, they extracted 139 samples from the DS118 dataset. See Figure 3.2 for a schematic representation.

As the basis of their automatic analysis, they computed eight fiducial facial points for each frame in the videos. These points were two points on each brow, and for reference, one point on the inner corner of each eye and one point on the outside of each nostril, see Figure 3.1 part B. These eight points were tracked through the entirety of each video sample to produce a time sequence of facial point positions. These points were the source of all information that would be used from the videos in the process of creating a classifier to distinguish posed from spontaneous behaviour. From this initial sequence of facial point positions, a set of basic features was calculated. These basic features tracked how each point was displaced from its original position and how distances between pairs of facial points changed over time. These features were then used to learn three separate detectors, one for each of AU1, AU2 and AU4.

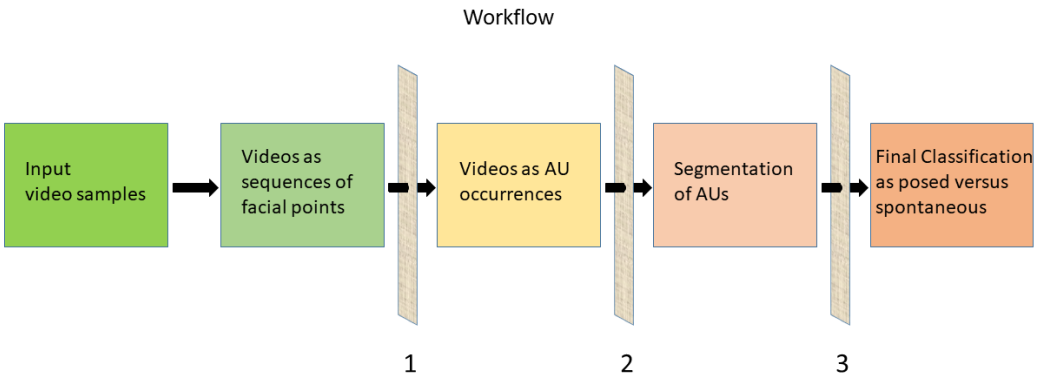

Figure 3.3: Multilayer process of building a classifier for distinguishing posed from spontaneous expressions. The three beige panes represent separate applications of boosting.

The basic features were then used in parallel with the AU detectors that had been made out of them to segment the occurrences of the three AUs into their onset, apex and offset phases, see Figure 2.7. After this step, any incorrectly segmented videos were removed. This left 119 videos of posed behaviour and 70 of spontaneous behaviour, see Figure 3.2. Mid-level features were then computed for these segments. These consisted of features that captured information about the speed and magnitude of eyebrow actions, their symmetry, and the order in which the AUs occurred. An important part of their methodology was the application of boosting to their features, both basic and mid-level, in order to determine which of the features were most relevant to the problem at hand. Boosting in combination with SVMs had previously been shown to be an effective combination for detecting facial expressions due to their speed and accuracy (Valstar and Pantic, 2006; Bartlett et al., 2004; Schapire, 1999; Vapnik, 1995). See Figure 3.3 for the workflow of their algorithm. The accuracy of the final classifier which classified behaviour as posed or spontaneous was over 90 percent. The authors refer to this as semi-automatic classification of posed versus spontaneous behaviour. This is due to their occasionally having to manually correct certain parts of the automatic detection. They emphasize that the purpose of the study was to determine if it is possible and feasible to automatically differentiate posed from spontaneous behaviour based on brow dynamics, not to present a classifier with a specific accuracy. They concluded that it is possible to distinguish posed from spontaneous behaviour based on brow motion alone. Their study confirmed that the speed, magnitude and duration of eyebrow actions are important in distinguishing posed from spontaneous behaviour. However, the authors did not find evidence that symmetry of eyebrow actions was useful for distinguishing posed from genuine behaviour.

3.4 Multimodal detection of posed versus spontaneous smiles

After having established that it is feasible to automatically detect deception by means of brow dynamics, Valstar's group followed up with a study on automatically discriminating posed from spontaneous behaviour based on geometric features of smiles (Valstar et al., 2007). This study used many of the same techniques as the brow study presented in the previous section, but it was more exploratory and broader reaching. As in the previous study, they again investigated the role of dynamics in distinguishing posed from spontaneous behaviour, however, they also investigated two other aspects of the problem: First, instead of using the single modality of the face they combined it with two additional modalities — head pose and shoulder dynamics — to investigate which of these modalities was most useful and whether these were more effective in combination or alone. Second, they compared three different ways of fusing features together before final classification. They compared the effectiveness of three fusion strategies: early, mid-level and late fusion. Their earlier brow study, described above, had used late fusion.

The basis of their study was a subset of the MMI-facial expression dataset, see Section 3.3, which was one of the three datasets used in the previous 2006 study. For the posed expressions, they used 100 near frontal view videos of individuals acting out a sequence of emotional expressions. For the spontaneous expressions, they used 102 near frontal view videos of individuals watching emotion eliciting cartoons or nauseating videos. In the posed samples, only that section where the actor demonstrated the smile was selected. In the spontaneous samples, only a single smile was selected. In both cases, the videos were trimmed to only contain the duration of the smile, from the smile's beginning to its end. For the first modality, head pose, they tracked over all video frames rigid head movement – horizontal, vertical and forward/backward – and three degrees of rotational movement – yaw, pitch and roll, see Figure 3.4, part A. For the second modality, the face, they tracked 12 facial points associated the action units AU6 (cheek raiser), AU12 (lip corner puller) and AU13 (cheek puffer), which are used to distinguish genuine smiles from posed smiles – four points for each eye (left and right corners, top of upper lid and bottom of lower lid), and four points for the mouth (left and right corner and top and bottom of upper and lower lip respectively). For the third modality, the shoulders, they

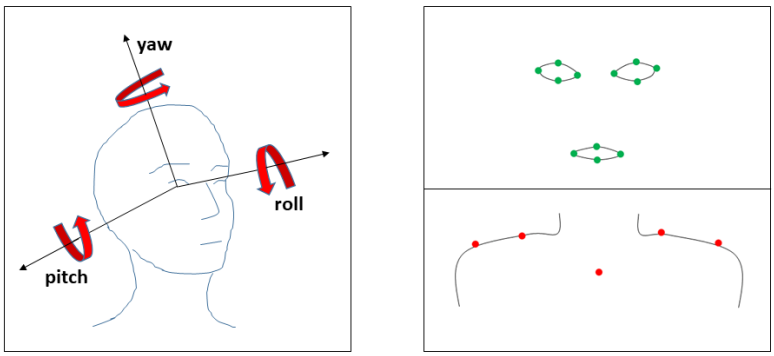

Figure 3.4: A, head pose yaw, pitch and roll. B, seventeen fiducial points tracked for the two modalities: face, with twelve facial points in green, and shoulders, with five shoulder points in red.

tracked five points – two on the left shoulder, two on the right and one in the middle to account for rigid motion of the torso, see Figure 3.4, part B.

Three different methods of fusing the data for final classification as either posed or spontaneous were tested to determine which was the most effective: early, mid-level, or late fusion. The different fusion methods represented different levels of abstractness in data representation, developing progressively from least abstract (early fusion) to most abstract (late fusion). For early fusion, basic features were computed over the tracked features to obtain speed and displacement of the features over time in a low-level, frame-by-frame fashion. Their 2007 paper continues along the lines of Cohn and Schmidt (2004) and Valstar et al. (2006), the main difference being that now there were three modalities instead of only one. This basic feature data for all three modalities was then fused into a single vector, and the final SVM classifier built for classification as either posed or spontaneous, see Figure 3.5. For mid-level fusion, the same type of mid-level features were used as in the brow work; three classifiers were built out of the primitive features that were used in early fusion, one for each of AU6, AU12 and AU13. Then the occurrences of AUs were segmented into onset, apex and offset phases, and the dynamics of these segments were then computed. These mid-level features, which represent a higher level of data abstraction and captured more dynamic and structural information than used in early fusion, were then fused into a single vector across all three modalities, before the final SVM classifier for these mid-level features was built. For late fusion, separate classifiers were built for each modality over both the early and mid-level features. These classifiers each classified behaviour as posed or spontaneous independently from

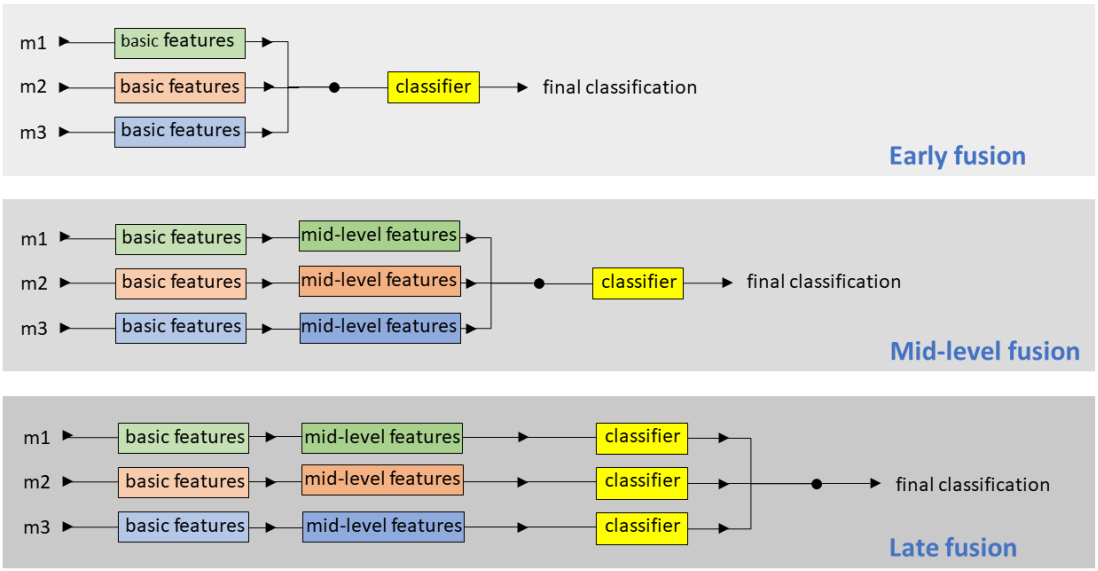

Figure 3.5: Workflow of early fusion (top), mid-level fusion (middle) and late fusion (bottom). m1, m2 and m3 are the three separate modalities.

the other two modalities. These three classifications for the three modalities were then fused using a sigmoid function into a single decision as to whether the behaviour was posed or spontaneous (Platt, 2000).

In their results, Valstar et al. (2007) show that combining all three modalities – head, face and shoulders - produces a better classifier than using any subset of these. In addition, due to its ability to join the results of specialized classifiers for each modality and phase, late fusion delivers the best classifier. The authors rank the relative importance of the three modalities, and, for the most part, head pose is more relevant for distinguishing posed from spontaneous smiles. They also compared the relevance of static versus temporal dynamics and concluded that temporal dynamics can better distinguish posed from spontaneous behaviour. As in the brow paper, they did not find any evidence that asymmetry is a good indicator of posed expressions. The accuracy of their best classifier for distinguishing posed from spontaneous smiles was 94 percent.

3.5 Detecting real high-stakes deception

In 2015, Pérez-Rosas and colleagues did a study on detecting deception in videos of real-life trials scenarios, which were collected from the internet (Pérez-Rosas et al., 2015). This dataset differs from posed or acted datasets, or datasets where the study subject is

instructed to lie, as the stakes were very high for the accused testifying. This affected their emotional state and arousal levels differently. The authors used the verdicts of the trials to establish the ground truth as to what was a lie and what was the truth. Altogether, there were 121 videos, each about 30 seconds long, 60 videos of 'honest' witnesses and 61 videos of 'dishonest' witnesses.

This was a multimodal study that involved verbal features as well as gestures, which included those related to facial expressions. It is being presented here because the authors extensively investigated the role of facial expressions, which they term 'facial displays'. They used the MUMIN coding system to describe eight displays involving face, head and hands, (Allwood et al., 2007). These are general face, eyebrows, eyes, gaze, mouth, lips, head movements, hands, and hand trajectory, leading to nine categorical values. Some facial displays correspond to action units and are signal-like, describing basic physical properties, such as 'open mouth' or 'lip corners up', 'lip corners down'. More complex actions can be described, though, such as 'laughing' to describe the general face, which is also more message-like.

The videos used in that study were not automatically annotated, but by two human annotators, and in a very simple way. For each video, for the facial displays, the annotators assigned the single value for each facial display that described the most prevalent state of that display in the whole video sequence. For instance, to describe 'mouth' for a video, the annotators had to choose one value, 'open' or 'closed', depending on which value in their opinion characterized the majority of the video. For each video, these values, along with verbal features, were used to build deception classifiers. The classifier models they tested were decision trees and random forests. They also explored classifiers built using individual features. They found that the best features for classifying deceitful behaviour were the facial displays. Facial displays alone led to a classification rate of 70% using decision trees and 76% using Random Forests.

3.6 How well can a computer spot a counterfeit crank?

In 2014, Bartlett and colleagues addressed the automatic discrimination of deceptive versus spontaneous behaviour in the context of physiological pain (Bartlett et al., 2014).

They chose this aspect of human behaviour because physiological pain is universally experienced by humans and it evokes a strong emotional response. There are also established and ethically acceptable means to induce both genuine pain in a laboratory setting as well as faked displays of pain. In their paper, they hypothesized that i) humans are bad at distinguishing real from faked displays of pain and ii) computers can distinguish real from faked pain significantly better than humans. To this end, they designed an experiment whereby they pitted humans directly against computers by giving them the same task of visually distinguishing real from faked pain. They then evaluated who performed the best. They base their assumption that faked and genuine displays of pain can be distinguished on the idea that there is a basic physiological difference between genuine and faked expressions of emotion because they are generated by two separate neuromotor pathways (Rinn, 1984). Genuine, spontaneous expressions of emotion are generated by the extrapyramidal pathway, which begins in subcortical regions of the brain. Faked, or volitional expressions of emotion are generated by the pyramidal pathway, which begins in the cerebral cortex. The extrapyramidal pathway is considered to be responsible for involuntary, reflexive motion and the pyramidal pathway for voluntary motion. Therefore, the question of visually distinguishing between deceptive and genuine behaviour is a question of being able to distinguish between the behavioural fingerprints of these two neuromotor pathways.

As in the previous three studies presented in this chapter, Bartlett et al. needed a dataset suitable for investigating their hypotheses. They created a database by using 45 human subjects, each of whom was given two tasks. For the first task, they were to submerge their forearm up to the elbow for one minute in a bucket of water at a temperature of 20 degrees Celsius, which was deemed to be not painful. While doing this, the subjects were instructed to make facial expressions that they believed could fool a professional doctor into thinking that they were in real pain and to continue this for the full duration of the minute. For the second task, the subjects were instructed to submerge their forearm up to the elbow for one minute in a bucket of water at a temperature of 5 degrees Celsius, the so-called *cold pressor test*, which is a common method to induce pain in a laboratory environment (Hines and Brown, 1932). While the subjects were performing these two tasks, videos of frontal views of their faces were taken at 30 fps. The order in which the experiments were done was the same over all participants: first, the faked pain (task 1) and then the real pain (task 2). The authors reasoned that, had the experiments been made the other way around, their subjects might have used their real experience of pain

to inform their fake expressions of pain, or might still experience residual pain from the cold water treatment.

3.6.1 Evaluating human performance

To evaluate how well humans can distinguish real from faked pain, 170 observers were chosen to watch videos of 25 of the 45 participants. The videos were randomly arranged in a sequence so that none of the subjects' pain videos and faked pain videos were directly next to each other. This was to prevent the observers from making direct comparisons between faked and real pain with the same subject. Using this setting, humans did not perform better than chance guessing at the task of distinguishing real from faked pain.

To level the playing field between humans and computers a bit, the authors designed another human experiment where 35 new observers, like the computer, were given a training phase to try and improve their ability to distinguish real from faked pain. The training phase they went through was as follows: they were shown the video pairs of 24 of the study subjects, but this time the real and faked pain videos of each subject were shown next to each other, but in mixed order. The observers had to decide which of the two videos for that subject exhibited the real pain and which exhibited the faked pain. They were given immediate feedback on the correctness of their decisions, so that they could inform their future decisions. After the training phase, their ability to evaluate real from faked pain was tested by randomly choosing one video per subject from 20 remaining participants. The authors concluded that training did not significantly alter human performance.

3.6.2 Evaluating computer performance

In addition to human observers, the 2014 study of Bartlett and colleagues also ran the *computer expression recognition toolbox* (CERT) action unit detectors on their videos. CERT is described in Bartlett et al. (2005). They used all 20 available action units. The outputs of the CERT detectors, like those of CNN-BLSTM and OpenFace, which I use in this study, estimate the presence or absence of each action unit in each frame of the video as a real number. Each of these 20 outputs was then put through eight different time-dependent filters ranging from three to 60 seconds to capture events at different

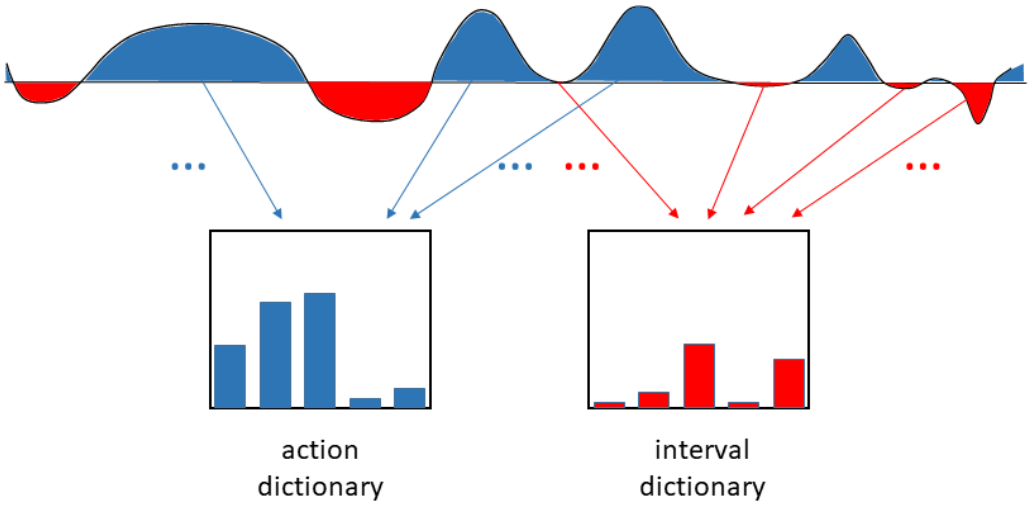

Figure 3.6: AU signal that has been put through one of the eight filters. Areas below curves representing actions (red) and areas above curves representing intervals (blue) are calculated and then the correct bin (word) for the value of each curve is incremented in the histogram dictionary, according to the 'bag of words' model.

time resolutions. Hence, each video was converted into 160 signal sequences containing 1800 frames, one sequence for each AU and one for each filter. Each of these action unit signal sequences was then segmented into 'actions' and 'intervals'. An action was defined to be the presence of the AU (section of the signal that had high values), while an interval was defined to be the absence of the AU (section of the signal that had low values).

To determine the size of each action, its area was calculated. These values were then recorded into a histogram with five bins. A similar histogram was made for intervals, as shown in Figure 3.6. These histograms were then treated as *bags of words* as done in Demirdjian and Wang (2009). This was based on a method of characterizing events in videos by different bags of features based on visual categorizations, taken over different timescales in order to detect long term and short term characteristics (Niebles et al., 2008; Laptev et al., 2008; Agarwal et al., 2004; Zelnik-Manor and Irani, 2001). In this way, the original signal of real values was converted into discrete 'words' and the problem of distinguishing between posed and spontaneous behaviour was converted into a 'bag of words' problem. After each video had been converted into 20*8*2 = 320 histograms, a feature selection method was used to build a final SVM classifier to distinguish posed from spontaneous behaviour; the histogram that produced the SVM classifier with the

best performance was then selected. The next histogram to be iteratively chosen was the one that produced the biggest improvement in classification. This greedy process was repeated until no more improvement could be obtained.

To compare the performance of their classifier with human performance, Bartlett et al. (2005) used *leave-one-out* on 25 video pairs; they removed each subject's pair of videos, trained on the rest of the subjects and then tested on the left-out subject (Bishop, 2006, page 33). They found that their classifier beat human performance and could distinguish faked from genuine pain with an accuracy of 85%.

3.7 A polygraph-like interrogation framework using computer vision

In 2018, Sen and colleagues presented an experiment with the aims to investigate dyadic human deceptive behaviour and to build a detector to distinguish between honest and deceptive behaviour (Sen et al., 2018). To carry out their experiment, they first needed to build a dataset. Here they focused on two aspects. The first aspect was that the dataset captures both deceptive and honest behaviour. The second aspect was that the video and audio was high quality to allow computer analysis and there should also be a large quantity of it to get statistical significance. They were particularly interested in investigating two hypotheses regarding deceptive behaviour. The first, set out by Ekman (1985), is 'duping delight', which states that deceivers enjoy the act of deceiving and this enjoyment might be revealed by their facial expressions, that is, they may smile more. The second, set out in *interpersonal deception theory* (Burgoon and Buller, 1996), focuses on the dynamics between the message sender and the message receiver. The theory posits that the temporal dynamics between senders and receivers is different depending on whether the sender is being honest or they are being dishonest to the receiver. Previous studies showed that synchronized head nodding between sender and receiver indicated more truthful interactions (Yu et al., 2015).

3.7.1 Creating a deception scenario

Sen et al. created a face-to-face interrogation scenario to be carried out by dyads communicating over a network. One member of each dyad was assigned the role of interrogator and the other the role of witness. This assignment was random. After this assignment, the witness was shown a picture and informed whether they were to be truthful or to lie to the interrogator about what was in the picture. The interrogator should determine whether the witness was lying or telling the truth. There were four phases of the interrogation, see Figure 3.7. In the first phase, the interrogator was instructed to ask a series of fixed questions. This phase was similar to polygraph questioning and was designed to get a snapshot of the witness's behaviour under different emotional circumstances - normal behaviour, slight confusion, memory recall, analytic thinking and discomfort. In the second phase, the interrogator was prompted to ask the witness a set of fixed questions about the photo they had been shown. In the third phase, the interrogator was prompted to ask the witness questions of their own choosing. After this, the interrogator made their first of two decisions about the honesty of the witness, if they were lying or telling the truth. In the fourth phase, the interrogator was given a hint about the photo the witness had seen and was then allowed to question the witness again. This last phase was based on the *guilty knowledge test* (Lykken, 1959), in which questions are asked that are designed to cause high levels of arousal in a person concealing knowledge. After this phase the interrogator made their second and final decision as to whether the witness was telling the truth. For motivation, the witnesses were promised $10 for each time they convinced the interrogator they were telling the truth. The interrogator was promised $10 for every time they correctly guessed whether the witness was lying or telling the truth.

3.7.2 Technical realisation

To carry out their experiment, the authors created the *automatic dyadic data recorder* (ADDR), which they propose to be a general purpose tool. The ADDR was designed to carry out their deception scenario while also allowing them to collect large amounts of high quality, annotated video and audio. The ADDR is a network program that brings dyads together over the internet. In this experiment, the researchers turned to two common sources for participants, Amazon's Mechanical Turk (Downs et al., 2010), which is

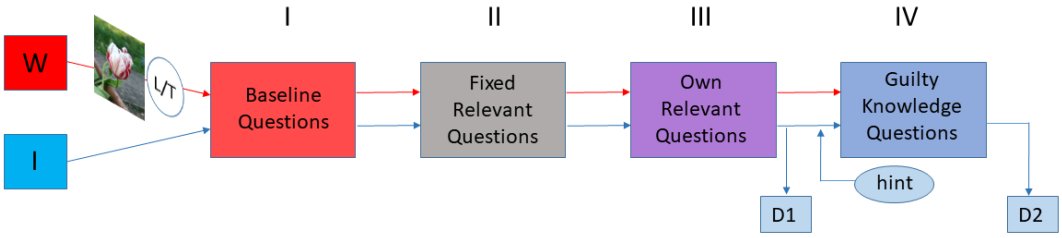

Figure 3.7: Interrogation workflow scheme. The witness (W) and interrogator (I) are randomly assigned. The witness is shown a picture and instructed to lie or tell the truth (L/T). The interrogation is four phases long (I-IV), with the interrogator guessing twice, D1 and D2, if the witnesses is truthful or not. One hint is given between the third and fourth phase of interrogation.

an online crowdsourced workforce, and email blasts, in this case to students of the School of Computer Science at their university asking them to participate (Sen et al., 2018). Participants were required to have a computer capable of transmitting audio and video. The ADDR first checked that potential participants had high enough quality equipment and that they were motivated. If the participants passed, they were then paired up and a session was scheduled.

The ADDR's mediated the experiment through interfaces. It set up telecommunication links between two participants, presented both participants with instructions through the interfaces, guided the course of the game, and recorded data. As each dyad carried out a face-to-face interrogation process over the internet in which they could hear each other and see each other's faces, the ADDR was able to capture clear video and audio of both participants, timestamp the data and annotate it with its game state and the participants' interactions, such as mouse clicks. Sen et al. collected data for 398 dyads, of which 151 were useable, thus fulfilling their aim of acquiring large quantities of data.

3.7.3 Experimental evaluation

To analyse their data, Sen and colleagues ran the action unit detectors of OpenFace on their videos of the dyads. They also created a baseline version of each AU for each witness as follows: from the first phase of baseline questioning, they took the segment that was intended to elicit normal behaviour, took the average value of the AU over that segment. They then subtracted this value from the rest of that witness's values for that AU in the following interrogation. This was intended to individually measure the change

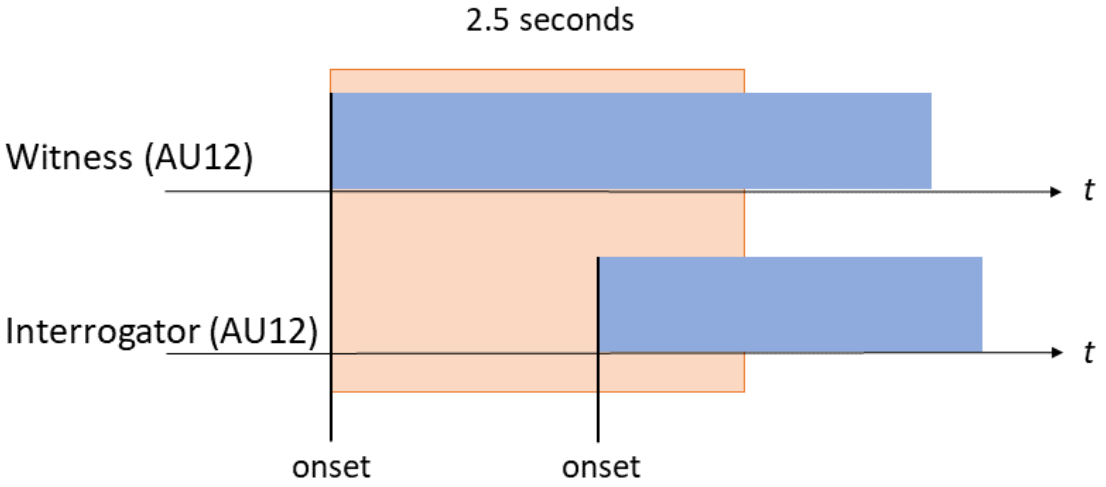

Figure 3.8: Diagram of a returned smile. Had the interrogator's smile begun outside the 2.5-second window (red rectangle), the witness's smile would have counted as 'not returned'.

in the witness's behaviour. They thus have both an original sequence of AU values as well as a baselined version for each participant.

To test the duping delight hypothesis the authors focused on AU12, lip corner puller. They looked to see if the average value of AU12 was significantly different between the truth tellers and the deceivers. They did this for both the original and the baselined AU values. They noticed that there was a tendency for the deceivers to show more occurrences of AU12 in the baselined AUs, but the difference was not significant. Thus, they did not confirm the duping delight hypothesis. The authors consider that this may be due to the fact that participants were informed before they were asked baseline questions as to whether they would be lying or telling the truth, and hence might have already been experiencing duping delight.

In a further set of experiments, the authors examined the temporal dynamics of AU12 between interrogator and witness to test interpersonal deception theory. They wanted to examine the frequency with which the interrogator returned the witness's smiles and vice versa to see if, in either case, there was a significant difference between honest and dishonest cases. They defined a smile, or more precisely an occurrence AU12, to be returned if, within 2.5 seconds of its onset, the other member of the dyad also smiled, see Figure 3.8. The authors found no significant differences in the rate at which the interrogator's smiles were returned based on whether the witness was telling the truth. However, perhaps counter-intuitively, they found that the rate at which the witness's smiles were returned was significantly higher for witnesses who were being deceptive. The authors

postulated that interrogators returned more smiles of deceitful witnesses than of honest witnesses, most likely because they are picking up on an altered behaviour of the witnesses that is not being detected by the AU detectors.

After investigating their two main hypotheses, the authors then carried out an exploratory investigation of the other action units computed by OpenFace, using both the raw and baselined values. They compared the average values of AUs between the truth tellers and the deceivers and looked for significant differences. The only statistically significant difference they found was in the baselined data, where there was a significantly higher occurrence of AU15, lip corner depressor, among deceitful witnesses than those telling the truth. The authors hypothesized that the significantly higher amount of AU15 among dishonest witnesses was probably caused by their simulating trying to recall details of a photo they never saw (Sen et al., 2018).

3.8 Refining detection of deception by creating AU contexts

A further study on the database presented in Sen et al. (2018) was done by Hasan and colleagues (Hasan et al., 2019). They were particularly interested in looking into the question, "Does looking into language patterns in light of facial expression contexts reveal any meaningful insight in understanding deceptive behaviour?" To investigate verbal patterns of deception they used the *linguistic inquiry and word count* (LIWC) (Chung and Pennebaker, 2012). LIWC contains around 4,500 words that are considered insightful to understanding a person's psychology and behaviour. These words are in turn partitioned into 64 categories such as 'conceptual' and 'cognitive'. To differentiate between deceptive and honest witnesses, Hasan et al. took counts of the number of LIWC words that occurred in each of the 64 categories in the spoken answers of the honest and deceptive witnesses and compared the two to see if there were any significant differences in any of he categories. They found that there were significant differences between truth-tellers and bluffers in the three LIWC categories 'seeing', 'conceptual' and 'cognitive'. They then looked at the values for the 17 AUs that were computed with OpenFace over the corresponding set of videos of the witnesses responding, and compared the averages of each of the AUs between honest and deceptive witnesses. Here, using the average AU

values as features in place of the LIWC features, they found no statistically significant differences between deceivers and truth-tellers.

To answer the question as to whether facial context can help reveal meaningful verbal patterns, they used the action unit values generated by OpenFace to create facial context for their LIWC features as follows. For each of the 17 action units computed by OpenFace and for each segment of video corresponding to a witness's answer they computed the average value. Then, for each AU and from all its average values per segment, they took the median. All segments that had an average value lower than the median were classified as low intensity occurrences of that AU, and all segments with an average value higher than the median were classified as high intensity occurrences of that AU. They thus partitioned their data into two groups - high and low intensity - and within each group looked once again to see how well they could differentiate between deceptive and honest witnesses using LIWC categories. They found that if they restricted their verbal analysis to those answers that were classed as high intensity for AU5, upper lid raiser (eyes wide open), the difference between average values in the 'seeing', 'conceptual' and 'cognitive' categories, which had already been statistically significant, increased. The authors concluded that averaged LIWC features alone could distinguish between deceivers and truth-tellers, however averaged facial features, as computed by OpenFace, could not. They attributed this to the fact that micro-expressions, which occur in a fraction of a second and could betray deception, are lost in averages that cover relatively long periods of time. Concerning their main research question, Hasan and colleagues concluded that facial context does sometimes strengthen signals when analysing linguistic features.

3.9 Conclusions and gaps in current research

The studies presented in this chapter indicate that AU detectors are useful devices for detecting and investigating human facial expressions in general. It is moreover becoming apparent that facial expressions can be valid indicators also for deceptive behaviours. There are, however, only a few works so far that apply AU detection of facial cues to decipher deception in human communication. In addition to the general paucity of knowledge in this area, my preceding review of the literature has identified several specific research gaps that pose obstacles to progress in the analysis of deceit using com-

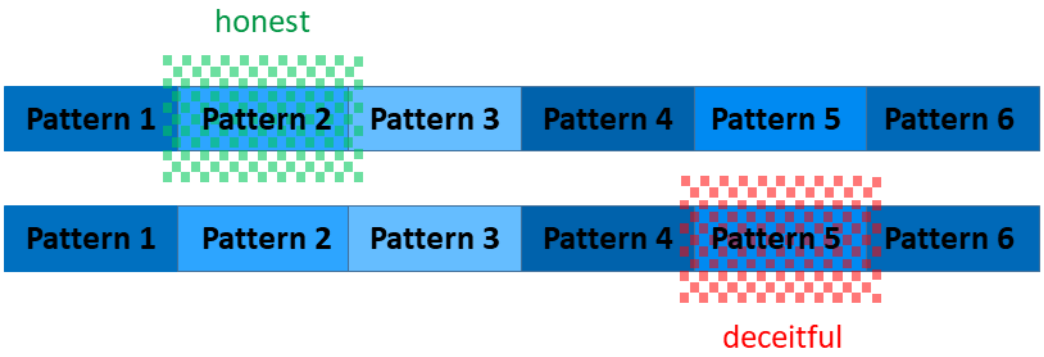

Figure 3.9: Selection of a single pattern from a sequence can introduce bias into the interpretation of an event. Pattern 2 could be associated with honesty and pattern 5 with deceit. Depending on which one is selected (red or green) to represent the whole sequence, the same event can be given opposing interpretations, namely honest (top) or deceitful (bottom).

puter vision. Given this small sample size, one such gap is the obvious question of how generalizable the current findings are, that is whether the deceptive cues that have been detected using AUs so far also apply to other deception scenarios. Another is the nature and quality of available data sets, such as the fact that most rely on posed rather than spontaneous behaviours, which is especially problematic when it comes to deception. Only a few of the past studies investigate deceptive behaviour where people lie without being instructed to do so and where they have an interest in not being caught (Pérez-Rosas et al., 2015). Others have a well-defined ground truth (Valstar et al., 2006, 2007). But rarely is there a convincing combination of both. I address these important limitations in Chapters 5 and 6, where I present a novel dice rolling experiment, which I designed together with behavioural economists Professor Roberto Hernan Gonzales and Professor Thorsten Chmura. This new experiment seeks to elicit genuine deceptive behaviours while also having an unambiguous ground truth.

In addition, several of the past studies do not consider videos in their entirety but extract the segments that display the deceptive behaviour they want to study (Ekman and O'Sullivan, 1991; Valstar et al., 2007). This method ensures that there is something to detect, but has the potentially serious downside that one cannot really be sure that the behaviour displayed reveals real-life deception. This problem is illustrated in Figure 3.9, which shows how a sequence of behavioural patterns can have two opposite interpretations depending on which pattern one selects to represent it. Thus representing a sequence based on a subsample is potentially erroneous. Looking at an entire sequence of behaviours, and not first manually selecting a revealing extract makes the task harder

of course, it becomes like finding a needle in a haystack. Sen and coworkers' ADDR takes this approach using the OpenFace detectors. But, as discussed in Section 3.7.3, they are not able to detect a significant difference in AU values between deceitful and honest behaviour (Sen et al., 2018). I address these problems in the next Chapter 4, where I similarly try to detect deception in a scenario where it is buried under many other behaviours in the context of the game of poker. However, I use a different approach to search for it, namely decision trees. Also, to gain a better understanding of the robustness of automated deceit detection, I use and compare two different AU detection systems, OpenFace and CNN-BLSTM.

Chapter 4

AUs and Decision Trees Identify Facial Cues Associated with Game Plays in Poker

4.1 Overview of how this chapter is organized

This chapter presents a computer vision study of the game of poker. It is organized as follows: First, I address the motivation behind the study, namely, there are many types of deception. I discuss what makes poker an interesting object of study. This is followed by a description of the poker dataset which is used in this study, in particular, how it was designed by the Institute of Creative Technologies at the University of Southern California to provide videos that are especially suitable for using computer vision to study human behaviour.

Next, I state my intentions to investigate this dataset using decision trees and discuss how I split the data into two classes, folds versus non-folds, and search a large space of different decision trees. The balanced classification rate is described, and I explain why I have chosen this as a measure of performance for determining good decision trees. Since there are two options for me to generate AU data, using CNN-BLSTM or OpenFace, before building decision trees I do a straight forward statistical analysis to see if splitting the data into folds versus non-folds (calls and raises) leads to any statistical differences for either detector. This also provides an opportunity to see how well the detectors agree with

one another which leads to a general statement about the performance of the current state-of-the-art in AU detection on spontaneous behaviour. This is an important consideration as AU detectors generally have difficulty in accurately detecting natural behaviour.

The results of the statistical analysis led me to use CNN-BLSTM for building decision trees. Searching a temporal space of decision trees led to a tree that classified folds versus non-folds with a balanced classification rate of 0.59. This tree also classified individual video frames as opposed to whole events, which usually consist of multiple contiguous frames. Applying a very simple and straightforward vote across all frames belonging to a single event provided an efficient way to join these individual classifications into a more intuitive single classification for an entire event and also improved the classification rate to 0.61. As this is nevertheless somewhat low, although not lower than expected given the nature of the problem, I once again compared OpenFace to CNN-BLSTM, mainly for the purpose of validating the CNN-BLSTM results.

To further reduce unwanted noise from AUs that weren't contributing to correct classification, I applied feature selection to another search for decision trees in order to find associations between AU displays and fold/raise/call behaviours.

Finally, to better understand how well the classifiers perform I designed an experiment to compare computer performance to human performance.

The work in this chapter substantially expands a preliminary study that I presented at Face and Gesture 2018 (Vinkemeier et al., 2018). This conference paper consisted roughly of the database description given here in Section 4.3, the discussion of the decision tree methodology and performance measure given in Section 4.4 and the initial decision tree search results given in Section 4.6. The statistical analysis in Section 4.5, the voting method in Section 4.7, the correlation test between OpenFace and CNN-BLSTM in Section 4.8, as well as the feature selection in Section 4.9 and the test of human performance on the poker dataset in Section 4.10 are new and appear for the first time in this thesis.

4.2 Motivation for studying deception in the game of poker

The first study carried out in this thesis is on the card game poker. Poker is well known as both a game of skill and a game of chance and it is one of the few settings in which it is socially acceptable to deceive. Poker is played for recreation, but it is also a gambling game that can be played for high stakes where each player has a chance to win or lose something of real value. Poker has grown into a multi-billion dollar industry with online gambling, television shows and several prestigious tournaments, such as the annual World Series of Poker (The Editors of The Economist, 2007).

As well as its entertainment and business aspects, there is also a lot of interest in poker from the perspectives of mathematics and game theory, as well as from the perspective of psychology. In 2015, the University of Alberta in Canada solved *Texas Hold'em*, the most famous poker variant, with a game-theoretic approach (Bowling et al., 2015). Shortly after, for the first time, two computer programs, DeepStack (Moravcik et al., 2017) and Libratus (Brown and Sandholm, 2018), separately beat professional poker players at Texas Hold'em. This represents a major advance in game theory and artificial intelligence and has applications in other fields like security and finance. In psychology, poker and other forms of gambling are considered to be "powerful tools for investigating risk-taking, decision making, and how the brain responds to personal gains and losses" (Jabr, 2010).

Poker is not only interesting to psychology, but psychology, like strategy and chance, is part of the game of poker. There is a strong social component to poker and players are constantly trying to guess whether their opponent's hand, which is at least partially hidden from them, is strong or weak. Among other things, players try to glean this information from their opponent's demeanour. Similarly, they themselves bluff and use their own demeanour to try and hide the strength of their own hand. The role of the appearance of a player's face during poker was studied in Schlicht et al. (2010), where it was discovered that a player deliberates more and makes more betting mistakes when they conceive of their opponent's face as trustworthy.

Slepian and colleagues tested the abilities of human observers to judge how strong a professional poker player's hand is (Slepian et al., 2013). The observers were not pro-

fessional poker players. They compared three different modalities - the upper body, the arms alone and the face alone. The observers were split into three separate groups and tasked with watching videos of the poker players in play. All videos were silent and approximately two minutes long. The observers then judged how strong the players' hands were based on the videos alone. It was found that a first group, who had been shown videos only of the upper body, performed as good as random but not better. A second group, who had observed the players' arms only, performed better than random, showing that there was meaningful information conveyed by the arms that the players did not manage to conceal. A third group of observers, who had relied on face-only cues, performed significantly worse than random. This implies that the professional players had successfully duped them with their facial expressions. This, in turn, suggests that there is meaningful information conveyed by the face but that, due to human subjectivity, human observers are deceived and may misinterpret it.

Perhaps it is possible that an unbiased computer focussing on facial cues might be able to detect deception in poker without falling into the trap of deceit. The study presented in this chapter investigates this question. It focuses on automatically distinguishing between when a player is about to 'fold' versus 'call' or 'raise' based on action units detected in videos. To my best knowledge, this study is the first study to use computer vision to analyse human behaviour in poker or any other card game.

4.3 Construction of a poker dataset for the purpose of using computer vision

To carry out the study, I first needed an appropriate dataset. This dataset needed to capture poker behaviour in such a way that good quality videos of the players' faces were provided for computer vision analyis. This study is based on such a set of videos of poker games and their corresponding metadata, which was donated to the Computer Vision Lab, UoN, by Professor Jonathan Gratch from the Institute of Creative Technologies of the University of Southern California. The poker games depicted in it were designed to produce a set of videos that could be used for machine learning analysis of dyadic human behaviour with special emphasis on facial cues. At least one other dyadic experiment in addition to poker was performed in the same video recording sessions, the *Iterative*

Prisoner's Dilemma (IPD). It is not part of this study. So far, four studies have been published on the IPD dataset, Stratou et al. (2015), Stratou et al. (2017a), Hoegen et al. (2017) and Stratou et al. (2017b). These four works investigate emotions elicited while playing IPD, a game where each member of the dyad can choose to either cooperate or behave selfishly. On the poker dataset, in contrast, although preliminary experiments were done on it (Lu and Pantage, 2015, unpublished), no other study has been completed yet, except for that presented in this thesis and published as part of this PhD (Vinkemeier et al., 2018).

4.3.1 The participants

The participants, who were not professional poker players, were recruited through Craig's List, which is a website that posts classified ads (Smith, 2019). They were offered a flat fee of $30 to play poker and the IPD; they did not play poker for real money, but instead they could win tickets for a lottery worth $100 if they played well.

4.3.2 Design of the poker game

The game played was a simplified version of poker. It consisted of ten rounds of poker, where, in each round, the players were each dealt a single card on the screen whose value was between 2 and 10 with the highest card winning in case of a showdown, also known as a call. As is typical in poker, the participants did not get to see their opponents' cards. A player won if either a call occurred and their card was higher or if their opponent folded, that is, conceded in order to avoid a bigger loss. Player A went first on odd numbered rounds, player B went first on even numbered rounds. The game was slightly complicated by the fact that the first bet in a round could be zero or positive, but a fold was not possible. Otherwise, the players had the choice to either call, raise or fold with the restriction that the total bet could not go over 25 and bet sizes were always multiples of five (0, 5, 10, 25). These details became clear after examining the database. Figure 4.1 shows a screen shot of the interface the player interacted with. The player's opponent is shown in the right half of the screen, their card on the other. Although from each player's perspective the card order seemed random, the sequence of ten cards was fixed according to Table 4.1. All A players saw the same ten cards as did all B players. This

Figure 4.1: The poker game as seen by a player. On the left, they see their card and details about the state of the game. On the right, they see their opponent and a thumbnail of themselves in the lower left corner.

Player\Round	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
A	2	7	9	3	4	9	6	4	8	10
B	7	6	10	2	8	4	3	5	10	4

Table 4.1: Fixed 10-round card order for players A and B. The player that goes first in each round is highlighted in blue.

fact remained hidden from the players and was not a problem since no player played in more than one game. After rounds three and seven, the players were given a brief multiple choice questionnaire about what they thought of their opponent's betting strategy, which lasted usually around 20 seconds. The entire game lasted around 5 minutes. A segment of a game can be seen in Figure 4.2.

4.3.3 Data capture: collecting players' videos along with their time stamped and annotated events

The players played in pairs, A versus B, over a local area network (LAN) where they could see each other by means of a computer monitor. Their faces were videoed at 30 frames per second by a webcam embedded in the monitors where the games were depicted and across which the players visually communicated with each other. Each game thus produced two near frontal view videos, one of each player's face, as they were naturally focused on the monitor for most of the game, see Figure 4.3. In order to

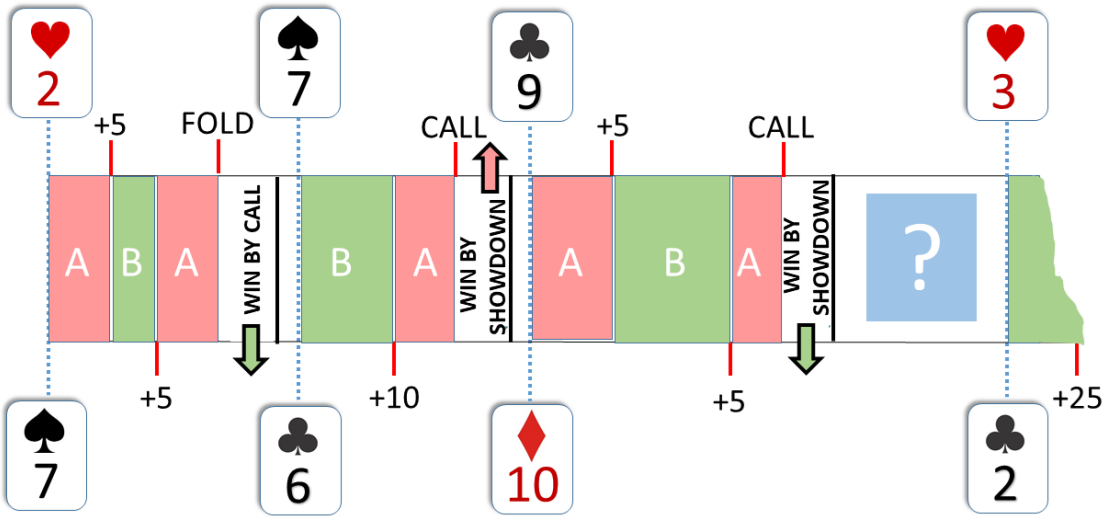

Figure 4.2: An excerpt from a game depicting rounds 1,2,3 and the beginning of round 4, as well as the questionnaire, represented by a question mark in a blue square, occurring between rounds 3 and 4. The upper sequence represents the sequence of cards and moves of Player A (red), the lower sequence represents those of Player B (green). Bet amounts are preceeded by '+'. The width of the red and green boxes is proportional to the duration of the corresponding play.

elicit non-verbal communication and to avoid confounding facial expressions with facial movement associated with speech, there was no audio and the players could not speak to each other.

Since the games took place over a LAN and were controlled by computers, players made bets by means of mouse clicks allowing many of the major events of the game to be automatically timestamped. There were timestamps for the beginning of each round, when each of the bets were placed, and when the two short questionnaires occurred. Timestamps for bets were activated by the players as they decided when to make their play by mouse click. The beginning of the game, when the players were shown their cards and when they were given the questionnaires were automatically generated by the program. Bet amounts and the answers to the questionnaires were also recorded. There were no other annotations to the database, nor were there later any added for this study, except for those of the automatic action unit detectors which I ran on the database, see below in Section 4.4.1. Therefore, the database as I received it, consisted only of the images of the videos and the timestamps and bet values, all of which were created automatically.

We were given 104 videos by the Institute of Creative Technologies, although originally there were more. This was because some participants had not given their permission to share their videos. In an additional spreadsheet I had been given, forty more of the 104

Figure 4.3: Frames from a near-frontal view video of a pair playing poker.

individual participant's videos were recommended by the Institute of Creative Technologies for possible removal. This was due to various reasons: they involved confederates, the video quality was poor, there had been technical difficulties such as lags in the timing that had bothered the participant, or the player did not understand the game. I removed all forty of these in one go before carrying out this study. Therefore, this study includes the remaining 64 videos. As a comparison, the four studies mentioned above that use the related IPD dataset reported using many more videos, ranging from 186 participants (93 pairs playing) to 604 participants (302 pairs playing). The poker set described here provided a dataset which could be analysed by means of computer vision and machine learning. The next step was to decide what methods and approaches to use for analysing it.

4.4 Methodology -detecting folds using facial AUs

In this chapter, action unit detectors are used to study the facial expressions of participants playing poker in the poker database. The concrete question asked is, *can a clas-*

sifier be built that automatically predicts if a person will fold or not based on their facial expressions? The respective facial expressions are not known and they are likely to be subtle and fleeting as they are spontaneous and associated with passing events. Different players may respond emotionally differently to the same situation and therefore, there might not be a single facial expression associated with a decision to fold or not, but rather a group of facial expressions. There will also likely be overlap with facial expressions with some occurring in both classes, such as the neutral face. For these reasons the baseline for detection is likely to be low.

To increase the likelihood of finding facial expressions associated with folds, calls and raises, this study focuses on the time window immediately surrounding the player's decision to fold or not. Park and colleagues also focused on the time window surrounding events for finding clues to negotiation outcomes (Park et al., 2015; Park et al., 2012). I am searching this entire time window with multiple classifiers, each covering a different subset of continuous video frames within this window. Individual frames of AU values, corresponding to video frames, are used as inputs and not aggregates of these such as averages, under which important information might get buried under multiple facial expressions or noise. Hence, I first approach the problem by classifying individual frames. Later, these multiple per-frame classifications get fused together into a single classification. This study focuses on static facial expressions as the time windows are short by necessity, since game plays last only a few seconds. Therefore, it is not advisable to collect data over long periods of time as these would span different events in the game. There is a good chance that a static, or instantaneous, expression will carry a lot of useful information as, according to Ekman, "any time slice within that apex (peak intensity of the facial expression) carries information about which emotion is being signalled" (Ekman, 2009).

The action unit detectors were combined with classifiers to discover if there is a facial expression, or a small enough set of facial expressions, common to enough different players in the dataset to make it possible to predict whether a player was going to fold or not. I wanted to address the question of whether such expressions exist and when and for how long their signals are strongest. Different classifiers were thus built to test different temporal positions relative to the decision to fold, call or raise and also for different spans of time. The idea is that if there is a common facial expression to detect at a certain point, the classifier covering that area will classify well, whereas classifiers

Action Unit	Description	Corresponding muscle group
AU1	Inner brow raiser	frontalis (middle part)
AU2	Outer brow raiser	frontalis (outer part)
AU4	Brow lowerer	procerus, depressor and corrugator supercilii
AU5	Upper lid raiser	levator palpebrae superioris, superior tarsal
AU6	Cheek raiser	orbicularis oculi
AU9	Nose wrinkler	levator labii superioris
AU12	Lip corner puller	zygomaticus major
AU15	Lip corner depressor	depressor anguli oris
AU20	Lip stretch	risorius, platysma
AU25	Lips part	depressor labii, mentalis, orbicularis oris
AU26	Jaw drop	masseter
AU45	Blink	levator palpebrae, orbicularis oculi

Table 4.2: The twelve action units used in this thesis and their associated muscle groups.

covering areas where there is no such occurrence will perform close to random. In order to increase the chances of detecting something, the search was centred around the player's mouse clicks, as presumably at this time they are thinking most intensely about the move they are about to make. The next step would be to decide how to prepare the data to be input into the chosen classifier.

4.4.1 Details of preparing the data for learning a classifier

Altogether, in the 64 videos of the poker games used in this study, there are 675,432 frames of video. Each frame consisting of 640×480 pixels. The twelve CNN-BLSTM AU detectors were run on all frames in this dataset to extract information about the action units. To gain a better understanding of the problem, which involved detecting low-level and spontaneous facial expressions, I also ran the OpenFace detectors. The twelve action units used are the same ones shown in Table 2.1 and repeated again here in Table 4.2. This extracted the relevant information about facial expressions from the videos while reducing the complexity of the frames and simplifying the problem. Each video frame of 640×480 pixels was replaced with a vector of 12 real-valued numbers between 0 and 1, representing the intensity of the 12 action units. Figure 4.4 shows the action unit intensities detected on a player at three instances, once when he folds, once when he calls and once when he raises. The timestamps that mark when players make their decision to fold, call or raise will be referred to as *FCR-events*.

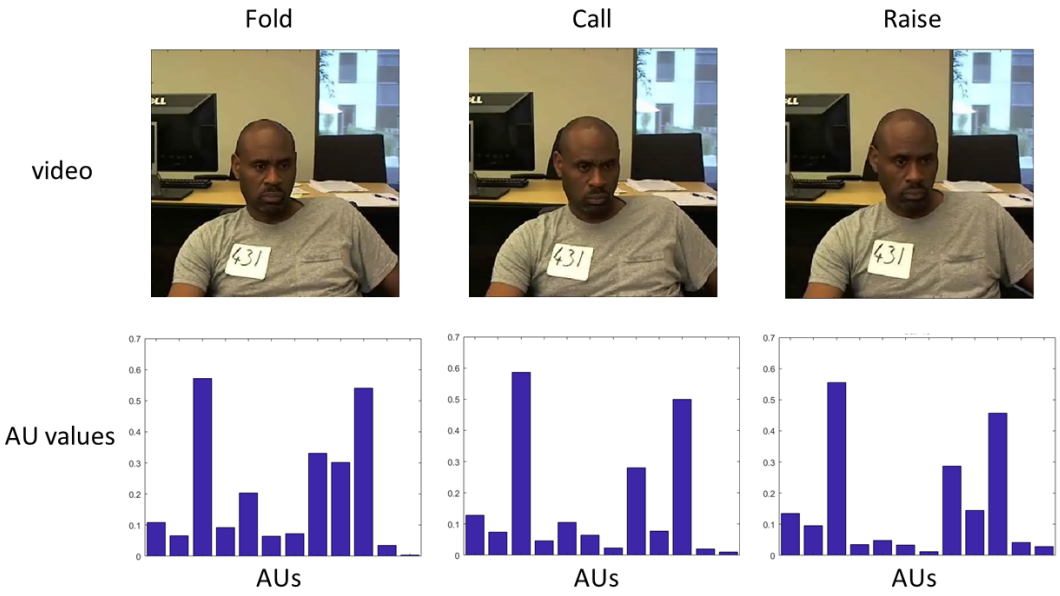

Figure 4.4: This figure shows video frames (top) with their corresponding action unit intensities for all 12 action units as computed by CNN-BLSTM (bottom). The three different events are shown: the player folds, calls and raises as indicated.

To prepare input for learning a single classifier, two parameters had to first be determined, namely *offset*, that is, when the search starts, and *duration*, that is how much time it should encompass. Once this was determined, it was applied to all players. The FCR-event timestamps are found for all players and used as the points of reference. The question 'when' is answered by choosing an offset relative to the FCR-event, such as 'one second before' the player moves (an offset of -30 as the videos are 30 frames per second), and the question 'how long' is answered by assigning a duration, such as for 'half a second' (duration = 15). Then the frames corresponding to this window are selected for each player and labelled class 1, a 'fold', if the FCR-event was a fold or otherwise class 0, 'not a fold', see Figure 4.5.

More precisely, let e be a video frame corresponding to a player's choice to fold, call or raise, meaning it is the one that is temporally closest to their mouse click. If the current offset is o and the current duration is d, then the d contiguous frames from $e + o$ to $e + o + d - 1$ are individually labelled 1 if e refers to a fold and 0 if e refers to a call or a raise. This labelled set of frames over all players and events is used to learn the current tree. Classifiers differ from each other only in their offset/duration parameters. These offset/duration values are always chosen so that all frames stay within a window of nine seconds. Having checked the data, I ascertained that nine seconds was the longest period for which no frame appeared in two sets simultaneously which would lead to it being an

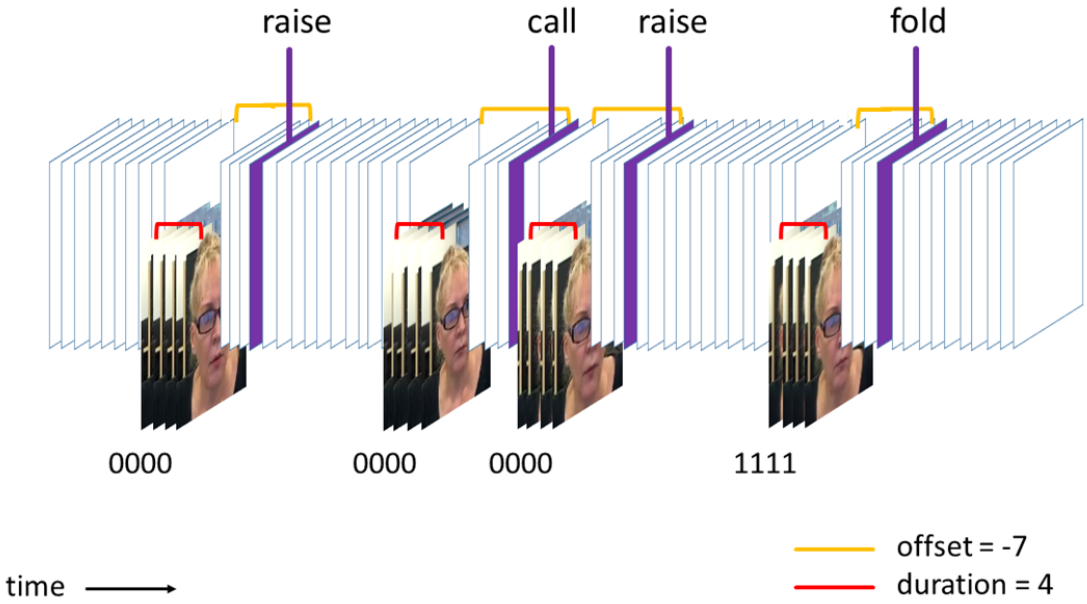

Figure 4.5: An illustration of which frames of each player's game are extracted for fixed offset/duration parameters in order to learn the corresponding tree. In this case, the tree is the one corresponding to the parameters offset = -7 and duration = 4.

input twice, possibly even having different labels if the two events it was associated with had different classes.

In order to find the period of highest detectability, the different classifiers generated by different offsets and durations were tested in an exhaustive fashion and compared. This is possible to do if the classifiers can be constructed quickly. The assumption is that if there is a classifier that detects well, it is likely that it has found a facial expression common among players associated with the decision to fold or not. For the first part of developing such a classifier, I focused on optimizing classifiers built over individual frames. Later, I looked to see if the classifications of the individual frames associated with an FCR-event could be turned into a single classification for the whole event. It turns out this could be done. In order to begin a concrete investigation, the next step was to decide on what type of classifier to use.

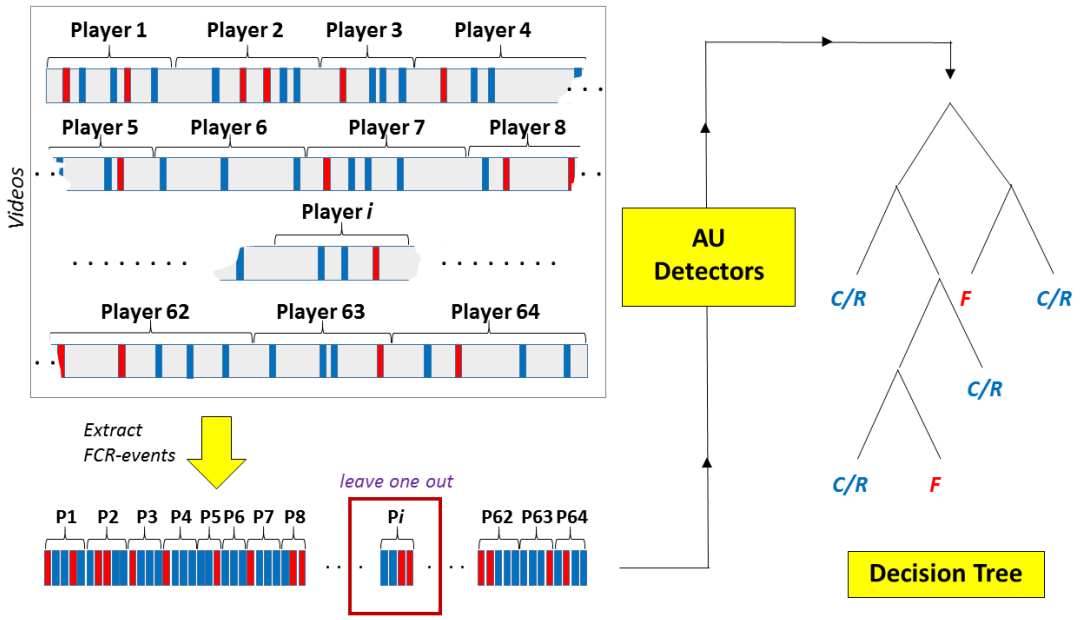

Figure 4.6: An overview of the method by which decision trees were created showing the FCR-event frames (blue and red rectangles) being extracted from the poker videos, converted frame-by-frame into 12 action unit values each and then being used to learn a decision tree. To approximate the performance of the tree, I used leave-one-subject-out.

4.4.2 Decision trees: an appropriate classification model for the problem

I chose the decision tree model for this investigation because of its clear conception and its ability to describe a configuration in terms of intervals. This corresponds to the way facial expressions are described in terms of AU values, which is intuitive. Decision trees can express any function and simultaneously can be built very quickly and they are appropriate for smaller amounts of data with fewer features. Additionally, the purity function, which describes the homogeneity of a set (to what extent its elements belong to a single class) used in constructing a tree, is a good way to filter out noise generated by the AU detectors. Decision trees are well established methods and explained in most books on machine learning, such as Mitchell (1997). There are many variations. The one used here is CART (classification and regressions trees) as implemented in MATLAB. The design and analysis of CART is described in Krzywinski and Altman (2017) and is laid down in detail in Breiman et al. (1984).

CART decision trees are recursively built. Starting with all the labelled training samples at the root node, a splitting criterion is applied and the data split into two sets. The

split that increases the purity of the child nodes the most is chosen. As a result, at each training node the training data is partitioned into two sets. These become child nodes and the splitting algorithm is then applied recursively to these. This continues until no improvement in purity can be achieved or stopping criteria have been reached.

The tree used here was built using splitting questions of the form, "is the value of a particular action unit greater than some threshold?", where a search is made over all possible thresholds over all 12 action units to find the best overall threshold, or split of the data, at the current node. The CART algorithm reduces the impurity at child nodes by maximizing the decrease in impurity (or increase in purity). It uses the following formula for calculating the change in impurity:

$$\Delta i(s, n) = i(n) - p_L \cdot i(n_L) - p_R \cdot i(n_R)$$

and calculates this over all possible candidate splits s. Here, the node which is being split is denoted n. The proportion of instances that will go to the left child according to split s is denoted p_L, the proportion of instances that will go to the right child according to s is denoted p_R. The impurity of a node n is denoted by $i(n)$. A node is pure if the instances it contains all belong to one class and is maximally impure if the instances it contains belong in equal proportion to both classes. The formal requirements for an impurity function over a node n are defined in Breiman et al. (1984), page 32, as

- The impurity function over n can have only one maximum, and that is when n contains equal proportions of both classes.
- The impurity function over n can have only one minimum, and that is when n contains only one class.
- The impurity function over n is symmetric with respect to the classes.

MATLAB's implementation of CART uses the Gini index of diversity to define the impurity of a node n. The Gini index fulfils the above requirements. It is defined as

$$i(n) = 2 \cdot p(1|n) \cdot p(0|n).$$

Here, $p(1|n)$ denotes the proportion of class 1 examples at node n and $p(0|n)$ denotes the proportion of class 0 examples at node n.

Therefore, the plan was to convert each of the video frames of the poker dataset into a labelled vector whose attributes were 12 real values between 0 and 1. The offset and duration parameters for each decision tree in the search space would then be applied and the corresponding vectors over all players would be used to learn a decision tree. For an overview of this, see Figure 4.6. Before actually beginning this experiment, however, I wanted to compare the data generated with CNN-BLSTM and OpenFace to help decide which would be better suited for the task.

4.5 Statistical look at the data

Before building the decision trees, I looked at some basic statistics of the data to get ideas for guiding the search and also to become more familiar with basic properties of the data. This was done to help decide which detector, CNN-BLSTM or OpenFace, to use, to determine how well they detect and also to look for any helpful patterns that might appear in the data. In this thesis, statistics is approached from the perspective of machine learning. To investigate basic statistics, I used the t-test, which is the most frequently applied test in the literature when detecting human behaviour using action unit detectors. I considered both the paired and the unpaired t-test. There is also an interdisciplinary aspect to the studies and in psychology and behavioural economics many other tests are considered in addition to the t-test such as ANOVA (analysis of variance) (Sirkin, 2006, Chapter 13) and regression models (Sirkin, 2006, Chapter 10) such as multilevel regression and OLS (ordinary least squares). In addition, in the psychological sciences family-wise corrections such as the Bonferroni correction are also used much more frequently than in machine learning. These will be left to future publications on these aspects where appropriate. Some of the t-test results, however, will be given as a full statistic in the style common for publications of the American Psychological Association (APA). In that case, I give the following parameters: The t degrees of freedom, which indicates how many independent values there are in calculating the statistic; the t statistic, which indicates how similar the data is to the null hypothesis; the p value, which describes how likely it is that the data was generated by the null hypothesis; and Cohen's d, which indicates the size of the effect (Cohen, 1988). I will note when I am using this style.

	Game duration	Round duration	Play duration
Average	5 min 39 sec	34 sec	13.5 sec
Maximum	11 min 2 sec	3 min 14 sec	2 min 44 sec
Mininum	2 min 52 sec	6. 2 sec	1.3 sec

Table 4.3: Durations of game events.

When regarding the data in the poker dataset for statistical analysis, I organize it into folds versus calls and raises as was planned for building the decision trees. Altogether there are 64 included participants with each participant's game consisting of 10 rounds. Each round contains at least two plays. The number of FCR-events over the 64 players is 481. These break down into 132 folds, 165 calls and 184 raises. I did not consider the first play of each round in the analysis because the participant who took the first move of the round did not have the choice to fold, but had to place a bet, a peculiarity which was mentioned earlier in Section 4.2.2. This reduced the number of plays in the dataset and might also have caused some confusion for the participants, especially those who might have wanted to fold immediately. The data showing durations of games, rounds and plays (fold, call or raise) is shown in Table 4.3. On average, the games lasted a bit longer than 5 minutes, but could go up to 11 minutes. Rounds also had a large range being as short as 6.2 seconds and as long as three minutes. A play, or decision to fold, call or raise, could be as short as 1.3 seconds or as long as nearly three minutes. It is likely that players who took much longer than the average 13.5 seconds to make their play, such as the player who took 2 minutes and 44 seconds, were trying to understand the game or were distracted by something else, such as another person in the room as this did occasionally happen. It therefore makes sense to focus on the time directly surrounding the FCR-event as this is when the participant is concentrating on their decision.

Average values of the 12 AUs were computed using the CNN-BLSTM and OpenFace action unit detectors. These are shown in Table 4.4 and in Table 4.5, respectively. The mean value for each of the twelve AUs was computed, for all the data, for a fold set and for a call/raise set. Fold and call/raise sets had to be defined before their averages could be computed. To split the data into a fold set, I used those frames defined by an offset -60 and duration 90 for the fold events of all 64 participants. So the AU values for a particular fold include all those from the frames between two seconds before the fold button is clicked and one second after. Each fold thus includes three continuous seconds of data. This way the data should be relevant to the fold, but short enough to not overlap

	AU1	AU2	AU4	AU5	AU6	AU9
all	0.23	0.19	0.23	0.13	0.09	0.10
fold	0.24	0.22	0.17	0.16	0.10	0.09
c/r	0.22	0.19	0.25	0.12	0.09	0.10
	AU12	AU15	AU20	AU25	AU26	AU45
all	0.10	0.12	0.08	0.18	0.09	0.08
fold	0.12	0.12	0.08	0.19	0.10	0.08
c/r	0.10	0.12	0.07	0.18	0.09	0.08

Table 4.4: Averages of all 12 AUs using the CNN-BLSTM detectors. First rows (all) show averages over all data. Second rows (fold) show averages over folds only, including two seconds before the fold to one second after. Third rows (c/r) show averages over call/raises, also two seconds before the call/raise to one second after. The larger value between fold and call/raise for each AU is highlighted in yellow.

	AU1	AU2	AU4	AU5	AU6	AU9
all	0.0386	0.0167	0.0817	0.0090	0.0588	0.0141
fold	0.0366	0.0151	0.0764	0.0111	0.0645	0.0150
c/r	0.0394	0.0174	0.0836	0.0082	0.0566	0.0137
	AU12	AU15	AU20	AU25	AU26	AU45
all	0.0575	0.0456	0.0278	0.0881	0.0872	0.0242
fold	0.0626	0.0414	0.0265	0.0926	0.0831	0.0216
c/r	0.0555	0.0472	0.0283	0.0864	0.0888	0.0252

Table 4.5: Averages of all 12 AUs using OpenFace detectors represented as in Table 4.4. Here, four decimal places are shown as there is very little change in the values within an action unit. The larger value between fold and call/raise for each AU is highlighted in yellow.

into other events. Call and raise data were made analogously. In both the Tables 4.4 and 4.5, the higher average between fold and call/raise is highlighted. The OpenFace and CNN-BLSTM detectors agree on which value is larger for AU4, AU5, AU6, AU12, AU15 AU25 and AU45. They disagree on which is larger for AU1, AU2, AU9, AU20 and AU26. This could be obtained if one or the other detector produced random results. Since it was not clear which of the values are statistically significant, I looked into this next.

It is difficult to come up with a clear and informative statistical test for the poker dataset since the data for an individual participant are highly dependent, but still one wants to distinguish between an individuals folds and non-folds and group them with the folds and non-folds of the other participants. The most natural way to view the data set is to

segment it into folds and call/raise events. This, however, produces a set of data that is not fully independent and not fully paired, as some players both fold and call or raise. On the other hand, giving each player a single averaged fold value and a single averaged call/raise value is also not ideal, as some players either never folded and in this case several players have to be thrown out and their data lost. For this reason, though neither is ideal, I looked at both options.

Independent t test. To see if there is any significant difference in AU values between fold and call/raise events, I first used an unpaired, two sided t-test on the data. The first value I tried was for an offset of -60 and a duration of 90 (three seconds of values starting two seconds before the fold, 90 frames of video). The average values of each of the 12 AUs were taken for each individual event and each of these values became a sample input to a separate t-test testing that specific AU. Therefore, if a participant folded three times and called/raised four times, they would produce three values for the fold sample and four values for the call/raise sample for each AU. In this case, each of these values would be averages of the 90 frames chosen as detemined by the offset and duration parameters. Averages were taken because individual frames that are neighbours in the video are highly dependent. For each of the 12 AUs, there were thus 132 values for the fold sample and 349 values for call/raise samples. For CNN-BLSTM detectors, the values for AU4 (brow lowerer) and AU5 (upper lid raiser) resulted in significant differences between the two classes. The APA style reporting of the statistics for AU4 are t(479) = 2.87, $p = 0.0043$, d = -0.28 and the statistics for AU5 are t(479) = 3.24, $p = 0.0013$, and d = 0.30. They are both significant at 5%, which means that there is a five percent or lower chance of randomly achieving these values, the lower the p-value, the stronger the evidence that the values are not random. In addition, their effect sizes are both small to moderate. Having gotten this result, I looked to see how this significance changed over different offsets and durations. The p-values were plotted for different offsets and a duration of 60 frames, as shown in Figure 4.7, left. The significances of AU4 and AU5 were fairly stable, but tended to be stronger near the FCR-event at offset 0. For OpenFace, also shown in Figure 4.7, right, there were no significances, except for AU25 (lips part) nine seconds before the event, which is probably too early to be associated with the event. The results obtained with the CNN-BLSTM detectors suggest that participants who folded raised their eyebrows significantly more and lowered them less than those who called/raised. The data generated by the two different detectors do not agree, though, and do not corroborate each other.

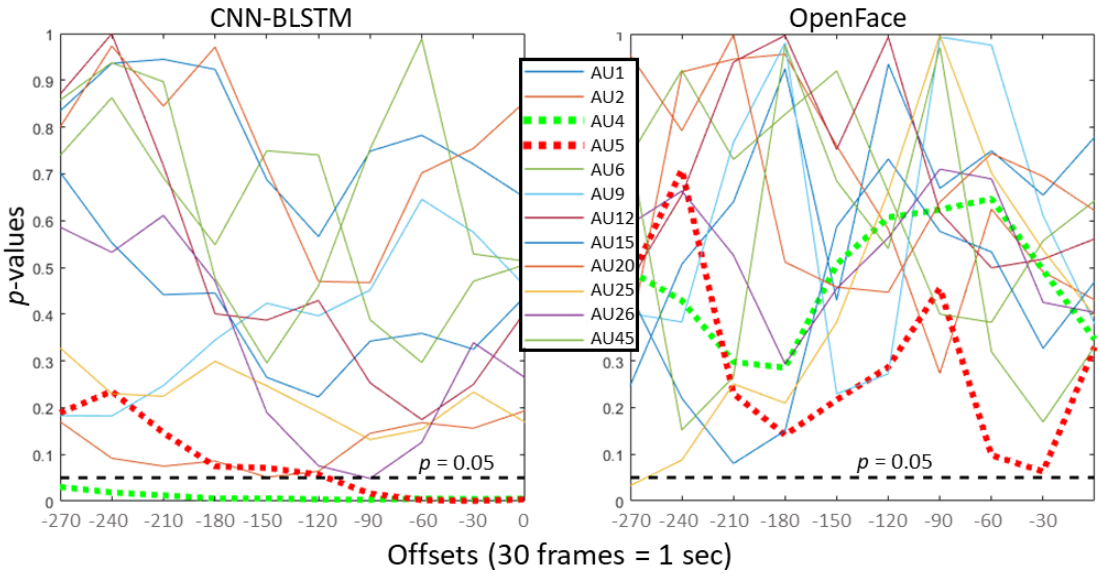

Figure 4.7: The p-values for the two sample unpaired t-test plotted according to offsets along the x-axis and duration of 60 frames (2 seconds). Higher p-values indicate less significance, lower p-values indicate higher significance.

Paired t test. The two sample t-test might not produce accurate results for this data, however, because the data are not normally distributed as seen by the histograms shown in Figure 4.8, where values generated by the CNN-BLSTM detectors are shown on the left and values generated by OpenFace are shown on the right. With 132 fold samples and 349 call/raise samples, there is enough data that this should not be a problem. However, the two samples overlap with 56 of the 64 participants contributing to both the fold sample and the call/raise sample, and eight participants being represented in only one of the two sets. This contradicts one of the main assumptions of using the t-test, namely that the data be either fully paired or fully unpaired, as the t-test for partially paired data might give invalid results (Moore et al., 2017). For this reason, I decided to also use the paired t-test.

To prepare the data for the paired t-test, those participants (eight in total) who did not both fold and call/raise at least once were removed as there seemed no reasonable way to split them into a pretreatment and treatment set, thus some of the data was lost. For each of the remaining 56 participants and for the different offsets and duration of 60 frames, the average of each AU was taken over all their folds and added to the fold sample and then taken over all their call/raises and added to the call/raise sample. This created a paired t-test with 56 paired samples. The p-values for different offsets and a

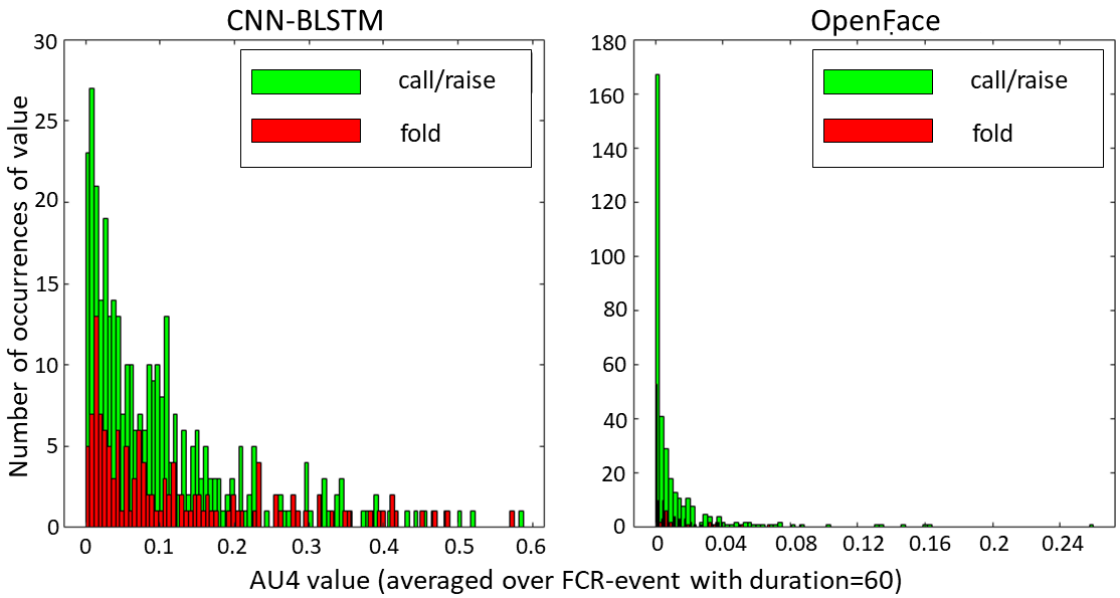

Figure 4.8: This figure shows histograms of the data used for the unpaired t-test for AU4. On the left is the data generated by CNN-BLSTM. On the right is the data generated by OpenFace. Values that make up the fold samples are red. Values that make up the call/raise samples are green. Each value is the average of a single FCR-event. The data do not look uniformly distributed.

duration of 60 frames are shown in Figure 4.9 for both detectors, for comparison. I also give a more thorough statistical picture of the data that is more in keeping with how statistics in psychological studies are presented for an offset of -30 and a duration of 60 frames. Here, there was statistical significance for AU5 and AU12: AU5 resulted in $t(27) = 2.35$, $p = 0.02$, $d = 0.31$. AU6 resulted in $t(27) = 2.20$, $p = 0.03$, $d = 0.29$. As the other AUs did not give statistical significance, I will not compute their fuller statistics. For AU5 and AU12, these statistics indicate they are both significant at the 5% level and have a medium to small effect. To investigate whether the data follow a normal distribution in order to determine how appropriate the t-test is, histograms were again made for visualization. The data shown in the histograms are the differences between each pair of values for each participant. Applying a one-sample Kolmogorov-Smirnov test for the right scale and mean showed that the hypothesis that the data comes from a normal distribution cannot be rejected. In Figure 4.10, one can see that the data look normally distributed but not like the standard normal distribution. To get a good scale for the standard deviation and to shift the mean, I subtracted the mean from both sets and scaled the data to different values before using the test. Most AUs passed the test for the right scaling factor, whereby the scaling factor was found by testing values linearly between the mean and the outermost values that the AU took on. It seems plausible,

therefore, that they have a normal distribution.

In conclusion, according to the paired t-test, the CNN-BLSTM detectors indicate that AU5 (upper lid raiser) and AU12 (lip corner puller) are significantly more pronounced when people fold, Figure 4.9, left. The results of OpenFace suggest something different, namely that AU6 (cheek raiser) and AU25 (lips part) are more pronounced when people fold, Figure 4.9, right. Although AU25 occurs very early. Here, I give the APA style statistic for AU6 at offset -60 and duration 90 as t(27)=2.41, p=0.02 and d = 0.32. Therefore, this is significant at 5% and has moderate to small effect.

It would have been a good confirmation of their ability to pick up subtle, spontaneous behaviours if the two detectors had agreed more. That they don't indicates that at least one of them is likely to be unreliable at detecting the type of facial behaviour in the poker dataset. I will return to this aspect again and compare the two detectors from a different perspective in Section 4.8. Altogether though, the tests indicate that splitting data according to folds and calls/raises might lead to results with CNN-BLSTM and that AU5 and AU12 might be particularly reliable for these detectors.

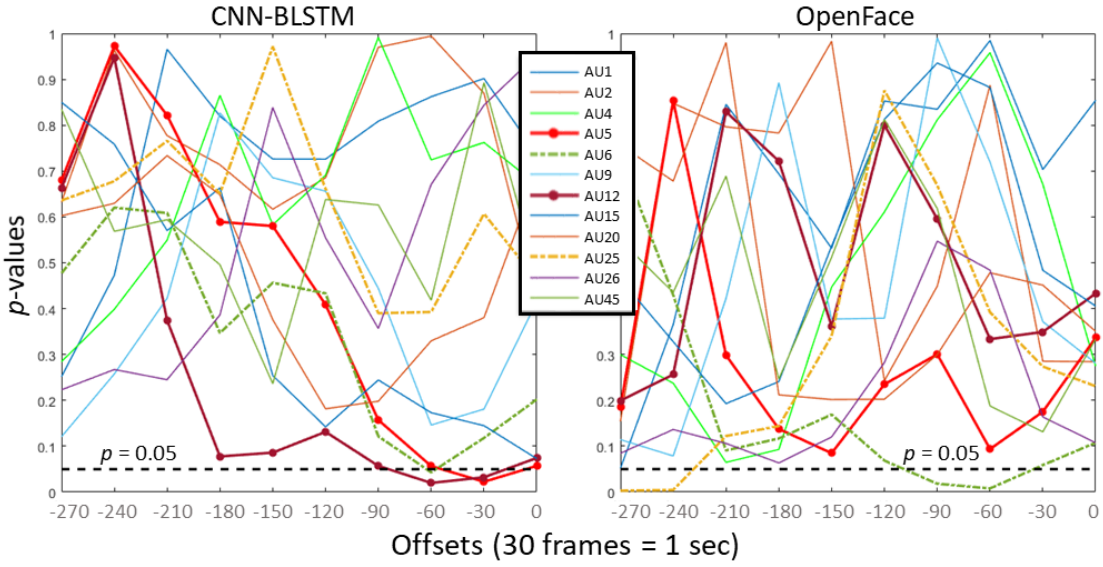

Figure 4.9: The p-values for the paired t-test plotted according to offsets along the x-axis and duration of 60 frames (2 seconds).

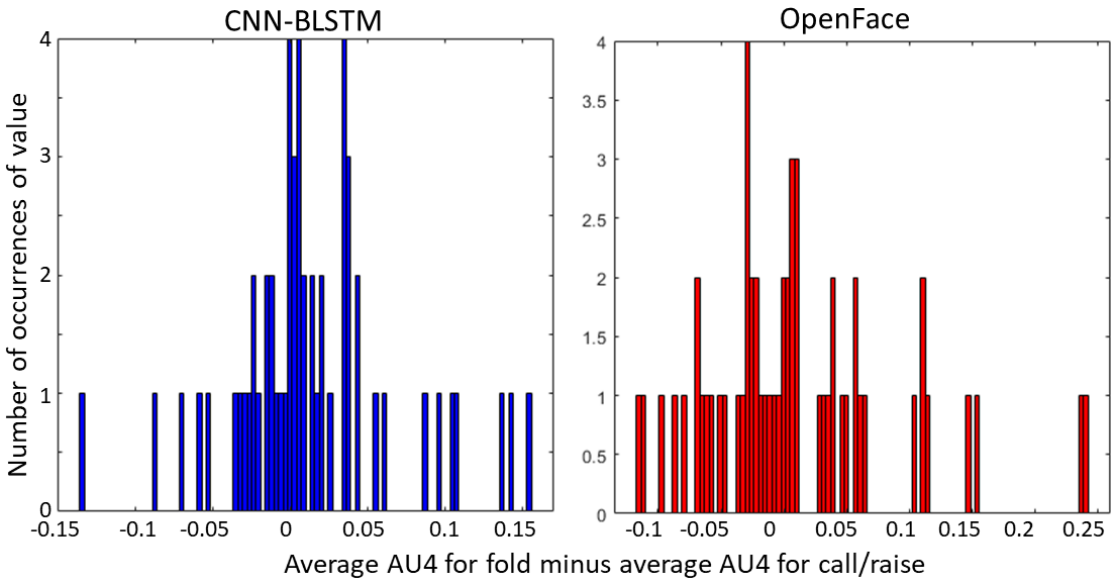

Figure 4.10: This figure shows histograms of the data used for the paired t-test for AU4. On the left, in blue, is the data generated by CNN-BLSTM. On the right, in red, is the data generated by OpenFace. Each data value consists of a player's fold-average minus their call/raise-average. The data look normally distributed.

4.6 Evaluation of first decision trees using CNN-BLSTM

The first application of decision trees began with all the 12 AUs computed with CNN-BLSTM even though only AUs 5 and 12 were significant in the tests of the previous section. It would not be reasonable to restrict analysis to these two AUs as many behaviours are signified by groups of AUs occurring simultaneously. I searched a space of decision trees $T_{offsets,durations}$ over different offset and duration pairs in order to discover if and when folds could best be detected. The distance between some FCR-events was just over nine seconds. In order to avoid overlapping events I focused on decision trees learned on frames that did not precede their corresponding FCR-event by more than nine seconds (-270 frames). One second after a round ended with a fold or call, the players were informed of the outcome (win or lose). Three seconds after this, they were dealt their cards for the next round, or the game was over. Therefore, I restricted the offsets to five seconds after the FCR-event as by then the participants were well into the next round or the game was over. The search space hence was restricted to offsets in the range of 9 seconds before the FCR-event to 5 seconds after and to durations of $\frac{1}{30}$ of a second (one frame) to 9 seconds. This is a broad enough range, most likely still encompassing too large a window, but it is not so large as to allow a single frame to be

included in two events, which is problematic, as previously mentioned.

In order to evaluate the performance of each of the decision trees, I used a leave-one-player-out protocol. So, for each offset/duration pair and for each of the 64 players p, a decision tree was learned without the frames for that player, making that particular decision tree independent of that player. This tree was then used to classify the left out player's frames. In this way, all the classifications for all the players were collected and used to estimate the performances of each tree $T_{offset,duration}$.

For the performance measure, I chose to use the classification rate R, which is defined as

$$R = \frac{C_1 + C_0}{N}.$$

Here, C_1 is the number of correctly labelled fold instances, C_0 is the number of correctly labelled call/raise instances, and N is the total number of instances. There are 481 FCR-events in the database. Of these, 132 are folds, 184 are calls, and 165 are raises. The ratio of fold events to call and raise events is therefore 1:2.65. This is also the ratio of fold to call/raise instances used to learn and test any given tree, since the number of instances used to learn a tree is just 481$\times$duration, as individual frames are input to the tree. Thus, the data is imbalanced. In this case, simply assigning every instance the class 0 (call/raise) gives a classification rate of 0.73. However, this only gives information about the distribution of the data and not how well a classifier can distinguish between a fold and a call/raise. Therefore, to get a more informative measure of this, the folds have been scaled to have equal weight as the call/raises. This has led to a balanced version of the classification rate, which can be interpreted as the classification rate in the case that, in the test data, the number of folds equals the number of calls and raises combined (Tharwat, 2018). The balanced classification rate R^b is defined as

$$R^b = \frac{P_1 + P_0}{2}.$$

Here, P_1 is the proportion of correctly labelled class 1 (fold) examples to all class 1 examples, and P_0 is the proportion of correctly labelled class 0 (call/raise) examples to all class 0 examples. Similarly, the balanced precision, recall and F1-measure for the trees were computed. For the precision, which measures the probability that an instance

	Regular	Balanced
Classification rate	0.6681	0.5938
Precision	0.4019	0.6398
Recall	0.4293	0.4293
F1-measure	0.4151	0.5138

Table 4.6: Performance values for decision tree $T_{15,3}$.

classified as a fold is really a fold, the balanced version gives

$$Precision^b = \frac{P_1}{P_1 + (1 - P_0)}.$$

For the recall, which measures the probability that a fold will be returned by the classifier as a fold, the balanced version is equal to the standard version and is given by

$$Recall^b = P_1.$$

The F_1-measure is the average of the precision and recall and combines them into a single value. The balanced version is given by

$$F1^b = 2 \cdot \frac{Precision^b \cdot Recall^b}{Precision^b + Recall^b}.$$

The results for the best decision tree, $T_{15,3}$, occur at $offset$ = 15 frames and $duration$ = 3 frames, see Table 4.6. In the case of this tree, there are 1443 frames. The method of leave-one-player-out labelled 170 of the 396 class 1 frames correctly and 794 of the 1047 class 0 frames correctly. The performance measures are listed in Table 4.6.

I also looked into the statistical significance of the results and compared the balanced classification rate of the classifier with a fair coin, that is one that has a probability of $\frac{1}{2}$ of landing heads, and used the binomial distribution with the statistical significance level .99. Instead of using individual frames, which are clearly dependent, I considered different FCR-events to be independent of each other, as they never overlap and usually have a gap of several seconds between each other. There were altogether 481 such events. Considering the decision tree with the best performance, which occurs at 15 frames (half a second) after the player makes their choice to fold or not, and has a duration of 3 frames (a tenth of a second), its balanced classification rate is .5938. If one were to classify 481 instances and get a .5938 proportion of these correct, then compared to random coin

flipping this would be significant. However, the data is imbalanced with .73 instances being classed as 0. This makes statements about significance difficult. If one considers the worst classifier, which occurs just short of five seconds before the event and has a duration of 3 frames, as a negative classifier, it has a balanced classification rate of 0.5651. Although this is lower than the classification rate of .5938 of the best classifier, it is also statistically significant when compared to a fair coin. This points out the difficulty of analysing spontaneous human behaviour data, which likely has a low baseline due to individual differences, like personality, where there are strong interdependencies and where the data is imbalanced as is the case here. It is worth noting that using a family-wise error correction here would have removed any statistical significance as thousands of trees were tested to find the best one.

I also created a heat map of the decision trees, see Figure 4.11, with offset values increasing along the x-axis and duration values increasing along the y-axis. This was done in order to discover if there were consistent areas where the detectors could better pick up folds versus calls and raises, which would indicate if and when there was a signal. The heat map could also indicate if the performance of the detectors was clustered at or near the FCR-event or if it was randomly distributed. Figure 4.11 shows that just more than four seconds before the FCR-event the classifiers begin to perform better, the performances then peak around half a second after the FCR-event and rapidly decline at 1.5 seconds after the event.

In order to view this from a different perspective, a bar graph was created, see Figure 4.12, which provides a cross section of the heat map shown in Figure 4.11. It also depicts the performance of the classifiers as time elapses, using the identical offset schema as in Figure 4.11. But this time, for each offset i, I plotted the best performance ratio from among the five classifiers made from durations of 1 to maximally 5 frames. This limit was chosen as otherwise adjacent classifiers made with longer durations begin to contain overlapping frames. Figure 4.12 thus shows nearly the same data as Figure 4.11, focused on the best classifiers at each offset, but restricted to durations in the range of 1 to 5 frames. One can see from both representations that the ability to detect the fold versus the call/raise increases four seconds before the event, peaks at half a second after and decreases rapidly at 1.5 seconds after. At 9 seconds before FCR-events, the players are once again in FCR-events. At 4 seconds after an FCR-event, the players are receiving their next cards. Since the card order for the players is fixed, it could be that the decision

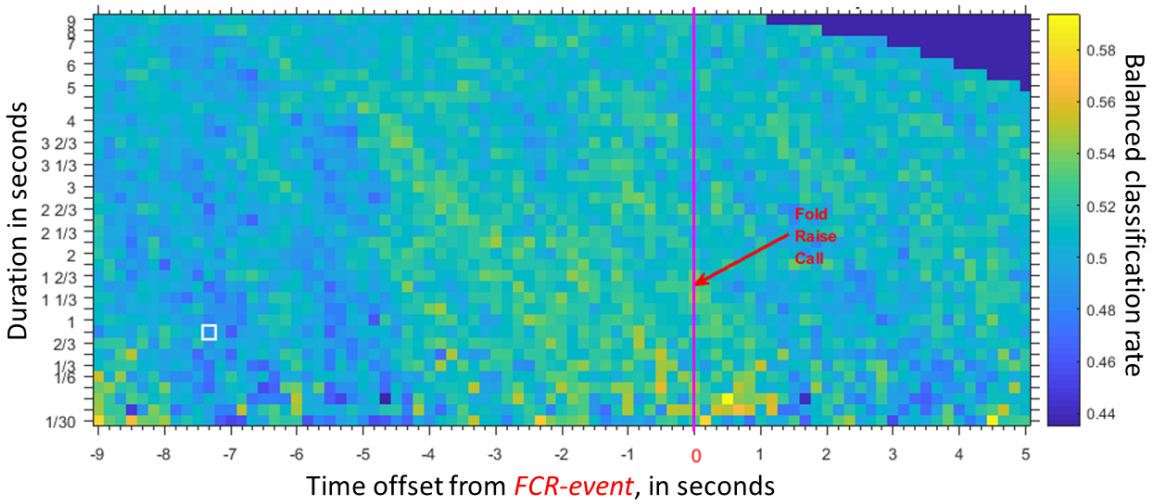

Figure 4.11: Heat map of the performance ratios of different classification trees. Each square in the heat map shows the balanced classification rate of a decision tree. Each entry represents a decision tree trained for a particular offset (x-axis) and window length (y-axis). The x-coordinate 0 represents the time of the FCR-event. Left and right of this are offsets in seconds (30 frames per second). The y-axis represents different durations, in seconds, increasing in ascending order. For example, the white-framed rectangle at $x = -7\frac{1}{3}$ and $y = \frac{5}{6}$ gives the balanced performance ratio of $T_{-7\frac{1}{3},\frac{5}{6}}$, which is 0.4856.

trees at these ends of the heat map are detecting other correlations between the games of the players.

The best decision tree, $T_{15,3}$, whose performance is given in Table 4.6, has been visualized and is shown in Figure 4.19 in Section 4.9. It has been placed in this later section in order to facilitate comparing it to the best tree found using feature selection. As shown in Figure 4.19, the first split in the tree uses AU5 at a threshold of .185 which splits the tree into two subtrees with larger AU5 values going right. The resultant right hand subtree tree contains data that is 64% fold events and 36% call/raise events, showing that this one split at the root already has good discriminative power. This also reflects the results of the paired t-test which showed that there were significantly higher AU5 values present when folding. The left subtree has high entropy and is only a bit better than maximum entropy. It is not clear if AU12 plays a particularly important role in the right subtree. In the left subtree, at a depth of three, high values of AU12 are associated with folding. This is also aligned with the paired t-test. Altogether, it is not clear if AU12 plays as important a role as AU5.

In summary, these analyses were an attempt to see if it was possible to detect, possibly beforehand, a player's objective actions in the game of poker from only their facial ex-

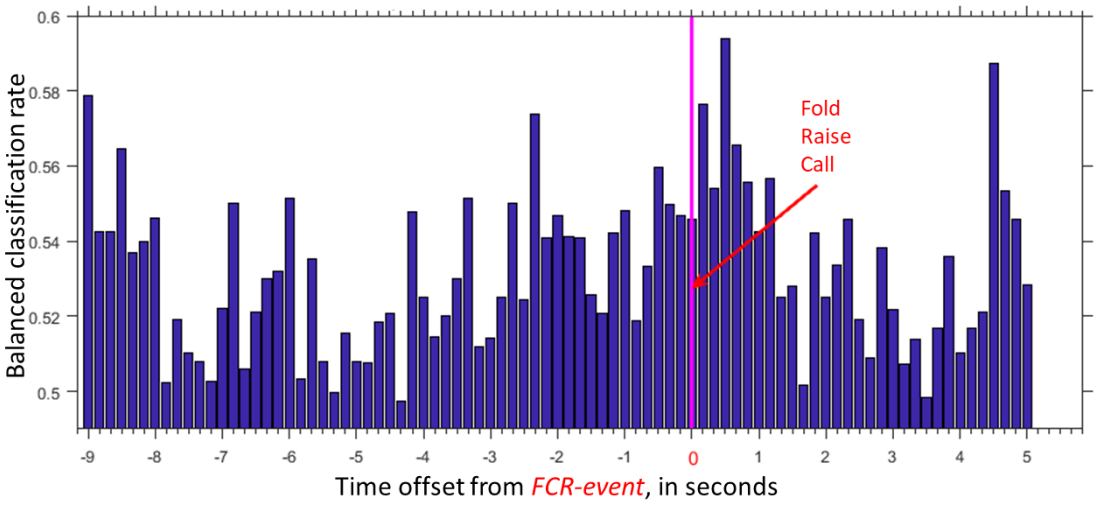

Figure 4.12: Bar graph showing best performance of classifiers over durations of 1 to 5 frames. The x-axis of this figure is identical to the x-axis of Figure 4.11. A single bar in the graph represents the best classification rate from the five classifiers at that offset with durations 1–5.

pressions. In particular, I wanted to see if folds could be detected. In the four seconds leading up to the FCR-event, classifiers were obtained with classification rates that were statistically significant at the 5% level; $T_{-100,4}$, just over 3 seconds before the event, had balanced classification rate 0.55, and $T_{-70,1}$, just over 2 seconds before the event, had balanced classification rate 0.57. However, the strongest detection obtained was half a second after the event at $T_{15,3}$ with a balanced classifiaction rate of .5938. Thus it seems feasible to detect subtle and spontaneous facial actions of humans that indicate future decisions using action unit detectors together with decision trees. The classifications thus far are for individual frames instead of events. It would be more intuitive and direct to classify a whole event rather than just its individual frames. This issue will be adressed in the next section.

4.7 A simple voting method for converting multiple frame classifications into single classifications

The classifiers made so far return a separate classification for each frame associated with an event. It would be more intuitive, though, to return a single classification for the whole event. In order to obtain a single classification for a FCR-event, the classifications of the

frames relating to that event were taken together in a vote and a cut-off value of 1/2 was chosen, see Algorithm 1. This cut-off was chosen as it is the most straight forward and simplest and it reflects the fact that the decision tree was trained using a balanced classification rate. Voting was done by taking the classifications for the individual frames of an event and summing them. If the sum of the frames was greater than $\frac{1}{2} \cdot (duration)$, then the event was assigned the class 1, otherwise it was assigned the class 0. This voting system altered the performance of the decision tree as shown in Table 4.7.

Data: input $(c_1, ..., c_{duration})$, $c_i \in \{0, 1\}\}$, classes assigned by decision tree to frames corresponding to a FCR-event e
Result: A single classification C for the event e
1 **if** $\sum_i c_i > \frac{1}{2} \cdot duration$ **then**
2 | C = 1;
3 **else**
4 | C = 0;
5 **end**

Algorithm 1: Find majority class.

	Regular	Balanced
Classification rate	(0.6681) 0.6840	(0.5938) 0.6079
Precision	(0.4019) 0.4265	(0.6398) 0.6628
Recall	(0.4293) 0.4394	(0.4293) 0.4394
F1-measure	(0.4151) 0.4328	(0.5138) 0.5282

Table 4.7: Performance values for decision tree $T_{15,3}$ after voting. Values for the same tree before voting, see Table 4.6, are shown in red. At this point, all 12 AUs have been used for the decision trees.

Comparing the values in Table 4.7 to those in Table 4.6 shows that voting and choosing the majority class provides a straightforward way to classify the whole event. It furthermore improves all eight values used in evaluating the performance of the classifier. This method of classifying a whole event as opposed to frames only also extends the previous decision trees conceptually.

4.8 Comparing the correlation of CNN-BLSTM and OpenFace on the poker dataset

The best classifier found so far had a balanced classification rate of 0.61. To find it, an extensive search had to be carried out that required calculating thousands of decision trees. Given that the classification rate of 0.61 is somewhat low, this raises the question of whether what is really being detected is dependent on noise. Watching the videos on the poker dataset along with their concurrent AU values shows that the detectors are indeed somewhat noisy for the subtle, low-intensity facial expressions that occur in the poker dataset. This, together with the fairly low classification accuracy is a motivation to take a closer look at the performance of the CNN-BLSTM detectors on this dataset. There is however no ground truth for the action units in the data set, so there is no direct way to evaluate this performance. For this reason, and also since OpenFace is freely available and widely used in research (Agarwal et al. (2019) and Rudovic et al. (2018) are recent examples), I did another comparison between OpenFace and CNN-BLSTM. Here, I looked at how well the two were correlated over the same signal. If the two agree with each other strongly, then it is likely that they both detect spontaneous behaviour well. This is a confirmation that CNN-BLSTM detects well. If they disagree, then one or the other or both might be unable to detect spontaneous behaviour well, but no strong statement is possible. Therefore, I used the Pearson correlation coefficient to calculate the linear correlation between the two detectors over the AU values they output for the poker database.

Pearson's correlation coefficient for samples is defined by

$$r_{xy} = \frac{\sum_{i+1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i+1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i+1}^{n}(y_i - \bar{y})^2}}.$$

Here, x_i is the AU value generated by CNN-BLSTM for the ith frame in the sequence and y_i is the the AU value generated by OpenFace for the same frame in the same sequence. Outputs for all the players were concatenated into one input for calculating the correlation coefficient. The computed correlation coefficients of the CNN-BLSTM and OpenFace detectors are shown in Table 4.8 for each of the twelve AUs under study. The correlation

coefficients, shown in column 1 of Table 4.8, were especially high for AU6 (cheek raise), AU12 (lip corner puller) and AU25 (lips part). As expected, randomly shuffling the order of the entire sequence of frames for one detector brought the correlation coefficients to near zero (these values are not shown). Since OpenFace replaces low AU values with zero as part of its normalization process to remove bias, I also tested the correlation between CNN-BLSTM and OpenFace after removing these zero valued frames. The results are shown in column 2, with the number of frames removed for each AU shown in column 3. The correlation coefficients of AU6, AU12 and AU25 remained high and it does not seem that OpenFace's normalization (which happens twice, see Section 2.6.2) had a large impact or caused the two detectors to deviate too much.

AUs	1 plain	2 zeros removed	3 frames removed	4 plyr shuffle	5 plyr avg
AU1	0.2005	0.2554	362,943	0.0615	0.3030
AU2	0.1552	0.2764	231,618	0.0454	0.2290
AU4	0.1203	0.1162	395,243	0.0729	0.1835
AU5	0.1138	0.1413	196,392	0.0144	0.1434
AU6	0.5078	0.5364	330,699	0.1986	0.4497
AU9	0.0635	0.0769	251,412	-0.0142	0.1599
AU12	0.6698	0.6283	291,112	0.3044	0.5253
AU15	0.1334	0.1633	378,796	0.0620	0.1066
AU20	0.0494	0.0654	305,024	0.0189	0.0933
AU25	0.4869	0.5243	525,308	0.1424	0.4485
AU26	0.2585	0.3057	526,414	0.0963	0.2757
AU45	0.1085	0.1232	341,351	0.0012	0.1304

Table 4.8: Correlation coefficients for OpenFace and CNN-BLSTM. Column 1 shows the correlation coefficients for each AU over the whole dataset, where the players' data has been concatenated together in sequence to form one input. Column 2 shows the correlation between CNN-BLSTM and OpenFace when all the frames were removed for which OpenFace assigned that AU a value of zero. Column 3 lists, for each AU, the number of frames removed from the total of 675,432 frames in the calculation of the previous column. Column 4 gives the correlation coefficient when the frames of individual players have been permuted among themselves. For column 5, the correlation coefficients were computed for each participant separately and the average taken. The three AUs with the highest correlation coefficients are constant across the columns and are highlighted red.

It is also possible that the detectors were correlating in their player biases and not in frame-by-frame activity based on behaviour, as they frequently have person-dependent bias (Baltrušaitis et al., 2018); the detectors could be agreeing in this as they change from participant to participant, see Figure 4.13 for a graphical explanation of how this might

look. Therefore, to investigate this possibility, for CNN-BLSTM, I shuffled the frames within each player but left the sequence otherwise intact, see column 4 in Table 4.8. The frame order for OpenFace was not altered. This generally reduced the correlation coefficients strongly, suggesting the detectors do indeed correlate in their detection of behaviour, but AU12 (lip corner puller) still remained quite high even after permuting a player's frames, indicating that bias might play some role in the correlation of the two detectors for this particular AU. To get another view of how well the detectors agree, I also computed the correlation coefficients for each player separately and then averaged these, see column 5 in Table 4.8. The values for the top three AUs remained high. I draw from this that both detectors detect these three AUs well even in the case of spontaneous behaviour.

To further assess the basis of the correlation of the CNN-BLSTM and OpenFace detectors, I made a bar graph of each of their outputs for an arbitrary participant, see Figure 4.14. I chose to visualize AU12 (lip corner puller) and AU45 (blink) because the first, AU12, has a high correlation coefficient and the second, AU45, has a low correlation coefficient, and, in addition, these two AUs are located in different parts of the face. To the naked eye, the bar graphs of AU12 for OpenFace and CNN-BLSTM appear to correlate well, similar to their coefficients in column 1 of Table 4.8. In contrast, the bar graphs of AU45 to the naked eye do not appear to correlate very much, in line with their low correlation coefficient in column 1 of Table 4.8. I therefore conclude that AUs 6, 12 and 25 are fairly reliable signals for both the OpenFace and CNN-BLSTM detectors. It is not possible to reach such a conclusion about the other AUs, though.

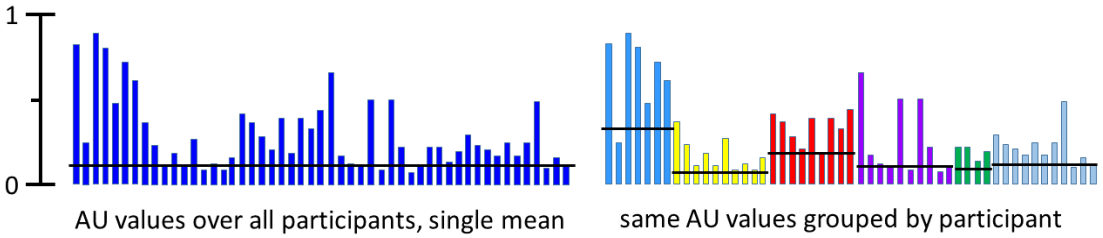

Figure 4.13: Bar graph, for visualization purposes, depicting the possible effects of person-dependent bias. The full signal used for calculating the correlation coefficients is the concatenation of the signals of six individual players. It is known that action unit detectors can have a strong subject specific bias. The signal on the left is identical to the signal on the right, but on the right the signals of the different players are highlighted, making it clear that they have different means, a factor which could also dominate the correlation coefficient. Black horizontal bars indicate signal means, on the left over the whole signal, on the right means are broken up per player.

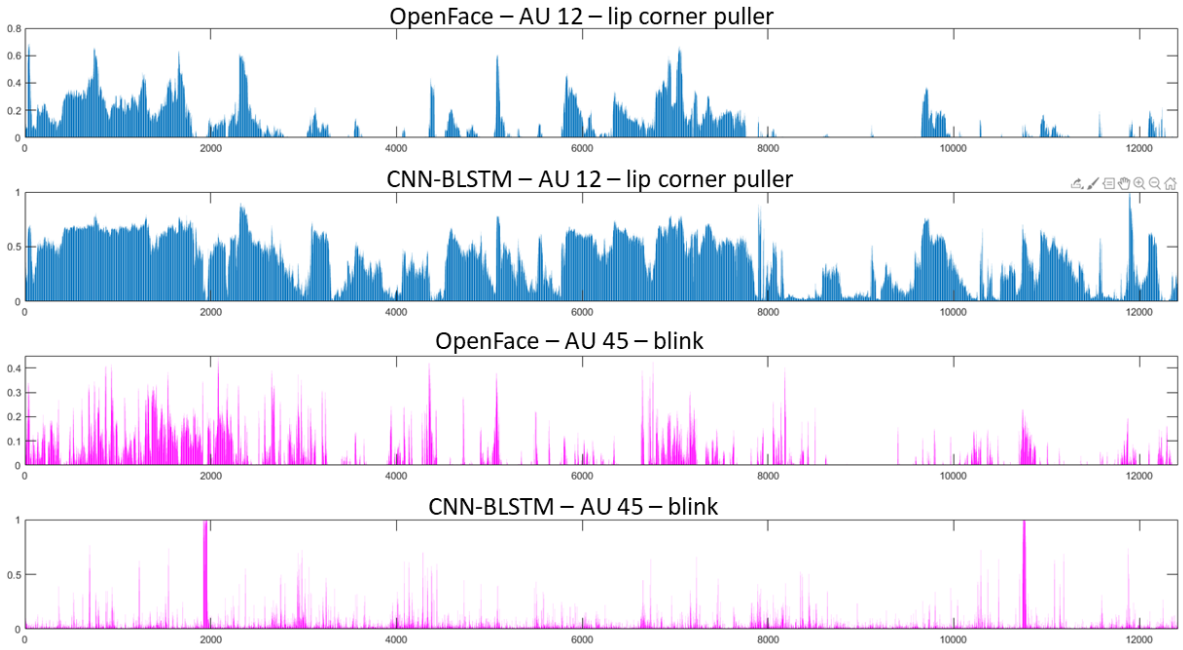

Figure 4.14: Comparison of a player's AU values. The top two bar graphs are of OpenFace's and CNN-BLSTM's detection of AU 12, an AU for which they have a strong correlation. The bottom two bar graphs are of AU 45 for which they have a low correlation. The bar graphs are of a single, randomly chosen player.

In addition, the observed correlation could be due to non-behaviour related events. Often, when the subject turns their head, or makes sudden large movements, the detectors lose track of their object. Figure 4.15, panels A and B, shows that the detectors react differently in these situations. I thus conclude that the observed correlations are not due to such non-behaviour related events.

I also examined the basic statistics for the two detectors over all the data in the videos. This was done in order to see how similar they were and also whether OpenFace's normalization process resulted in too much information loss, or conversely, if there was too much person-based bias in the CNN-BLSTM detectors. The statistics were computed as follows: for *min*, the minimum value for each player was computed and then these values plotted for all players. To calculate *1stQ* for each of the 64 players, the value of the first quartile was calculated and the 64 values thus obtained plotted. The same method was applied to the other eight statistics: *median*, *3rd quartile*, *maximum*, *interquartile range*, *mean*, *mode*, *variance* and *standard deviation*. The results for AU5 and AU12 are shown as box plots in Figures 4.16 and 4.17, respectively. For AU5 (upper lid raiser), which has a low correlation coefficient between the two detectors, a large difference between the statistics of OpenFace and CNN-BLSTM is found. Open-

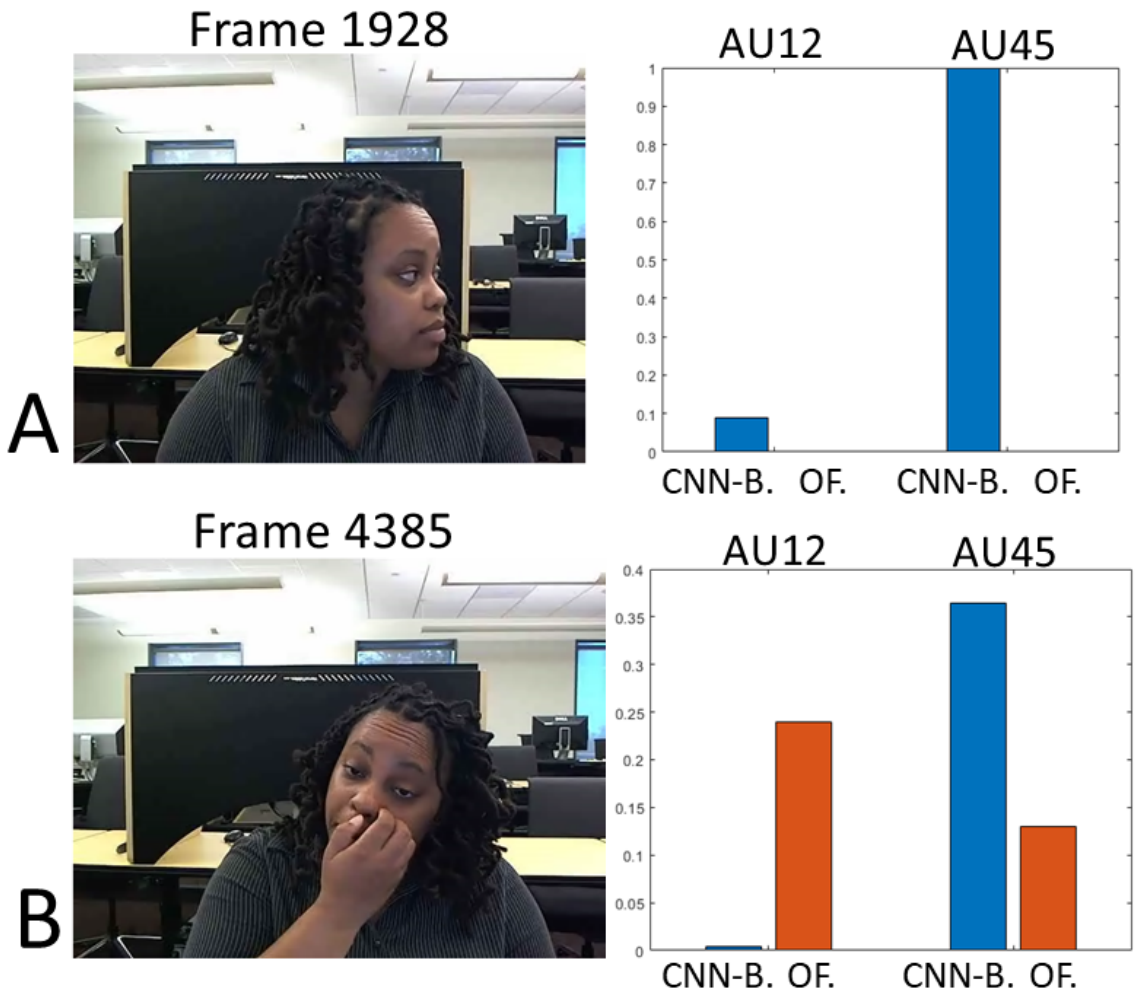

Figure 4.15: Detector responses to highly non-frontal views (A) and facial occlusions (B). OpenFace (OF.) and CNN-BLSTM (CNN-B.) detectors are thrown off track by the participant looking to the side as in A, or putting their hand to their face, as in B. The detectors give different results in these cases. In A, CNN-BLSTM gives a low value to AU12 (lip corner puller) and the maximum value 1 to AU45 (blink), while OpenFace assigns them both a value of zero. In B, the values output by the two detectors also do not agree.

Face shows very little activity with all statistics being near zero except for the *maximum*, while CNN-BLSTM has a much larger spread and AU5 takes on a wider range of values. The flip side is that the *median* and the *mode* also take on so many values, which may indicate a strong influence of person-specific bias. In contrast, for AU12 (lip corner puller), an action unit for which the two detectors have high correlation, the two detectors also demonstrate highly similar statistics. For box plots comparing the other ten action units, see Appendix A. The two AU pairs shown here are typical of the types of contrasts between the statistics of CNN-BLSTM and OpenFace.

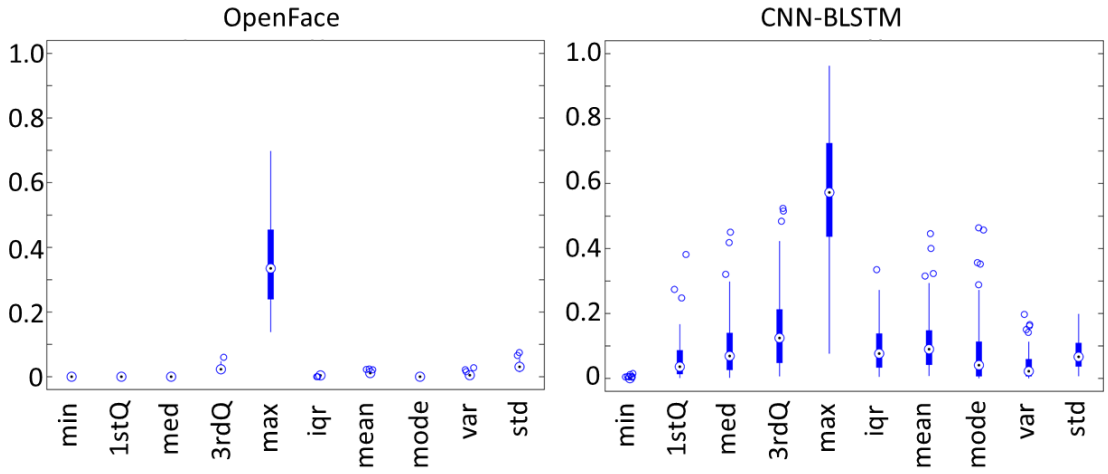

Figure 4.16: For AU5 (upper lid raiser) a box plot was made of the min, minimum; 1stQ, first quartile; med, median; 3rdQ, third quartile; max, maximum; iqr, interquartile range; mean; mode; var, variance and std, standard deviation values for each of the 64 players.

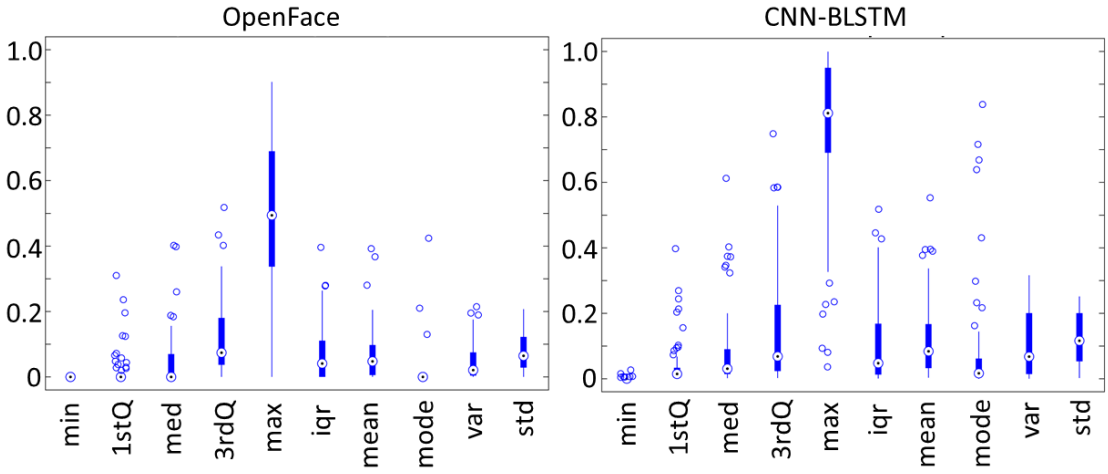

Figure 4.17: For AU12 (lip corner puller) a box plot was made of the min, minimum; 1stQ, first quartile; med, median; 3rdQ, third quartile; max, maximum; iqr, interquartile range; mean; mode; var, variance and std, standard deviation values for each of the 64 players.

From these comparisons, it seems clear that both CNN-BLSTM and OpenFace detect well on AU12 and that their correlation is not due to subject bias alone or their behaving similarly when the detectors get off track. It also appears that OpenFace might be less sensitive to low-level facial expressions when compared to CNN-BLSTM, as some of OpenFace's statistics seem too low, such as those shown in Figure 4.17 for AU12. On the other hand, OpenFace seems to suffer less from person-specific bias as there are fewer outliers in its statistical values for the mean and median, values which often give away a person specific bias. The experiments in this section shed more light on how the different detectors work. They also indicate that, at least for AU6, AU12 and AU25, CNN-BLSTM is picking up a meaningful and reliable signal representative of human behaviour. Since CNN-BLSTM was picking up human behaviour it made sense to continue trying to improve the classification rates of the decision trees. As a next step towards this, I implemented feature selection.

4.9 Searching for better decision trees by implementing feature selection

That CNN-BLSTM and OpenFace correlate so well over at least some AUs was encouraging. It is also counter-intuitive to expect that all action units should be equally important when classifying facial expressions. Therefore I decided to look again at building decision trees, this time using feature selection, which would help to weed out action units that only contribute noise. Knocking out irrelevant action units prevents them from interfering with classification. Although decision trees are considered good at ignoring irrelevant attributes, this might not be so in the case of complex and subtle facial expressions, limited data and possible noise. Therefore, removing the unwanted attributes altogether is a more absolute way of avoiding them. The idea of feature selection is to create classifiers based only on subsets of the available attributes and find those that classify best. In this case, the attributes are the values for the 12 different AUs. It is too time consuming to build decision trees for each subset of the 12 attributes as there are 2^{12} such subsets, so feature selection was done in the usual heuristic and greedy way. Here, I began with the empty set and added each attribute in turn to see which led to the best classifiers. From among these, the best were taken and in the next iteration, these sets of attributes were extended by one attribute to get 2-tuples. Those leading the best

classifiers were kept and out of them 3-tuples were built. This process was iterated until the 12-tuple containing all attributes was reached. I therefore did not require that the new set of tuples necessarily improved upon the old set, see Algorithm 2, but took the best ones even if they did not lead to an improvement and continued to the next set. For the following, I restricted the time window to be between two seconds before the FCR-event and one second after, as this was large enough and had a much better runtime than if I had used a larger window and thus calculated trees for many more different offsets.

```
Data: AU values for videos, list of 12 AUs (features)
Result: set of best subsets of features
initialization;
current_tuples ← all (12 2) two element subsets of 12 AUs (2-tuples);
while current_tuples is not the tuple containing all 12 AUs do
    for t ∈ current_tuples do
        calculate set of decision trees over offset/duration using only attributes in t;
        mark t with best b_classification rate over all offset/durations;
    end
    next_tuples ← tuples t ∈ current_tuples with best b_classification rate;
    best_tuples ← best_tuples ∪ next_tuples;
    current_tuples ← {};
    for ∀t ∈ next_tuples do
        for ∀a ∈ AU, a ∉ t do
            current_tuples ← current_tuples ∪ (t,a);
        end
    end
end
return best_tuples;
```

Algorithm 2: Feature selection pseudocode. This is a simple variation on the Feature Selection heuristic that is frequently used in the literature, for instance Bartlett et al. (2014).

The best classification with feature selection was found at tuple (AU4, AU5, AU9, AU12, AU15, AU25, AU45) with offset -30 and duration four. The balanced classification rate was 0.6014. The voting scheme in Algorthm 1 was applied. This resulted in a balanced classification rate of 0.6365.

	Regular	Balanced
Classification rate	(0.6840) 0.7152	(0.6079) 0.6365
Precision	(0.4265) 0.4803	(0.6628) 0.7096
Recall	(0.4394) 0.4621	(0.4394) 0.4621
F1-measure	(0.4328) 0.4710	(0.5282) 0.5597

Table 4.9: Performance values for the best decision tree with voting after feature selection was performed. Values for the previous best classifier, built over all attributes, see Table 4.7, are shown in red for comparison.

Once again, all eight values for measuring performance were improved, as can be seen in the comparison shown in Table 4.9.

There were altogether 82 feature selection tuples picked out by the feature selection algorithm. Voting was done on all of these. For the best offset/duration of these tuples, voting improved the balanced classification rate for 66 of the 82 trees (80%). Voting improved the real classification rate of 80 of these trees (98%). If one chooses the offset/duration parameters that maximize the balanced classification for the trees with voting, then the best offset/duration parameters are the same ones chosen for trees without voting applied for 59 of the 82 tuples (72%). Therefore, voting generally improves results over trees without voting. Detailed results for 82 tuples selected by feature selection are given in Appendix A.

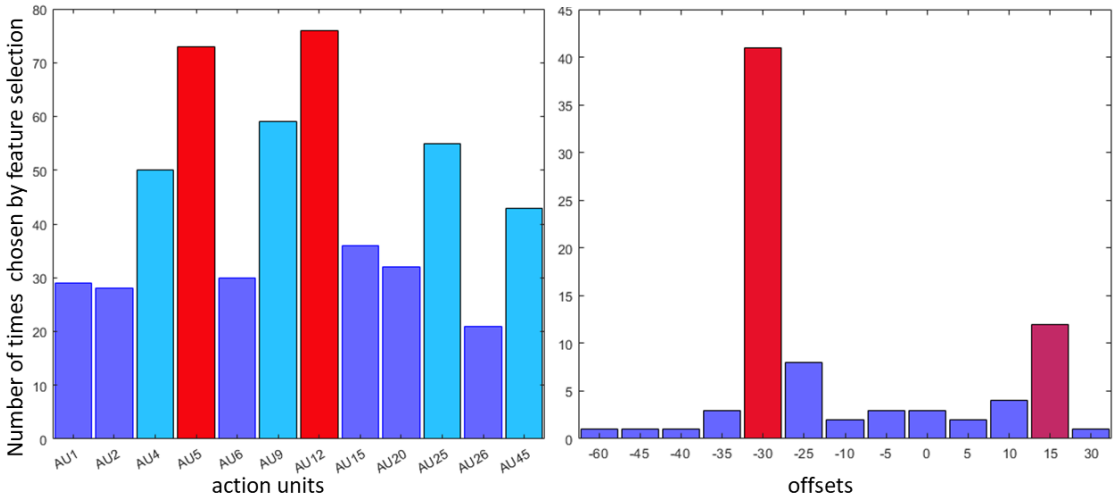

Figure 4.18: The left panel shows the frequency with which action units were chosen by feature selection. Red (high frequency AUs, AU5 and AU12); light blue (medium frequency AUs); dark blue (low frequency AUs). The right panel shows the frequency with which offsets produced best classifiers. Red (high frequency offsets, -30 and 15); dark blue (low frequency offsets).

To gain further understanding of the videos in the database, I looked at the features

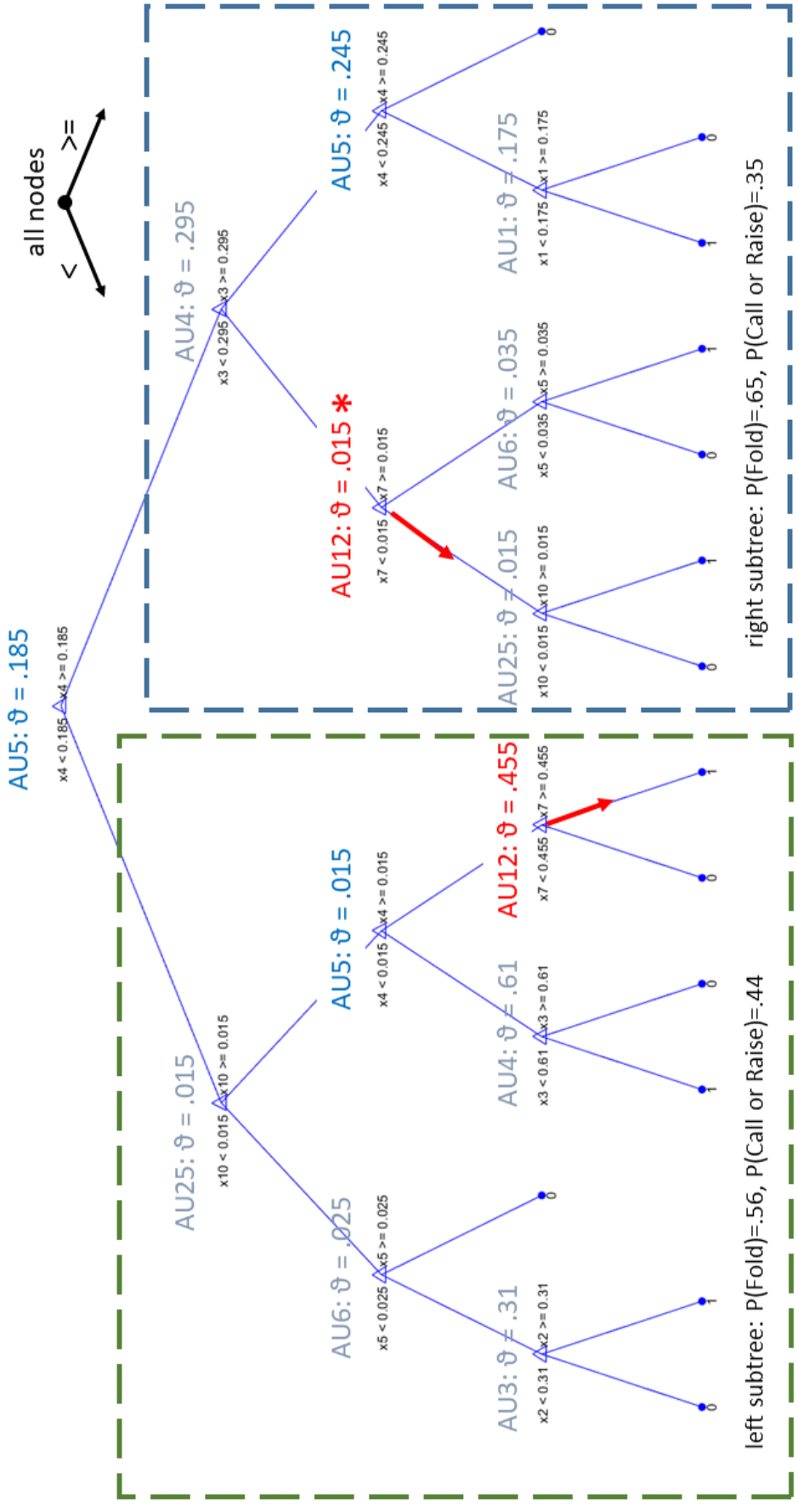

Figure 4.19: This figure shows $T_{15,3}$, which was computed in Section 4.6 over all 12 AUs computed by CNN-BLSTM. Here, its depth was restricted to give a better overview of the most important features of the tree. AUs used for splitting, as well as their selected thresholds, θ, are shown at each node. The right branch at each split represents the $\geq$ branch while the left is the $<$ branch. As they were significant in the statistical tests, AU12 is in red type and AU5 is in blue type. The red arrows at AU12 splits indicate the direction to follow to increase certainty of a fold having occurred. The asterix at split AU12, $\theta = .015$ indicates the only split that uses AU12 where lower AU12 values increase the certainty of a fold.

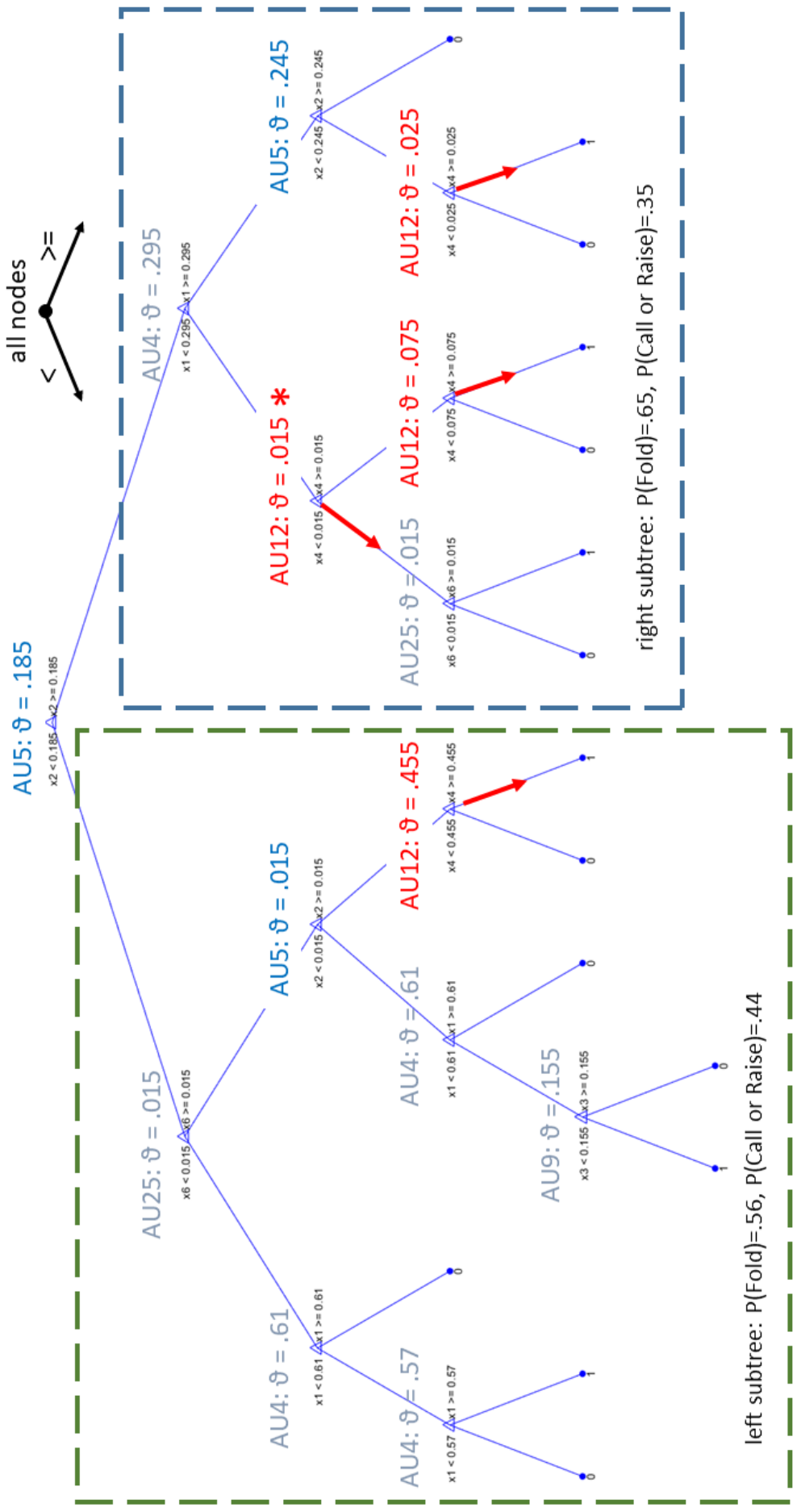

Figure 4.20: This figure shows $T_{-30,4}$, which was the best classifier returned by feature selection. The selected features were AU4, AU5, AU9, AU12, AU15, AU25 and AU45. As in Figure 4.19, the depth of this tree was restricted. Also as above, AUs used for splitting, as well as their selected thresholds, θ, are shown at each node. The red arrows at AU12 splits indicate the direction to follow to increase certainty of a fold having occurred. The asterix at split AU12, $\theta = .015$ indicates the only split that uses AU12 where lower AU12 values increase the certainty of a fold.

selected for the best classifiers by the feature selection search. The bar graph in Figure 4.18, left, was made to illustrate how many times each feature was selected. AU5 and AU12 were selected with very high frequency and much more frequently than the other features, suggesting they are important. These are also the two action units which were significantly higher for folds versus calls and raises in the paired t-test done in Section 4.5, page 64. A bar graph was also made of the most frequent offsets chosen, Figure 4.18, right. The largest spike occurs at an offset of -30, with a second, smaller spike occurring at an offset of 15 frames, suggesting that shared facial expressions are occurring at these times. Of note, Section 4.5 also shows the statistical significance of AU12 and AU5 as occurring at -30 frames, or one second before the FCR-event. The tuple chosen by feature selection was located at -30 and used both AU5 and AU12, implying once again that people who fold have higher levels of AU12 (lip corner puller) and AU5 (upper lid raiser).

To facilitate an understanding of how the best tree returned by feature selection works, it is visualized in Figure 4.20. The parameters for this tree were offset -30, duration 4 and AUs 4, 5, 9, 12, 15, 25 and 45. This figure has been placed next to Figure 4.19, which was returned as the optimal tree in Section 4.6 where all 12 AUs were used and only the influence of different offset and duration parameters were investigated. Even after feature selection is performed, the first split in the tree is the same for both trees, and occurs at AU5 with a threshold of .185. Despite the removal of 5 AUs during feature selection, the topmost three levels of both trees are identical regarding what AU is chosen and what its threshold is. The only exception is the leftmost split on the third level where, in Figure 4.19, the split is on AU6 with $\theta = .025$, as opposed to Figure 4.20, where the split is now on AU4 with $\theta = .61$. It is only on the fourth level where the trees begin to differ as AU3, AU6 and AU1 were removed by feature selection. One occurrence of these has been replaced by AU4, the other three have been replaced by AU12. AU5 plays a large role in both trees and appears therefore to be important and stable. Also, in the trees, as in the paired t-tests, higher AU5 values are associated with folding. This is because the split at the root produces a subtree, the right subtree, in which there is a higher probability, .65, of an event chosen randomly from this subtree being a fold. While the role of AU5 as being important for distinguishing fold from call or raise seems to be upheld as it remains stable across the search, it is not so easy to understand the role of AU12. It becomes much more prominent after feature selection was applied, implying it is not as strong as AU5 and is confounded with noise more easily. The red arrows in the

trees in both Figure 4.19 and Figure 4.20 indicate the branch to follow from the split to increase certainty of a fold occurring in the subsequent subtree. All splits on AU12 but the one marked by an asterix point to larger AU12 values indicating folding, which is also corroborative of the paired t-test results. Therefore, the trees can also be seen to possibly indicate that higher AU5 and AU12 values are associated with folding.

4.10 Comparing human performance to the performance of the classifier

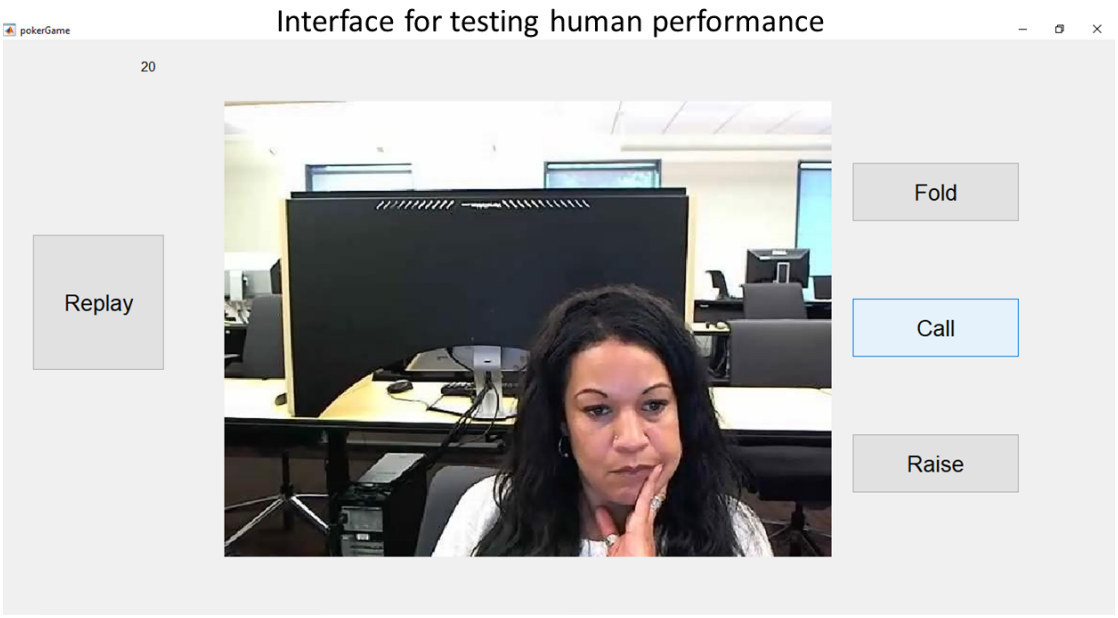

Figure 4.21: Example of the interface used to test how well humans can distinguish between folds and call/raises in the videos of the poker dataset.

After having developed a classifier, it can be useful to compare it to human performance, as was done by Bartlett et al. (2014). This can provide insight into how well an automatic detection method works and also whether there is a detectable ground truth. For this reason, I created an experiment to test human performance at distinguishing folds from calls and raises. The interface for this experiment was built by myself in MATLAB. The experiment consists of, for each study subject, randomly selecting ten of the 132 videos of a player folding, five of the 184 videos of players calling and five of the 165 videos of their raising. These 20 videos are then shuffled before being presented to the test subject by means of the interface shown in Figure 4.21. They are then asked to judge for each short video segment (four seconds ending in a FCR-event) whether the poker player is

folding, calling or raising in the video. They can replay each video as many times as they want before making their decision and moving on to the next video. To evaluate the subject's performance in a way similar to the classifiers, they were not penalized for confusing calls with raises, but only for failing to distinguish folds from calls or raises. Funds were not available to carry out this experiment, so I did this informally by asking friends and co-workers if they would participate. Although data collection was going well, Covid-19 lockdown began shortly after I began running the experiment and I only managed to get 14 subjects. The average performance of the 14 test subjects was .52, with .50 being chance. This was not significantly different from chance, but the number of participants is too low to reach a final conclusion. This trend suggests, however, that the classifiers perform better than humans at detecting folds in this dataset.

4.11 Concluding remarks

I have searched a space of decision trees and attributes to find areas that lead to good classification of facial expressions. The main assumption made here was to focus on the event, the decision made per mouse click, as this is when the participant is most likely to be considering their decision. The search space was thus focussed tightly around the events. Originally, the search space encompassed nine seconds before the event and five seconds after. Decision trees were built over individual frame values and not statistics or aggregates over the values. This was to avoid losing small or fleeting signals. I next added a voting method to combine the classifications of the individual frames associated with an event into a single classification for the event as a whole. Aggregating the outputs of the decision trees by a simple voting method improved the classification rates. After this, to improve the classification rate further I performed feature selection on the AUs. To cut down on the runtime during feature selection, which is very repetitive, I restricted the search space to between two seconds before the event to one second after the event. Feature selection resulted in better classification with fewer attributes. These results were also improved by applying voting, including for the tree with the best set of features returned by feature selection. The criterion used for driving the search throughout was improving the balanced classification rate.

A comparison was made between CNN-BLSTM, detectors recently developed at the University of Nottingham, and OpenFace, which are established detectors and frequently

used in the literature. In the best case, the two detectors would have agreed. Statistical tests seemed to indicate that CNN-BLSTM was better suited for this dataset therefore I used these for constructing decision trees. After doing an initial search for a decision tree classifier, I looked to see how well CNN-BLSTM and OpenFace correlated over the dataset. Comparing the two detectors gave strong evidence that they are both good at detecting AU6, cheek raiser, AU12, lip corner puller, and AU25, lips part. The behaviour of the detectors on some of the other AUs differed strongly, however. This second comparison of CNN-BLSTM with OpenFace was mainly done to confirm that CNN-BLSTM was really detecting meaningful behaviour. For future reference, comparing the basic statistics of the two detectors showed that OpenFace may not be as sensitive to low levels of expression as the CNN-BLSTM detectors. That the OpenFace detectors might not work well with low levels of expression or when the neutral face is not the most frequently occurring face, is something that the authors have pointed out (Baltrušaitis et al., 2018). Also, as I found while performing these experiments, the OpenFace detectors are about 30 times faster than CNN-BLSTM (it took OpenFace two minutes to run on a video, but one hour for CNN-BLSTM). Some of the sensitivity of OpenFace might have been sacrificed for speed.

The results shown here on the poker dataset indicate that combining decision trees with a static representation of the face as computed by AU detectors provides a plausible way of finding associations between facial expressions and actions in the game of poker. There was consistent support throughout these experiments for the idea that players exhibit more AU12 (lip corner puller) and AU5 (upper lid raiser) when folding than when calling or raising. These two AUs stood out in the paired t-test analysis, in the feature selection process, and figured prominently in the construction of the best trees, which were shown in Figures 4.19 and 4.20. There is therefore evidence that higher levels of AU5 and AU12 are associated with a person's intention to fold.

Chapter 5

A Virtual Dice Rolling Experiment Reveals that Gender and Stress Modulate Deception

5.1 Using the poker database design as a springboard for a new experiment

The poker database was a good preparation for designing my own study of deception. Many of the ideas behind the design of the poker study were transferable to a new setting and were therefore important to consider. First, an interface needed to be created through which participants interacted with the experiment and the flow of the experiment needed to be controlled by the computer. It should be designed in such a way that there were many meaningful timestamped events in order to later segment the data into events of interest. For this reason, important computer generated events and also events generated by participant decisions needed to be annotated and timestamped to establish a descriptive and accurate ground truth, much as was done in the poker data set. While the interaction was being moderated and recorded by computer, a frontal view video of each participant had to be made. It should be as good quality as the poker videos and also have timestamped frames to allow for cross reference between frames and events for correct segmentation. Given these similarities, I was free to create a new deception scenario. My intent was to create a more antagonistic experiment than poker that might get

closer to the idea of malintent by creating a situation in which it is not socially acceptable to deceive.

This study in deception was designed together with behavioural economists Professor Roberto Hernán-González, now at Burgundy School of Business, Dijon, and Professor Thorsten Chmura, now at Nottingham Trent University (Corgnet et al., 2016; Chmura et al., 2017). It approaches an important problem in behavioural economics with computer vision and machine learning such that both schools of thought can benefit. To my knowledge this is the first such study. There is a large overlap of interests between these two fields and the study was designed in such a way that, seen from either perspective, it should be valid and complete; it is an investigation into human decision making while simultaneously allowing high quality digital data to be collected for computer analysis and research.

5.2 The investigation

This study investigates the role of stress on decision making, in particular the role of stress on human lying behaviour. Stress is considered to interfere with a person's cognitive abilities. It has been shown that stressed humans confronted with a task rely more on habit to solve the task than their unstressed counterparts, who react instead in a goal-directed way (Schwabe and Wolf, 2009; Valentin et al., 2007; Dickinson, 1985). Lying is considered to be cognitively difficult (Zuckerman et al., 1981; Vrij et al., 2008). We therefore hypothesize that stress alters the way people lie, both quantitatively, that is, how much they lie, and qualitatively, that is, their physical behaviour when they lie. It follows that we should be able to detect these both by recording their decisions and behaviours, including facial expressions, while they are making their decisions.

According to standard economic theory, a person will always lie if it will lead to material gain and there is no fear of punishment (Lewicki, 1984). These means that things like moral behaviour, for instance, must be explained in terms of fear of punishment or reprisal. In fact, the first to note this selfish behaviour was John Stuart Mill (1806-1873) who coined the term $homo\ economicus$ to describe this strictly rational wealth-maximizing human (Persky, 1995). However, this seems to be an oversimplification. A recent investigation found, contrary to expectations, that even when there is a profit to be

made and no fear of punishment, people do not lie maximally (Fischbacher and Föllmi-Heusi, 2013). In their experiments, participants privately rolled a die in such a way that no one could know what they rolled. They then reported what they rolled to receive a monetary reward. The size of the reward depended on what they rolled so they had an incentive to lie. Simultaneously, since no one could know what they rolled, as it was a blind experiment, they had no fear of punishment. Nonetheless, the researchers found that the study subjects did not fully lie to maximize their rewards. After testing different controls, they concluded that lying behaviour is robust and they formulated three characteristics that define lying behaviour in their die rolling experiment:

1. There is a positive number of truly honest people who will report a pay-off of zero if that is what they rolled.
2. There is a positive number of people who lie maximally and report the value with the largest pay-off, even if they didn't roll it.
3. There is a positive number of people who lie partially, that is, they lie but not to the fullest and report a higher but not maximal pay-off.

Fischbacher and Föllmi-Heusi could explain point (1) by considering these participants to have a preference for truth-telling and an aversion to lying, and (2) by considering these participants to be cases of homo economicus. However, (3) seemed more elusive. Perhaps these subjects did not lie maximally because they did not want to appear greedy, or perhaps they were trying to disguise their lies. Because of its simplicity and powerful hypothesis-testing possibilities, the Fischbacher and Föllmi-Heusi die rolling experiment has become a paradigm for studying deceptive behaviour.

The study presented here follows their work, albeit with important modifications and extensions. Here, participants virtually roll a die a sequence of times and report what number they rolled. For each roll, they get a monetary reward that depends on the value they report. Note that the maximum reward in this study is similar to the reward offered in Fischbacher and Föllmi-Heusi (2013). Thus, participants likewise had an incentive to lie as the different faces of the die were worth different amounts. While this was happening they were being recorded, a fact that they were told. As another major extension to the Fischbacher and Föllmi-Heusi experiment, here the participants were subjected to physiological stress. For this, the participants had to submerge their hand in ice cold

water (1–3°) directly before carrying out the die rolling experiment. For the control, the participants went through the exact same procedure except that they put their hand in body-temperature water instead of ice water. The details of the experiment are given in the next section.

5.3 Experimental design

5.3.1 Participant recruitment

Participants were University of Nottingham students who volunteered to participate in experiments of the Centre for Decision Research and Experimental Economics (CeDEx) laboratory for studying human decision making in the School of Economics. Their motivation was to contribute to science and also receive a monetary payment. They were assured that data protection laws would be adhered to and that their decisions and any other data kept on them would be only used in anonymized and aggregate form. CeDEx keeps a database of volunteers and randomly chooses people from it for experiments. If a person is chosen, they are sent an email inviting them to a session. If they accept, an appointment is made for them and they are informed that they can withdraw at any time. Very little is said about the nature of the experimental scenario before they show up for their session. The ethics approval to perform this experiment was granted by the UoN's School of Economics.

5.3.2 Facilities

These experiments took place at the Cribs Laboratory for Experimental Behavioural Economics, which is part of CeDEx and is located in the Yang Fujia building on Jubilee Campus at the University of Nottingham, UK. The Cribs lab consists of an anteroom, where a network controller is kept to manage the computers in the adjoining lab. The lab contains 40 identical cubicles, see Figure 5.1, left, numbered 1 through 40, which are each equipped with a desk, a chair and a personal computer (PC). All 40 computers are connected to a network which can be controlled from the network controller in the anteroom. As such, the network can be utilized to run software and collect data. In addition to this standard set up, webcams were mounted on the monitors of the PCs. The room is

illuminated by eight pairs of bright fluorescent lights. I put a layer of white tissue paper over them as an inexpensive way to make the light more diffuse to reduce interference with the AU detection. Only those cubicles directly under the lights were used for this experiment to make the lighting as consistent as possible across all participants. A row of chairs was put at the front of the room, see Figure 5.1, right, for the participants to sit in when they did the water treatments. These were numbered the same as the PCs, so each participant had a fixed seat (the one with the same number as their PC) for the water treatment. The chair to the left of each seat held a bucket of water for the participant. It was also given the same number to avoid any confusion. Thus, each participant had a pair of chairs. Due to space limitations, only up to ten participants could take part in a given session. Figure 5.2 shows a map of the lab setup.

Figure 5.1: The Cribs Lab. Left, the participants sat in their own cubicles while doing the dice rolling experiment. Right, view from the anteroom. A row of chairs was placed along a wall of the lab for participants to sit on for the water treatments. Note the black water buckets on every other chair.

5.3.3 Stress treatment

The main hypothesis of this experiment is that stress will alter a person's lying behaviour. To induce stress in participants before they carried out the die rolling part of the experiment, a version of the the Socially Evaluated Cold Pressor Test (SECPT) described in Schwabe and Schächinger (2018) was used, which is similar to the cold pressor test (Hines and Brown, 1932) and is a simplified version of the Trier Social Stress Test (Kirschbaum et al., 1993). The participants submerged their hands, including the wrist, into cold water and held it there for three minutes. For this experiment, the temperature was kept between 1-3° C. A researcher was in the room and controlled that the partici-

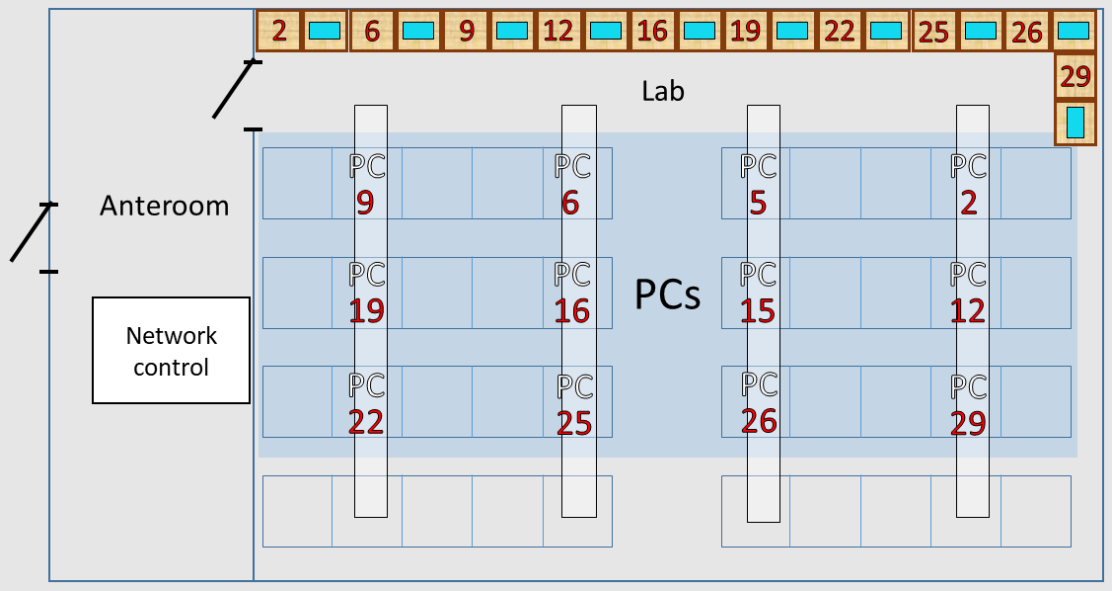

Figure 5.2: Map of the Cribs Lab. The chairs used for the water treatment are shown as red rectangles, water buckets as solid blue rectangles. Only the indicated PCs were used as these were all located under the four rows of lights (white rectangles). Cubicles are indicated by blue rectangles.

pants kept their hands in the water. When participants tried to take their hand out of the water due to the discomfort, which happened occasionally, they were asked to try and hold it in or return it to the water as soon as they could. The procedure for the control group was the same, except that the temperature of the water was kept at 37-39° C. This temperature range was chosen, as it is close to body temperature and should feel comfortable, whereas water at room temperature can be perceived as cold and uncomfortable. Since the effects of the treatment might be small, I chose to stick strictly to these temperatures with as little variation as possible. For details on how the temperatures were kept constant, see Appendix B.1.

5.3.4 Die experiment software

The die rolling experiment was introduced by Fischbacher and Fölmi-Heusli as a way to study honesty and deception. In their work, a real die was used by the participant. The participant received a monetary reward in Swiss francs (1 CHF $\sim$ 0.84 GBP) based on what they reported as follows: 1 dot = 1 CHF, 2 dots = 2 CHF, 3 dots = 3 CHF, 4 dots = 4 CHF, 5 dots = 5 CHF and 6 dots = 0 CHF. Here, a similar reward system is used in Sterling, see Table 5.1. In the Fischbacher and Fölmi-Heusli experiment, there was no way for the experimenter to know what the participants had actually rolled. Instead, the

researchers could only estimate from the cumulative results how much lying took place. In this case, they knew there was lying because the distribution of reports did not match the uniform distribution as would be expected from rolling a fair die. The experiment presented here, the software for which I built using MATLAB, is conceptually similar except that a video of a rolling die is used and also reports are made using the computer interface such that the ground truth is known, making it impossible to disguise a lie.

Die face	Reward	Die face	Reward	Die face	Reward
1 dot	£0.05	3 dots	£0.15	5 dots	£0.25
2 dots	£0.10	4 dots	£0.20	6 dots	£0.00

Table 5.1: Rewards associated with each face of the die. This follows the schema in Fischbacher and Föllmi-Heusi (2013), with the maximum cumulative reward over 20 rounds being comparable.

Due to the nature of the experiment, the software had to fulfil two main functions: The first function, the *game module*, needed to present and moderate the experiment to each participant while recording and timestamping the participant's decisions made by mouse clicks and dice roll outcomes. The second function, the *video module*, had to make a timestamped recording of the participant's face via the webcam mounted on the monitor while the experiment was in progress. The mouse click events of the game module interfered with the video recording in the video module causing gaps and delays in the videos, so these two modules had to be built separately and run in two separate MATLAB instances to parallelize them. The two modules communicated with each other and coordinated their activities by writing and reading into a common directory. See Figure 5.3 for an overview of this system.

The video module

The video module consisted of a loop that waited for the game module to trigger it to begin recording. Once triggered, it then moved into the recording phase which was a loop that created a video one frame at a time together with the timestamp for that frame. This loop repeated itself adding a frame and a timestamp with each iteration until the game module informed it by means of a shared directory to stop recording. When this

signal came, the video module quit recording and stored the video in an AVI (Audio Video Interleave) file and the timestamps, given in epoch time, in an array in a .MAT file. The number of frames in the video was the same as the number of timestamps and the ith timestamp was the timestamp for the ith frame. For examples of videos made during the experiments, see Figure 5.6, in Section 5.4.

Output of the video module.

- A video of the participant taken by a Logitech HD 1080p at a resolution od 640X480 and an achieved frame rate of 30 frames per second stored as an AVI file.
- A .MAT file containing the array of timestamps for the video.

The game module

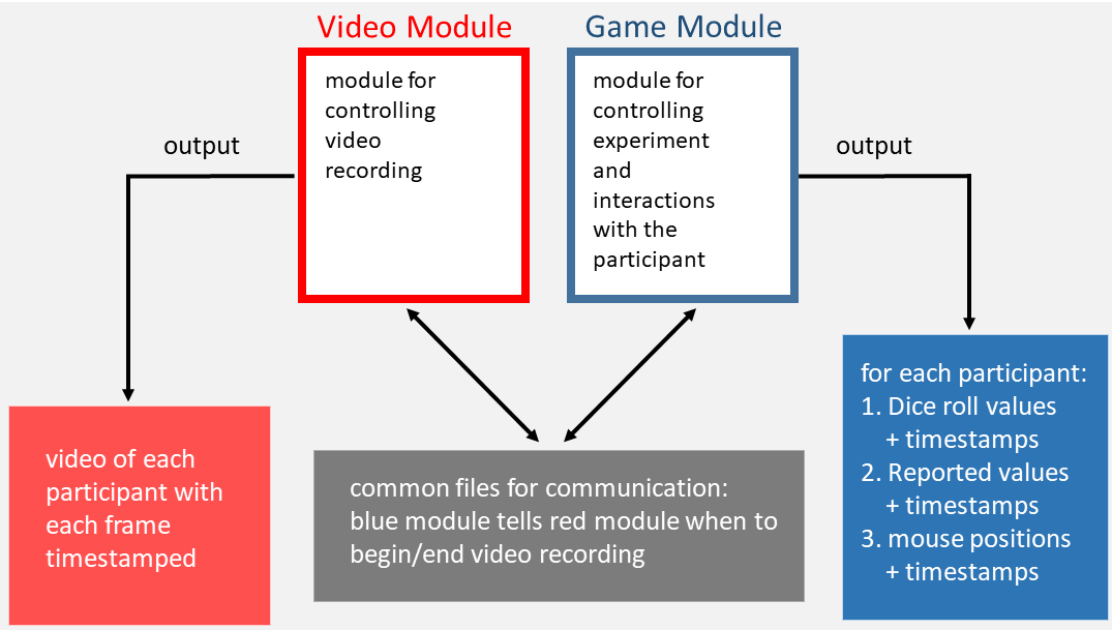

Figure 5.3: Flow chart of dice rolling experiment software. There are two modules, the video module and the game module, which run in parallel and communicate through writing and reading from a common directory.

After the participant returned from their water treatment, the game module greeted them with a screen that prompted them to begin the experiment when ready by pressing a 'start' button. When this button was pressed, the game module informed the video module to begin recording and showed the participant a screen that asked them to roll a die by pressing a 'roll die' button. This button press was timestamped. The computer then

selected and played a random video of a die being rolled. The number rolled in this video was annotated and the end of the video was timestamped. The participant was then prompted to report the outcome of their roll by pressing a report button. The timestamp for this button press, along with what was reported, were also recorded. As in the video module, timestamps were in epoch time. All values were stored as arrays in .MAT files.

Instead of a single roll as in the Fischbacher and Fölmi-Heusli study, in my experiment the participant rolled the die 20 times, each time reporting what they got by pressing buttons in the computer interface. Therefore, the above procedure was repeated 20 times, with the rolling of the real die being replaced by virtual die rolling. This sequence of interfaces is shown in Figure 5.4. For more detailed images of the interface, see Appendix B.2. After the twentieth and last role, the game module messaged the video module to stop recording, informed the participant what their cumulative reward was and directed them to fill out the questionnaire (on paper), which had been placed in the upper left hand corner of their desk.

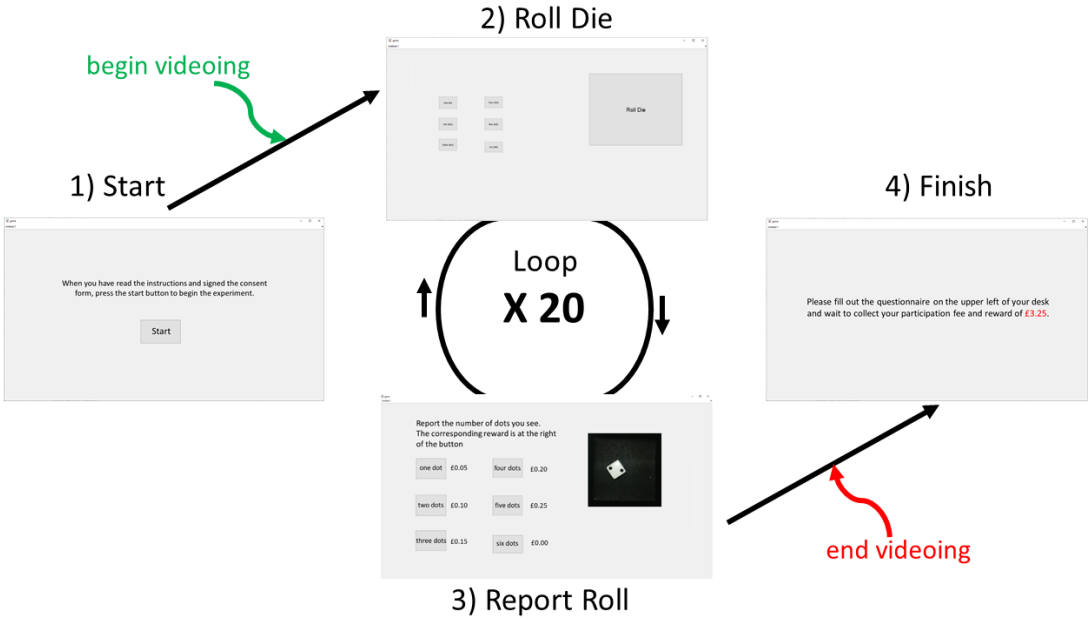

Figure 5.4: Game module. The flow of the interface with which the participant interacted. The first screen was 1) Start. After pressing 'start', the video module was informed to begin videoing and the next screen was 2) the roll die Screen. After pressing the 'roll die' button, the randomly chosen video was shown and the participant prompted to report what they rolled on screen 3) Report Roll. After this was repeated 20 times, the experiment ended, the video module was informed to stop videoing, and the participant was taken to the final screen 4) Finish. Larger images of the interface are shown in Appendix B.2.

In order to control the flow of the program, button activation and deactivation were used.

The buttons for reporting what was rolled were deactivated when it was time to press the 'roll die' button. Upon pressing 'roll die' a random video was shown. The videos were all the same length and ended when the die came to a standstill. When the video ended, the report buttons were reactivated and the user prompted to report what they had rolled. The participants were thus not able to report on the outcome of the round until the die video came to an end. This was to prevent them from taking shortcuts to the rewards instead of making twenty separate decisions. Once the participant had selected their choice, the 'roll die' button appeared again as active along with the prompt to roll it and the report buttons were deactivated. This controlled the flow of the experiment and also helped keep the participants engaged as they activated many of the events themselves.

In addition to the data described above, the game module also collected mouse tracking data. This was not part of the original design of the experiment, but I added it because the opportunity presented itself and it did not negatively impact any of the other functions I had built, such as video acquisition. Analysis of the mouse tracking data will be part of a later study.

Output of the game module.

- A .MAT file containing
  - 20 roll die timestamps in epoch time
  - 20 values for the dice rolled (1–6)
  - 20 values reported by participant (1–6)
  - 20 timestamps for when the participant made their report, in epoch time
  - 20 timestamps marking the ends of the 20 videos, in epoch time
- A text file (.txt) containing the x- and y-coordinates of the mouse positions along with their timestamps in epoch time.

Die videos. Although it might not be important to the decision making process of the participant, effort was made to make the die rolling seem as realistic as possible to keep the participant engaged. This was done in the hope it might dampen their awareness of their being surveyed, which otherwise might discourage deceptive behaviour. Therefore, I recorded 120 separate videos of an actual die being rolled into a black box, so no video needed to be shown twice to a participant. Also, the participants activated the events

themselves to keep them engaged and paying attention, for instance, they had to press the 'roll die' button to get the die to roll.

5.3.5 Questionnaires

After the dice rolling part of the experiment, the participants completed a questionnaire, which, for practical reasons, was printed on paper. It consisted of the cognitive reflection test (CRT), a modified self-assessment manikin SAM, the MACH-IV test of Machiavellianism, and demographic questions. The CRT (Frederick, 2005) tests how well the participant overrides the urge to answer intuitively in order to correctly analyse a problem. SAM asks a person to rate their emotional state using graphical representations of the three fundamental emotional dimensions, pleasure, arousal and dominance (Bradley and Lang, 1994). As we were especially interested in discovering how stressed the participant felt, I invented a fourth manikin to represent stress, shown in Figure 5.5, and added it to the three standard SAM manikins. Of note, the SAM test was only added to the questionnaire half way through the experiment, so not every participant filled it out. The MACH-IV (Exline et al., 1970) tests to see to what extent the participant agrees with the ideas stated by Machiavelli, namely that one should focus on achieving one's ends without heed to morality or empathy for others. The final part of the questionnaire consisted of demographic questions. For the complete modified SAM and the demographic questions, see Appendix B.3.

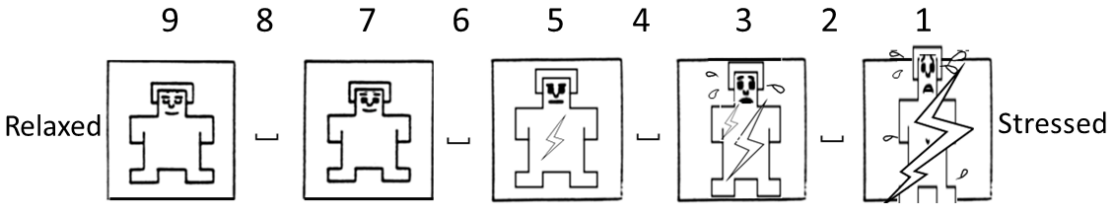

Figure 5.5: The fourth manikin for self-reporting stress. Numerical values associated with the conditions are located above.

5.3.6 Experimental protocol

An extended version with a detailed step-by-step description of procedures can be found in Appendix B.1.

Arrival

Upon arrival, participants waited outside the Cribs lab in the waiting area. There were two experimenters and at the designated time, Experimenter 1 opened the door, greeted them and checked their names and IDs against the list of participants enrolled for this session. Experimenter 1 let the participants into the anteroom, see Figure 5.2, where they were greeted by Experimenter 2, who was waiting at the door of the lab. Experimenter 2 asked them to draw a card from a cotton bag to ensure random seat assignment. The participants were then told to go to the cubicle, see Figure 5.1, left, with their number on it where they could read the consent form and the instructions, see Appendix B.4. The participant's cubicle contained a table with a PC, a keyboard and a webcam. There were also a pen and an A4 envelope containing the questionnaire.

Induction

When everyone had arrived, Experimenter 1 explained to the participants that they were there on a voluntary basis, that they would be recorded by the webcams mounted on their computers and that images produced were for research purposes only; however, if they were uncomfortable with this or anything else, they could leave at any time. The webcams were clearly visible on the tops of their monitors, and their lights indicated they were turned on, however, they were not yet recording at this time. Participants were then given a few minutes to read and sign the consent forms.

Water treatment

After everyone had signed their consent form, they were told by the experimenter to come to the front of the room and sit in the chair labelled with the same number as their PC. The chairs formed a U shape, so the participants could all see each other, see Figure 5.2. They were told they would need to submerge their hand, including their wrist, in the bucket in the 1–3° C cold (stress) or the 37–39° C warm (control) water for three minutes. Subsequently, after drying their hands, they were told to return to their cubicles. They were told by the experimenter to now follow the instructions on their computer screen and that if they needed any help, they should raise their hands. From then on, the experimenter remained at the door of the room looking in in case anyone

raised their hand with a question.

Die rolling

When the participants returned to their cubicles, the die game program had been maximized. The participants now interacted only with the computer. The computer program asked the participants to confirm they had signed the consent form and were ready to begin by clicking the 'start' button, which also started the video recording and mouse tracking. The experiment consisted of rolling a die by pressing a 'roll die' button followed by reporting what had been rolled. On the instruction sheet, see Appendix B.4, the participants had been told there was a reward associated with each number of dots and this was also shown on the computer interface. This was repeated 20 times in order to be able to consider the rolls as independent while increasing the amount of data and video we could collect. The die rolling part of the experiment usually took just over three minutes. Details of the computer interface are given in Appendix B.2.

Questionnaires and payment

At the end of the die rolling experiment, the screen prompted the participant to fill out the questionnaire on their table. When the questionnaires were filled out, Experimenter 1 called them one-by-one to get their payment (£3.50 participation fee + reward) in the anteroom. Experimenter 2 handed over payments in a sealed envelope. After a participant received their payment and left, the next participant was called.

5.4 Outcome of the experiment

The experiment took place in 41 sessions spanning the time period of September 2018 to July 2019. I organized and carried out each session and was either Experimenter 1 or Experimenter 2. Altogether there were 384 participants of which 11 did not complete the experiments for various reasons and were therefore removed from the study. This leaves 373 participants who completed this study. A video was made for each of these as they carried out the dice rolling part of the experiment. Frames from three randomly chosen videos are shown in Figure 5.6.

Figure 5.6: Images from videos of three different participants. The participants' faces have been blurred and their eyes covered to conceal their identities.

A total of 16 hours of videos were made at 30 fps. Collectively, these videos consist of 1,716,711 timestamped frames at a resolution of 640×480. In addition to each video, there are timestamps for when each participant rolled the die and reported what they rolled. Furthermore, the mouse movements were recorded and timestamped over exactly the same amount of time.

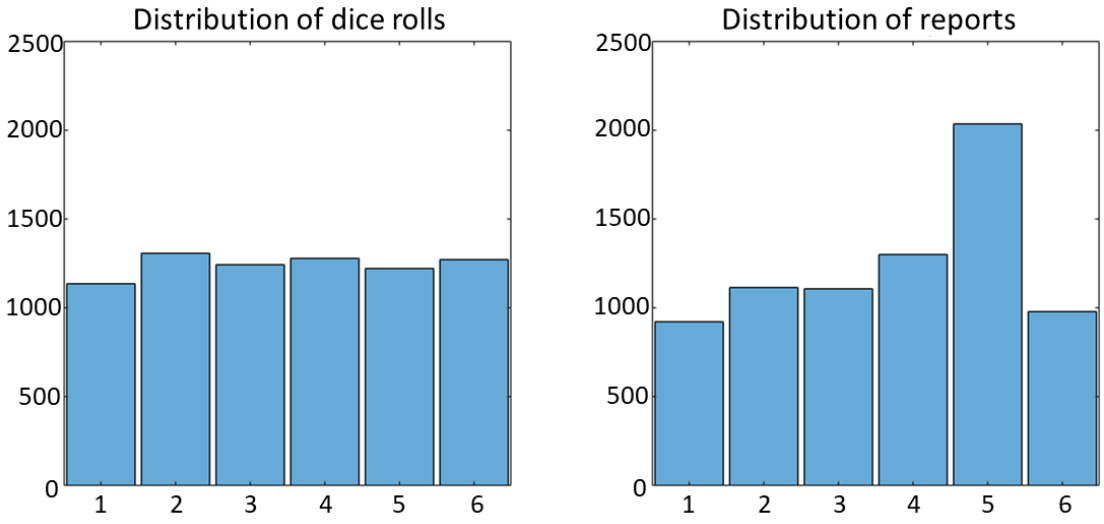

Figure 5.7: Distribution of rolls (left) versus reports (right).

In their study, Fischbacher and Föllmi-Heusi could only infer by means of analysing the distribution of reports that both honest and dishonest behaviours had occurred. Here, in contrast, the ground truth is available. In this experiment, 7460 rolls were recorded. In Figure 5.7, left, one sees that there was a uniform distribution across all die faces, as one would expect from a fair die. However, the distribution of reports is different, see Figure 5.7, right. Altogether there were 1,016 misreports (13.6%) whereby the highest paying face (five) is strongly over represented while the lowest paying (one and six) are under represented.

I next looked into the truthfulness of the 373 participants. The clear majority (249) was entirely truthful, however, a sizeable minority (124) made at least one misreport, see Figure 5.8, left. I also examined the number of misreports that the 124 misreporters made. The most frequent value was one misreport (23 participants), see Figure 5.8, right. This leaves 101 participants (27%) who repeatedly misreported. Among these, there were 19 participants, which is only 5% of all, who exclusively reported the pay-off-maximizing five. For detailed visual presentation of rolling and reporting data see Appendix C.2.

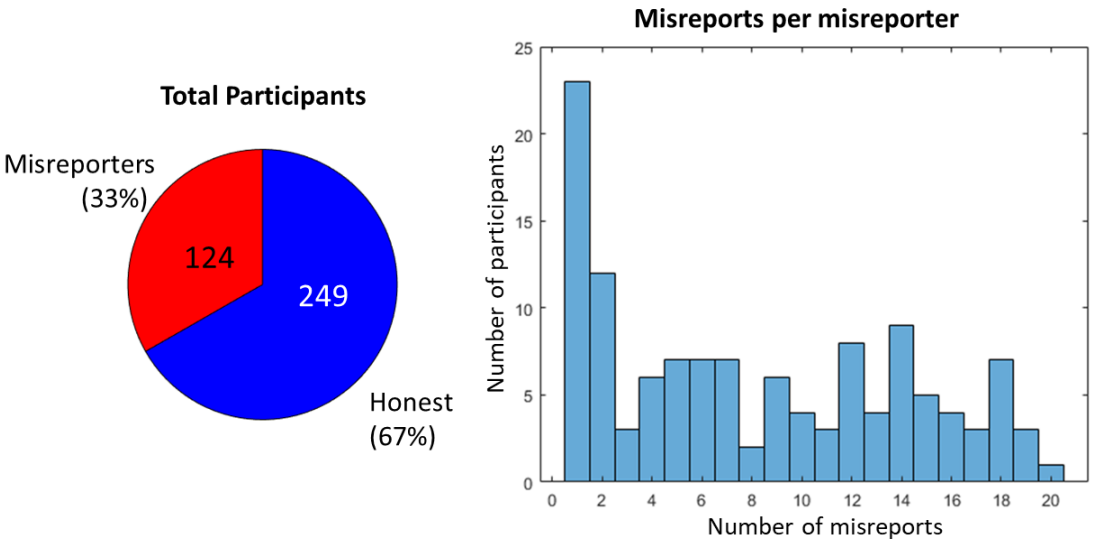

Figure 5.8: Proportion of misreporters (left) and frequencies of the number of their misreports (right).

5.4.1 The effects of the stress treatment on deception

Next, I looked at the effects of the stress treatment. Overall, 99 (32%) of the 308 participants who received either the cold water treatment (stress) or warm water treatment (control) entered at least one misreport, Figure 5.9, left. Only 308 participants received a water treatment due to a change in experimental procedure which will be explained later. Among the 151-strong control group, this was seen for 58 (38%) participants. Remarkably, the proportion of misreporters dropped to 41 (26%) in the similarly sized experimental group that had received the cold water stress treatment, see Figure 5.9, middle and right. For completion, I note that there were an additional 65 participants without water treatment, which had a very similar proportion of misreporters (38%) compared to the control group. However, since the conditions were so different between

the water-treated and untreated cohorts I do not include the latter in this comparison of treatments. I examined if the observed treatment-associated reduction in misreporters was statistically significant. To do this, I did a two sample t-test at a significance level of 5% comparing warm and cold treatments. I only took into account whether a participant was a misreporter or not, that is, the test did not distinguish between a misreporter who misreported only once and a misreporter who misreported a number of times. The difference between the treatments was statistically significant. In the APA style, the relationship between cold and warm water reporters is given by t(306)=-2.32, p=0.02, Cohens d=-0.26.

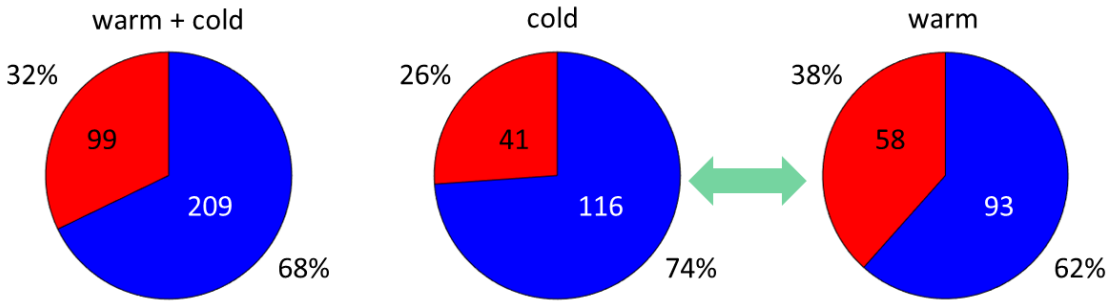

Figure 5.9: Pie charts of misreporters showing cold water (middle) and warm water treatment (right) and combined (left). Red sections represent the proportion of people who misreported once or more. Blue sections represent the proportion of honest people. Numbers in the pie are participant numbers. The green double arrow indicates that the difference in proportions of misreporters and truth-tellers was significant for the two sample t-test with a significance level of 5%.

In addition, I did a similar analysis to assess how the stress treatment affected the amount of misreporting each participant did as opposed to just whether or not they had misreported. Overall, 821 (13%) of the 6,160 reports in the combined treated and control (warm water) groups were misreports, Figure 5.10, left. In terms of the impact of stress, the same pattern emerged as above. Of the 3020 reports in the control group, 533 (18%) constituted misreports, whereas in the stressed group with 3140 reports this was halved to 288 (9%), see Figure 5.10, middle and right. Thus, not only were there fewer lies in the stress group, but there were also fewer misreports per misreporter. In this case, I also did a two sample t-test at a significance level of 5% as above, however, this time I took into account how much a misreporter misreported, that is, each participant was now represented by a number 0–20 according to the number of times they lied. Once again, the difference between the treatments was significant. In the APA style, the relationship between cold and warm water lies told by misreporters is given by t(306)=-2.89, p=0.004, Cohens d=-0.33.

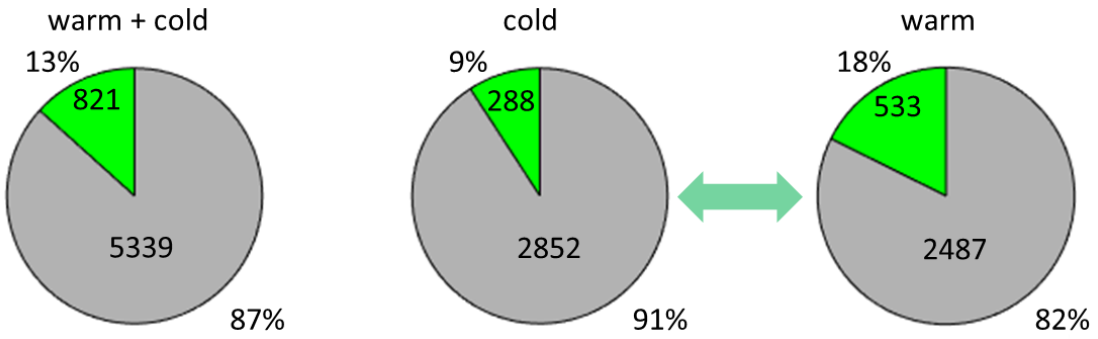

Figure 5.10: Pie charts of misreports according to treatment as in Figure 5.9. Green sections represent the proportion of misreports. Gray sections represent the proportion of truthful reports. Numbers in the pie give the exact number of misreports and honest reports. The green double arrow indicates that the difference in proportions of misreports and truthful reports per reporter was significant for the two sample t-test with a significance level of 5%.

5.4.2 Effects of gender on deception

I then differentiated reporting according to gender, as this has previously been found to play a role in decision making and deception (Croson and Gneezy, 2009). As shown in Figure 5.11, bottom left, 73 (42%) of the 173 male participants were misreporters, whereas among the 200 females only 49 (25%) misreporters were found, top left, which is significantly lower according to the two sample t-test with significance at 5%. For this comparison, I included all 373 participants regardless of treatment as I was not interested in treatment effects but gender effects. However, I did compare the effects of gender within the different treatments. I found that, restricted to the cold water treatment, the effects of gender were statistically significant, Figure 5.11, centre left, as they were when restricted to the warm water treatment, Figure 5.11, centre right. Only the 'no water' treatment group showed no difference between males and females, Figure 5.11, right. Importantly, however, restricting the analysis to females only, the reduction in misreporting between the cold (stress) and warm water treatments was not significant, see Figure 5.11, top centre. The same observation was made for the male participants, see Figure 5.11, bottom centre. This is not in line with results in the previous section, where I found that cold water stress increased truthfulness. This discrepancy could be explained by the lower participant numbers and/or an increased representation of females in the cold water treatment. I therefore listed the distribution of males and females across treatments and indeed found an over-representation of females especially in the cold water treatment, see Table 5.2. This raised the possibility that the gender effect supersedes

the stress effect. However, this was ruled out by an ordinary linear regression analysis, performed by Professor Hernán-González, which confirmed that there was indeed a gender effect, with females lying significantly less than males, but that even taking into account the gender effect, the treatment effect remains strong with a p-value less than 5% (results not shown). The analysis is similar to that done in Corgnet and Hernán-González (2019).

I then did the same analysis taking into account how much people misreported, that is I took into account the number of times a person misreported, not just whether they did or didn't ever misreport. The outcome was essentially identical to the above, with males entering twice the number of misreports (18%) that females did (9%), see Figure 5.12, left. One notable difference to the comparison of misreporters is that now, when restricted to the males alone, cold water treatment significantly reduced the number of misreports compared to the control group, see Figure 5.12, centre bottom.

Gender	Cold	Warm	No Water	Total
Female	96 (61%)	80 (53%)	24 (37%)	200 (54%)
Male	61 (39%)	71 (47%)	41 (63%)	171 (46%)
All	157	151	65	373

Table 5.2: Percentages of males and females across the treatments.

5.4.3 Effectiveness of the cold water treatment to induce stress

To assess how effectively the cold water treatment induced stress in the participants, I used the results of the manikin self-assessment questionnaire, which is used to rate a person's fundamental emotional state on a scale from 9–1. This was added to the investigation late and was thus only filled out by 170 people. The average values reported by treatment for each of the four manikins are shown in Table 5.3. A two sample t-test with a significance level of 5% showed that on the Happy—Unhappy and In control—Controlled dimensions, participants reported no significant differences between treatments. On the Relaxed—Stressed dimension I observed a shift towards 'Stressed' in the cold water-treated cohort to 6.2 versus 6.6 in the control group, however this was not

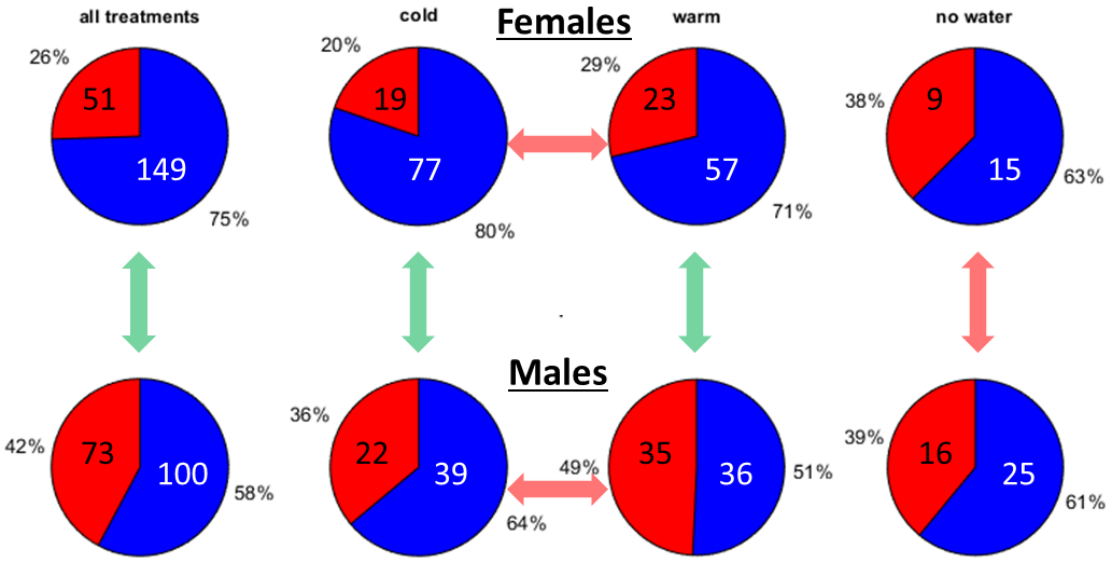

Figure 5.11: Proportions of female and male misreporters by treatment. Red sections represent misreporters, blue sections truthful reporters. Double pointed arrows indicate a t-test was performed, green signifies significance, red no significance.

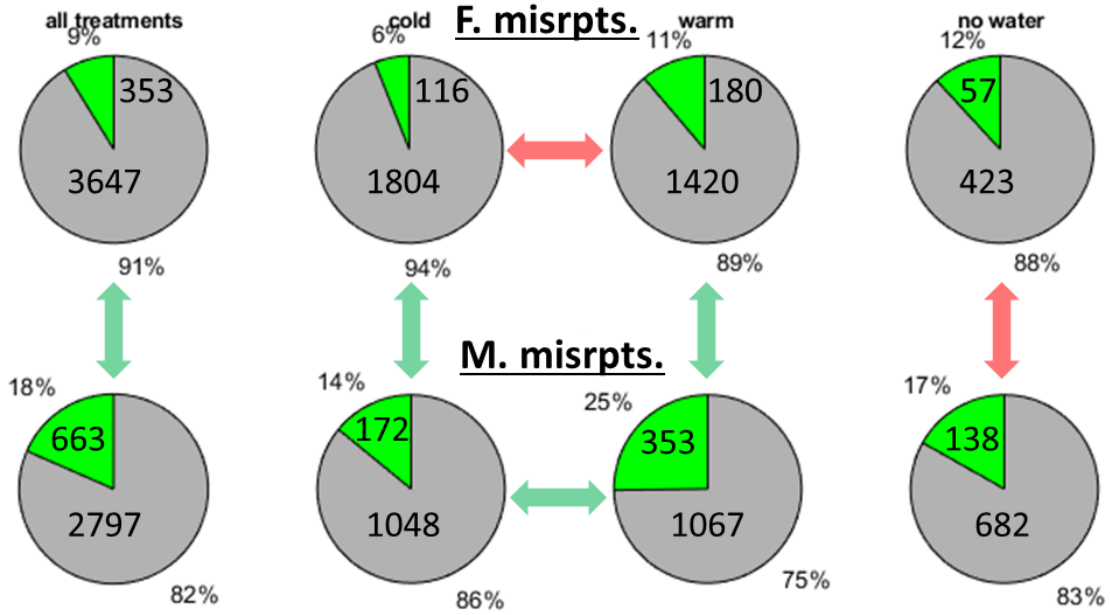

Figure 5.12: Comparison of gender effects on misreporting by treatment. Top, misreports by females; bottom, misreports by males. Green sections represent misreports, gray truthful reports. Double pointed arrows indicate a t-test was performed, green signifies significance, red no significance.

significant. On the Calm—Excited dimension, in contrast, the cold water-treated participants self-assessed significantly more towards the 'Excited' pole (4.6) compared to the control group (3.7). As the Calm—Excited dimension showed statistical significance between the cold water treatment and the control group, I give the more detailed APA

style t-test statistic as t(172)=-2.77, p=0.006, Cohens d=-0.42. I thus conclude that the cold water treatment indeed resulted in the desired stress inducement, albeit weakly.

	Happy	Excited	In control	Relaxed
Control	5.9 ±1.8	3.7 ±2.2	5.4 ±2.3	6.6 ±2.0
Cold water	6.0 ±1.7	4.6 ±2.1	5.6 ±2.1	6.2 ±1.9

Table 5.3: Average values (± standard deviation) for each of the self-report manikins for both the warm (control) and cold water treatments. Green highlighting indicates a statistically significant difference.

5.4.4 Validity of the ground truth assumption

For all participants, what they rolled as well as what they claimed to have rolled are known, as this was recorded by the computer program they interacted with. The assumption being made in this experiment is that misreports, which can be easily verified, are intentional lies. One can question whether this assumption is valid. Participants might have misreported by accidentally mistyping or by misunderstanding the experiment. In this case one would expect misreports to happen randomly, that is, without a preference for profit or loss. To get an impression of the prevalence of intentional misreporting, aka lying, I analysed the relationship between misreporting and profitability. The results are summarized in Table 5.4.

Misreports	**Cold**	**Control**	**No water**	**All sessions**
All	288	533	195	1,016
Profitable	283 (98.26%)	530 (99.44%)	193 (98.97%)	1,006 (99.02%)
Unprofitable	5 (1.74%)	3 (0.56%)	2 (1.03%)	10 (0.98%)
Misreporters				
All	41	58	25	124
Accidental	1 (2.44%)	2 (3.45%)	0	3 (2.42%)

Table 5.4: Proportions of misreports and misreporters. Accidental misreporters are participants who made one false report that led to a loss.

Among a total of 7,460 reports across all treatments, I recorded 1,016 misreports. Only ten of these (0.98%) were loss-making misreports, while 1,006 (99.01%) were profitable misreports, whereby this figure did not significantly differ between the cold water, control, and 'no water' treatments. If one assumes that non-profitable misreports were accidental, then one could also estimate there were just as many profitable but accidental misreports. This would then make under 2% of all misreports accidental. I thus conclude that at least 98% of all misreports were intentional and hence constitute incidences of deceptive behaviour.

Table 5.4, bottom, lists the relationship between all misreporters, namely people who misreported at least once, and accidental misreporters, which are people who made just one misreport that did not lead to a profit. There were 124 misreporters among the 373 participants across all treatments. Of these, three are accidental (2.42%), whereby once again this figure did not differ significantly across treatments. Thus, assuming the same number of accidental misreporters who made profits, this would mean that at most 6 participants (5%) who misreported did not actually mean to.

Therefore, it is safe to say that the vast majority of misreports were intended to deceive and the majority of misreporters were intending to be deceptive. I therefore think the experiment successfully elicited real lying behaviour and it is reasonable to say with high probability, that that which is recorded as a misreport is really a lie.

5.4.5 Mouse positions

Although it was not part of the original conception of the experiment, I built the die experiment software to also record the mouse positions of the participants along with their timestamps. Recording of mouse positions began when the start button was pushed and ended when the final reward was displayed after the completion of the 20^{th} round. As this additional feature did not interfere with the other data collection, I incorporated it into the final software design so that it could be used for future studies beyond analysis of facial cues. The rationale for this is as follows: there has been considerable research into the connection between cognitive states and movement trajectories such as mouse movements. It has been shown, for instance, that a higher cognitive load influences mouse trajectories when a study subject makes decisions that involve navigating with their mouse to click buttons to log these decisions. Dale and Duran showed that mouse trajectories

reveal cognitive stress in evaluating negative sentences (Dale and Duran, 2011); Freeman and colleagues showed that mouse trajectories reveal a heavier cognitive load for a study subject distinguishing the race of mixed race individuals if the subject had less interaction with mixed races (Freeman et al., 2016); in a deception study, Monaro and colleagues showed that mouse trajectories could be used to distinguish people who were lying about their identities from those who were being truthful (Monaro et al., 2017). These and similar studies have taken the view that "the dynamics of action have become a valuable signature of ongoing cognitive activity, revealing finer-grained characteristics of these processes", (Dale and Duran, 2011). In analysing mouse trajectories, these studies use such characteristics as curvature, speed, acceleration, distance from the shortest path, x-flips and y-flips (changes in direction along the x- and y-axes respectively), length of trajectory and reaction time, information that is contained in my database. Already a preliminary analysis of mouse trajectories shows distinctive patterns appearing to emerge for honest participants, homo economicus, disguised and accidental liars, see Figure 5.13. While this constitutes only an initial snapshot of research that is outside the scope of the current thesis, this phenomenon nonetheless warrants further in-depth research using the wealth of data that has been gathered in this study.

5.5 Concluding remarks

This experiment was successful at collecting a large amount of high quality data from 373 participants in the form of more than 1.7 million timestamped video frames (cumulative duration of just under 16 hrs) and annotated button clicks, plus additional timestamped mouse tracking data collected in parallel. The experiment also provided a thorough and complete ground truth. About the video data, more will be said in the next chapter. The experiment was also a successful investigation of human deception with several interesting results. First, two-thirds of participants were totally honest (67%), see Table 5.5. The remaining third were liars (33%), the vast majority of which were partial liars with just one misreport being the most frequent number of misreports for this group, see Table 5.8. The abundance of maximal liars, in contrast, was very limited (5%). Homo economicus, thus appears to be more an exception than the rule. Notably, this research confirms the patterns of lying that Fischbacher and Föllmi-Heusi first observed in their paradigmatic 2013 study. The same three patterns are present here too – maximal liars,

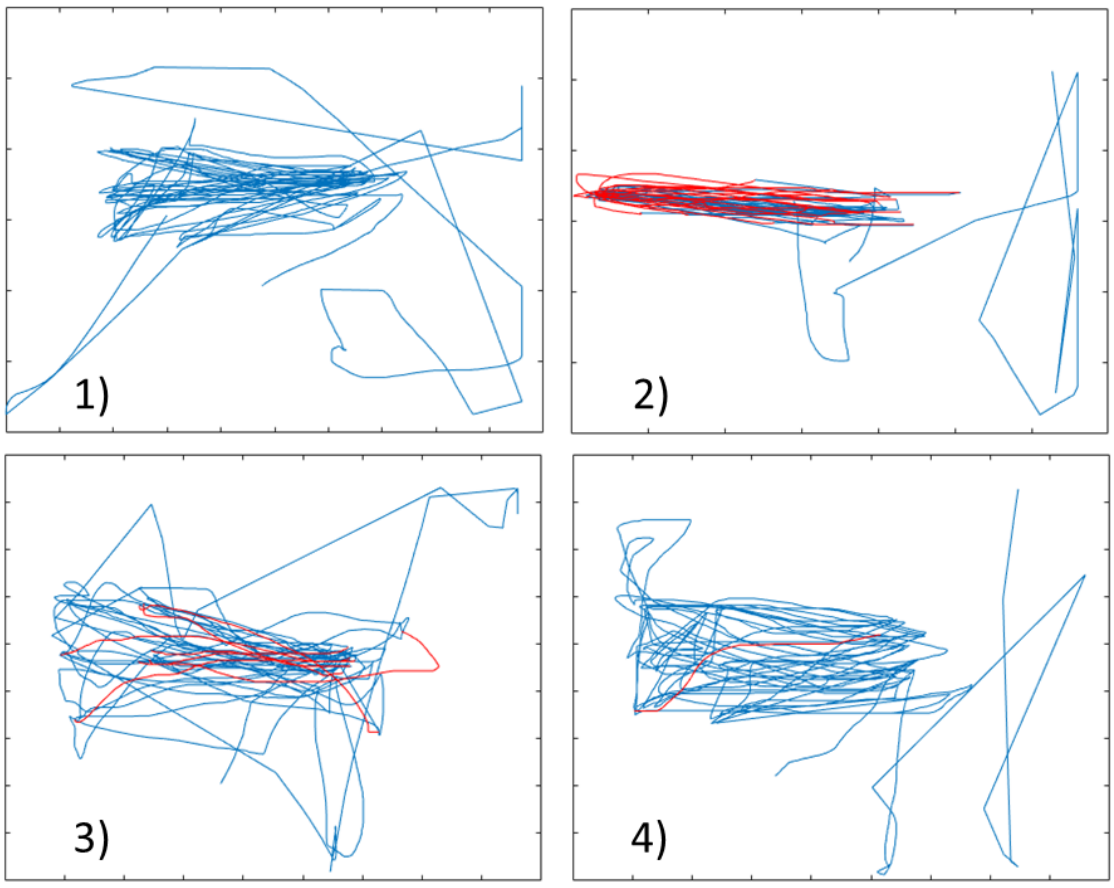

Figure 5.13: Example mouse positions for four different participants over the entire course of their die rolling experiment. A red line indicates a path from the click of the 'roll die' button to the click of a report button that was a false report. All other lines are blue. There are four potential 'types' displayed here, 1) honest 2) homo economicus 3) disguised liar and 4) accidental liar.

partial liars and totally honest participants. The numbers and percentages with which they occur in this experiment are given in Table 5.5. This was despite the fact that here the interaction was mediated by a computer and despite the fact that the participants knew they were being recorded.

	Control	Stress	No water	All
Honest	93 (62%)	116 (74%)	40 (62%)	249 (67%)
Maximal	10 (7%)	6 (4%)	3 (5%)	19 (5%)
Partial	48 (32%)	35 (22%)	22 (34%)	105 (28%)
Total	151	157	65	373

Table 5.5: Table classifying study participants as honest, maximal, and partial liars according to Fischbacher and Föllmi-Heusi across treatments. See Appendix C.2 for complete rolling data.

Studying the consequences of stress on deceptive behaviour is a major tenet of this study. In fact, the study provided evidence that the cold water treatment successfully induced stress in the participants. This study further indicates that the three types of lying behaviours are robust traits that are found irrespective of stress and the different treatments used. Strikingly, it was shown that the cold water stress reduced the amount of lying. Moreover, there was a clear gender effect, with there being significantly fewer female misreporters than male, and the amount of lies per dishonest participant was lower for females. Finally, the known ground truth of this experiment makes it conceivable to distinguish more types of lying, for instance, splitting the groups of partial liars into those who disguise their lies and those who do not want to appear greedy. The concurrently collected information on mouse movements might prove useful to refine such distinctions in future research.

Chapter 6

Results - Computer Vision Analysis of the Dice Rolling Experiment Links Head Pose to Deception

6.1 Investigating facial expressions in the dice rolling experiment

The die experiment produced just over 1.7 million frames of video covering the experiments of 373 different participants. The first thing I did was to run action unit software on these videos for the 12 AUs for CNN-BLSTM and OpenFace. It took around two minutes for OpenFace to process each of the 373 videos and around one hour for CNN-BLSTM to achieve the same. As was observed before for the poker database, CNN-BLSTM is also more difficult to run and requires frequent restarts, which added to the time required to collecting the AU data over all videos.

The dice rolling experiment, unlike the poker study, does not mainly involve human-human interaction. Instead, the participants only interact with the computer during the experiment and so one might expect that this affected the types and intensities of emotions displayed. After having watched several of the videos, I got the impression that the level of facial expression was indeed generally lower than with the poker data set. Several of the participants seemed to have very little expression, nonetheless, there were

facial expressions occurring. It is hard to judge just by watching, as the poker dataset also contains participants with low levels of expression. To get a more objective idea of expressiveness in the dice rolling experiment, I computed the means and standard deviations for the 12 AUs as detected by both CNN-BLSTM and OpenFace. I interpret the mean as the strength of an AU and the standard deviation as the amount that AU changes while producing different facial expressions. Table 6.1 shows the comparison between the dice rolling experiment and poker for the CNN-BLSTM detectors. In line with visual inspection, all of the values, except those of AU9 (nose wrinkler) and AU45 (blink), are either lower for the dice rolling experiment or very close to equal. That AU9 and AU45 appear to be markedly higher for the dice rolling experiment was unexpected, given their inconspicuous role in the poker data set. On the other hand, AU12 and AU5 that dominate the detection of facial expressions in the poker experiment, were represented only weakly in the dice rolling data set. This does, however, reflect the videos upon inspection, as it is obvious that there is much less smiling taking place than in the poker dataset. The respective outcomes for the Openface detectors are shown in Table 6.2. As with the poker data, the values are very low across all AUs, whereby no particular AUs were detected particularly well. Although some of OpenFace's AUs were marginally higher for the dice rolling experiments compared to the poker data, including AU9 and AU45, I do not consider these differences as meaningful detections of actual facial expressions.

	Mean		S.D.	
	Dice rolling	Poker	Dice rolling	Poker
AU1	0.1702	0.2363	0.1855	0.2216
AU2	0.1748	0.1967	0.2094	0.2124
AU4	0.1047	0.2616	0.2281	0.3151
AU5	0.0957	0.1145	0.1160	0.1350
AU6	0.0498	0.1067	0.1068	0.1574
AU9	0.1413	0.1139	0.2253	0.1921
AU12	0.0573	0.1146	0.1213	0.1805
AU15	0.0671	0.1209	0.0632	0.1109
AU20	0.0729	0.0714	0.0588	0.0692
AU25	0.1551	0.1754	0.1644	0.1801
AU26	0.0737	0.0923	0.1189	0.1171
AU45	0.1489	0.0994	0.2124	0.1647

Table 6.1: CNN-BLSTM detectors: Comparing mean and standard deviation (S.D.) of AU values for dice rolling and poker experiments.

Nonetheless, it was surprising that AU9 (nose wrinkle) and AU45 (blink) went up by 24% and 52% respectively for CNN-BLSTM, as these had not previously stood out in

	Mean		S.D.	
	Dice rolling	Poker	Dice rolling	Poker
AU1	0.0384	0.0386	0.0832	0.0847
AU2	0.0207	0.0167	0.0640	0.0486
AU4	0.0843	0.0817	0.1216	0.1226
AU5	0.0112	0.0090	0.0360	0.0247
AU6	0.0185	0.0588	0.0528	0.1074
AU9	0.0163	0.0141	0.0514	0.0378
AU12	0.0352	0.0575	0.0728	0.1079
AU15	0.0295	0.0456	0.0637	0.0811
AU20	0.0703	0.0881	0.1012	0.1250
AU25	0.0716	0.0872	0.1035	0.1089
AU26	0.0384	0.0242	0.0800	0.0506
AU45	0.1489	0.0994	0.2124	0.1647

Table 6.2: OpenFace: Comparing mean and standard deviation (S.D.) of AU values for dice rolling and poker experiments.

the poker dataset. Therefore, I decided to manually check the dice rolling videos to see if there might be an explanation for this. For each of AUs 9 and 45, I inspected the ten videos of those participants who had the highest means for these values and made the following observations. For AU9 (nose wrinkler), I did not note appreciable nose wrinkling. However, all ten participants were wearing large, dark rimmed glasses. Because of this striking commonality I conclude that the high values for AU9 are false positive signals and do not result from genuine nose wrinkling activity. Similar observations were made concerning AU45 (blink). Here, the ten participants with the highest averages fell into two categories; six did not display noticable blinking but had major occlusions (five dark rimmed glasses and one baseball cap), while the other four were clearly blinking and had their faces very close to the camera. Therefore, it is likely that for people sitting at a normal distance to the camera, the detectors cannot detect blinks well but are prone to reporting artefacts (glasses, hat), as, for instance people were in general sitting farther away in the poker dataset and not necessarily blinking less. There were similar problems with OpenFace regarding glasses, hands on the face, and hats causing occlusion, head motion affecting the detection of action units involving eyes and eyebrows and higher AU45 values for people with their faces closer to the webcam. Examples of participants are not shown for the dice rolling experiment due to data protection requirements. In order to better understand the difference between occlusions in the poker study and the dice rolling study, I viewed all of the videos and counted the number of participants with glasses and the number of participants who kept their hands on their face for at least 10%

of the video. The results are shown in Table 6.3.

	Glasses	Hand on face
Dice rolling (373 participants)	169 (42%)	141 (38%)
Poker (64 participants)	14 (22%)	12 (19 %)

Table 6.3: The number of occlusions caused by glasses and hands on face in the dice rolling and poker databases.

Here one can see that nearly double the percentage of participants of the dice rolling game are wearing glasses. At 42%, this is approaching half of the participants, so there will be considerable occlusion and interference particularly with AUs involving eyes and brows (AU1, AU2, AU4, AU5, AU45) and also AU9 (nose wrinkler), where I have shown that glasses can cause false positives. Additionally, the percentage of participants in the dice rolling experiment who have their hands on the face for a longer period of time is twice as high than for the poker data set. It is interesting to note that touching the face has also been investigated as a possible sign of deceitfulness (DePaulo et al., 2003). As the expressive levels detected by the AU detectors, both CNN-BLSTM and OpenFace, are much lower for dice rolling than for poker, and as the occlusions are much higher for dice rolling, I did not think that investigating this dataset with the 12 AU detectors used so far was a very promising way forward. Yet, the dice rolling videos are notably different from the poker videos in another regard. While there is less facial expressiveness, the manual analysis suggested that some participants seem to display more head motions, which is a property detected by OpenFace. I therefore decided to rather assess this aspect of emotion detection.

6.2 Investigating head pose in the dice rolling experiment

As part of its detection of facial action units, OpenFace 2.0 (Baltrušaitis et al., 2018) computes head pose in terms of x-, y- and z-rotations (pitch, yaw and roll), as shown in Chapter 3, Figure 3.4. In OpenFace, head pose is computed directly from the 68 facial points that form the basis of the toolkit's AU detection system (Zadeh et al., 2017). While head pose has so far not been the primary focus of automatic emotion recognition (Gunes and Pantic, 2010b; Ramirez et al., 2011) it has been recognized as a valuable source of

information about a person's emotions (Adams et al., 2015; Gunes et al., 2015; Tracy and Matsumoto, 2008). To gain an understanding of whether head pose might reveal behavioural differences in the dice rolling study, one might split the data of each participant's reports into two groups, lies and honest reports. I decided against this approach because it may fail to reveal global behavioural differences between the different types of reporters. Therefore, I decided to focus on an alternative approach, namely the idea that truthful and dishonest reporters can be distinguished, and have assumed that even a dishonest participant's truthful reports are part of their overall deceptive behaviour. To test this possibility, I sorted participants into two groups according to their lying behaviour, entirely honest and dishonest.

6.2.1 Preparing the data

In the poker study, I focused on FCR-events. Here, in the dice rolling experiment, I similarly focused on the moments when the participant decides to tell the truth or lie, that is when they report the face of the die. I extracted data from the time each participant pressed the 'roll die' button to the time they pressed the button to log the number they had rolled. For each participant there were 20 such segments. To characterize head pose for each segment, I determined the x, y and z angles their head pose passed through during each of their decisions. These values were computed as follows. For each of the participant's ith round, $i = \{1 \dots 20\}$, let $f_{i,j}$ represent the jth frame in the sequence of frames $j = \{1 \dots n\}$ recorded from the time they clicked the roll die button to the time they click their corresponding report. For each image, OpenFace computes head pose relative to the three axes x, y and z. Let x($f_{i,j}$) be the angle of the head around the x-axis as computed by OpenFace. Then the maximum head angle for a round was taken to be the absolute value of the difference between the maximum and minimum values occurring in this range, or

$$\textbf{max rotation around x-axis} = \mid \max_j(x(f_{i,j})) - \min_j(x(f_{i,j})) \mid \text{, for } j = \{1, \dots, n\}.$$

The values representing the angular rotation around the y- and z-axes were defined analogously. Additionally, I also defined general face motion to characterize head movement by choosing a facial point, represented by (x, y) pixel coordinates, and computing the length of the two-dimensional path it followed by using facial points detected by Open-

Face. This is closely related to angle displacement, but captures the amount of movement better as moving back and forth more might not increase the maximum angle, but will increase the total length of the path the points traverse. To compute this, of the 68 facial points computed by OpenFace, I selected a single one on the upper right hand cheek, see Figure 6.1. I chose this point, facial point 3, because it was rarely occluded or off screen. I used a straightforward definition of path length P of this point

$$\textbf{length of path that P follows} = \sum_{j=1}^{n-1} \|P_j - P_{j+1}\|_2 \,.$$

So for each participant, four values were calculated for each of their twenty decisions: the three angles their head poses spanned around the x-, y- and z-axes, plus the path length that facial point 3 followed during their decision. Of these twenty sets of four values, I tested the mean and maximum to look for significant differences between honest and dishonest participants.

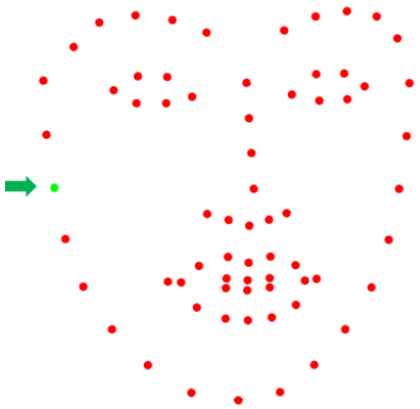

Figure 6.1: The 68 facial points computed by OpenFace. The green arrow points to the third facial point, which is used for calculating the length of path that the head traverses.

6.2.2 Basic statistics

Averaged participant features

To probe if there were significant differences between the *average* head motions of honest participants and those of dishonest for the control and stress treatments, I looked at three groups of dishonest participants: all dishonest participants, homo economicus and partial liars and compared each of these groups to the truthful reporters. I also did sta-

tistical tests on all but homo economicus to check if these differences were significant. Homo economicus, although an interesting group, was left out of the statistical analysis because of its small size (10 in the warm water treatment and 6 in the cold water treatment). See Table 6.4 for the comparison of averages for the warm water controls and Table 6.5 for comparison of averages for the stress (cold water) group.

Warm water	pitch (x-axis)	yaw (y-axis)	roll (z-axis)	path length
All (151)	$7.6° \pm 5.1°$	$7.3° \pm 4.1°$	$5.8° \pm 3.8°$	71.9 ± 47.1
Honest (93)	$7.2° \pm 4.8°$	$7.7° \pm 4.1°$	$5.6° \pm 3.4°$	69.8 ± 44.9
Liars (58)	$\bullet 8.1° \pm 5.5°$	$\bullet 7.3° \pm 4.1°$	$\bullet 6.2° \pm 4.5°$	$\bullet 75.3 \pm 50.6$
H. e. (10)	$7.5° \pm 6.6°$	$5.2° \pm 3.3°$	$4.8° \pm 3.1°$	65.0 ± 27.3
Partial l. (48)	$\bullet 8.2° \pm 5.4°$	$\bullet 7.7° \pm 4.1°$	$\bullet 6.5° \pm 4.7°$	$\bullet 77.5 \pm 54.2$

Table 6.4: Warm water treatment. Average values for the four features pitch, yaw, roll and path length. An unpaired t-test was performed for each of the four variables, once comparing 'honest' to 'liars' and once comparing 'honest' to 'partial liars' (partial l.). If the test was significant, a dagger was placed in the cell for 'liars', or respectively, 'partial liars'. If the test was not significant, a bullet was placed in the cell. The lowest value in each column is highlighted blue, the highest red. H. e., homo economicus.

Cold water	pitch (x-axis)	yaw (y-axis)	roll (z-axis)	path length
All (157)	$7.8° \pm 7.2°$	$8.4° \pm 5.4°$	$5.9° \pm 5.0°$	76.1 ± 54.0
Honest (116)	$6.6° \pm 4.5°$	$8.1° \pm 5.2°$	$5.2° \pm 3.3°$	67.8 ± 34.6
Liars (41)	$\dagger 11.0° \pm 11.5°$	$\bullet 9.2° \pm 6.1°$	$\dagger 7.8° \pm 7.8°$	$\dagger 99.5 \pm 84.8$
H. e. (6)	$10.8° \pm 7.4°$	$8.0° \pm 5.9°$	$7.9° \pm 7.4°$	69.3 ± 51.8
Partial l. (35)	$\dagger 11.0° \pm 12.1°$	$\bullet 9.4° \pm 6.2°$	$\bullet 7.8° \pm 8.0°$	$\dagger 104.6 \pm 88.8$

Table 6.5: Cold water treatment. Average values for the four features pitch, yaw, roll and path length. Legend as described in Table 6.4.

From the tables, one can see that for the warm water treatment there were no statistically significant differences between the honest group and either of the tested dishonest groups, although there is a tendency for honest participants to have smaller values, that is they move their heads less than dishonest participants. After cold water treatment, however, this tendency clearly solidifies, and the separation of honest and dishonest groups widens, with honest participants tending to move less now than under control conditions before. With the dishonest participants the opposite occurs, as their head movements tend to be larger after cold water stress in comparison to the control conditions. As a result, there are now significant differences between the honest and dishonest groups for most variables except for 'yaw' (both 'liars' and 'partial liars') and 'roll' ('partial liars'), which

did not reach significance at the 5% level.

Maximum participant features

For each participant, the averaged features above capture information from all 20 of that participant's rounds. I also assessed each player's *maximum* value feature over all 20 rounds. Here, the maximum value is kept and information about their other rounds is disregarded. I interpret the maximum value as meaning the event happened at least once. As for averaged features, I also determined if the maximum features differed between 'honest' and the various dishonest groups. The results of the comparison for the warm water treatment are shown in Table 6.6, and for the cold water treatment in Table 6.7.

Warm water	pitch (x-axis)	yaw (y-axis)	roll (z-axis)	path length
All (151)	$40.5° \pm 48.5°$	$30.3° \pm 23.0°$	$33.2° \pm 49.1°$	357.9 ± 619.7
Honest (93)	$36.0° \pm 41.6°$	$27.2° \pm 20.5°$	$29.4° \pm 39.8°$	346.7 ± 685.3
Liars (58)	•$47.6° \pm 57.7°$	†$35.2° \pm 25.9°$	•$39.4° \pm 61.1°$	•375.8 ± 502.0
H. e. (10)	$35.6° \pm 32.2°$	$26.7° \pm 14.6°$	$29.3° \pm 18.7°$	264.6 ± 154.9
Partial l.(48)	•$50.1° \pm 61.6°$	†$37.0° \pm 27.5°$	•$41.5° \pm 66.5°$	•399.0 ± 545.8

Table 6.6: Warm water treatment. Maximum values for the four features pitch, yaw, roll and path length. An unpaired t-test was performed for each of the four variables, once comparing 'honest' to 'liars' and once comparing 'honest' to 'partial liars'(partial l.). If the test was significant, a dagger was placed in the cell for 'liars', or respectively, 'partial liars'. If the test was not significant, a bullet was placed in the cell. The lowest value in each column is highlighted blue, the highest red. H. e., homo economicus.

Cold water	pitch (x-axis)	yaw (y-axis)	roll (z-axis)	path length
All (157)	$37.0° \pm 45.1°$	$30.3° \pm 24.0°$	$29.3° \pm 38.5°$	326.1 ± 436.7
Honest (116)	$28.5° \pm 25.5°$	$25.6° \pm 17.3°$	$22.9° \pm 23.4°$	237.2 ± 162.7
Liars (41)	†$60.9° \pm 72.6°$	†$43.7° \pm 33.8°$	†$47.2° \pm 61.3°$	†577.5 ± 761.5
H. e. (6)	$34.6° \pm 23.4°$	$31.3° \pm 24.9°$	$28.4° \pm 20.1°$	223.3 ± 188.4
Partial l. (35)	†$65.4° \pm 77.3°$	†$45.8° \pm 35.0°$	†$50.4° \pm 65.5°$	†638.2 ± 806.8

Table 6.7: Warm water treatment. Maximum values for the four features pitch, yaw, roll and path length. Abbreviations as in Table 6.6.

For the control (warm water) group, the differences between 'honest' reporters and 'liars' and 'honest' reporters and 'partial liars' were minor and not significant with the exception of yaw, as seen before for the averaged values. Reassuringly, the cold water treatment

again brought about significant differences between honest and dishonest participants in all assessed features of head movements. Thus, both the averaged and maximum head movements of dishonest participants were larger than of their honest counterparts. This difference, however, required cold water stress treatment to become apparent.

Since the cold water values for maximum pitch, yaw, roll and path length were statistically significant, I made a table of the full APA style t-test statistics of each of the four attributes comparing all liars to the honest participants, see Table 6.8. Here, one can see that the p-values are small and the effect sizes are medium to large, with the effect size for maximum path length being largest.

	degrees of freedom	p-value	t-stat	Cohen's d
max pitch	155	0.000054	-4.153336	-0.754616
max yaw	155	0.000022	-4.372364	-0.794411
max roll	155	0.000411	-3.611194	-0.656115
max path length	155	0.000011	-4.551033	-0.826873

Table 6.8: APA style t-test statistics comparing honest to dishonest participants in the stress treatment according to each of the four attributes maximum pitch, maximum yaw, maximum roll and maximum path length.

6.2.3 Decision trees

To research if these differences could be translated into a classifer, I investigated the data with decision trees. As in Section 4.6, I used a leave-one-out procedure. Instead of single frames as input as I had done in poker, I used the computed average and maximum features for each participant as input. This data was easy to glean from the die rolling data set for a number of reasons. The head pose detectors are more accurate than the AU detectors, there is enough time between dice rolling and reporting and there is also a much lower level of distraction and irrelevant interaction. This is also in keeping with the view that, in this experiment, a person is either dishonest or honest throughout, which also contrasts with poker where the same participant can either fold, call or raise at each turn. The first decision trees were built over all eight features (four average values and four maximum values). Removing some of the features increased the performance. To get a better understanding of the data, I made a scatter plot of the control group and one of the stress group to see if there were any obvious visually evident differences, see Figure 6.2. I chose two variables: maximum pitch and maximum path length. Maxi-

mum values were chosen over averages because they differentiated between honest and dishonest reporters most in the statistical tests. Among these, path length was chosen, because it encapsulates information about head movement most completely - all head motion causes the path length to increase. Combining path length with pitch, yaw or roll produced similar results (not shown). Therefore, it is enough to show the scatter plots for pitch versus path length in Figure 6.2.

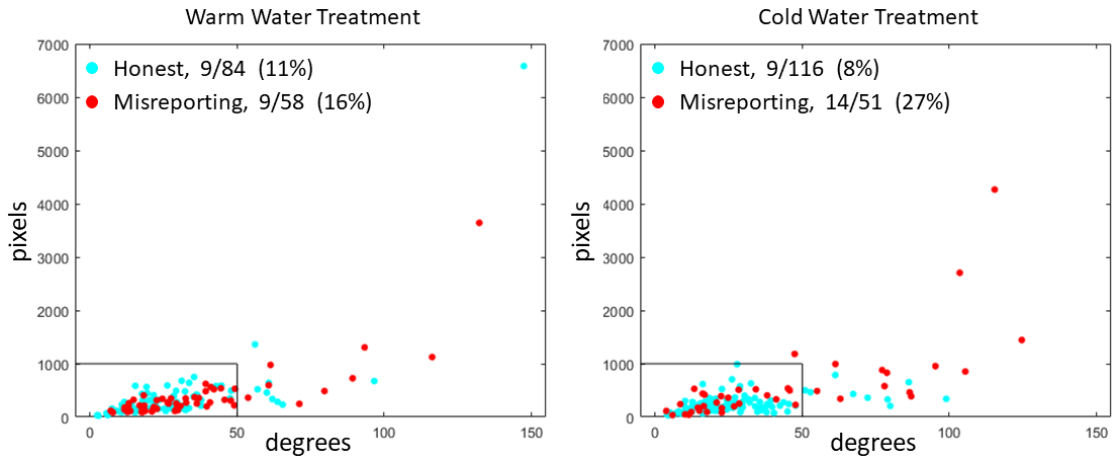

Figure 6.2: Control group scatter plot, left. Treatment group scatter plot, right. For both scatter plots, the x-axis is participant's maximum pitch in degrees, the y-axis is their maximum path length in pixels. Red points are dishonest participants, blue points are honest participants. A rectangle, which was chosen by visual inspection, was placed around a region which is meant to correspond roughly to low head motion and outside it to high head motion. The legends in the upper left corners give the ratio of honest participants outside the rectangle to all honest participants in the plot and dishonest participants outside the rectangle to all dishonest participants in the plot, respectively. In the control group, 89% of honest participants are positioned within the rectangle and 11% outside while 84% of dishonest participants are positioned within the rectangle and 14% outside. In the stress group, 92% of honest participants are positioned within the rectangle and 8% outside while 73% of the dishonest participants are positioned within the rectangle and 27% are positioned outside.

A rectangle, which was chosen by eye, was placed in the same position in each of the scatter plots to demarcate an area of low head motion from a region of high head motion. From the scatter plots one can see that between the two treatments, the dishonest subjects tend to fall more frequently outside the rectangle after exposure to cold water stress — 27% lie outside the rectangle in the stress treatment, while only 16% lie outside the rectangle in the control group. In contrast, the location of the honest subjects does not change much between the two treatments, with 8% lying outside the rectangle in the stress group and 11% lying outside the rectangle in the control group. Therefore, the treatment does seem to separate honest from dishonest subjects. It seems that within

the area of the rectangle there is not much chance that the data can be better separated based on these variables and that here the majority class should be chosen. Any further separation would appear to just cause noise effects. Although this rectangle was chosen by eye, it suggests that when building a decision tree, it might be best to restrict the tree to only two or three cuts. Using this strategy, I obtained the following decision trees shown in Table 6.3, once for the control treatment and once for the stress treatment. Here, one can see that, in the control group, the decision tree was only able to separate a small sliver with the second attribute x_2 (path length) being between 5.27 and 7.93 pixels, to be classified as dishonest, indicating that the data is not separable. In contrast, in the cold water treatment, the decision tree was able to section off a rectangle defined by $x_1 > 15.70$ and $x_2 > 377.97$, to classify as dishonest and the rest as honest. So in this case, the behaviour between the two treatments is different and also leads to different classification behaviour, which is not surprising given the previous statistics of Tables 6.6 and 6.7 and also Figure 6.2 where the classes look harder to separate in the control group than in the stress group. To compare the rectangular areas found by the decision tree with the hand chosen rectangle in Figure 6.2, scatter plots have been made again in Figure 6.4, this time using the attributes roll and path length instead of pitch and path length, together with the decision boundaries found by the decision trees.

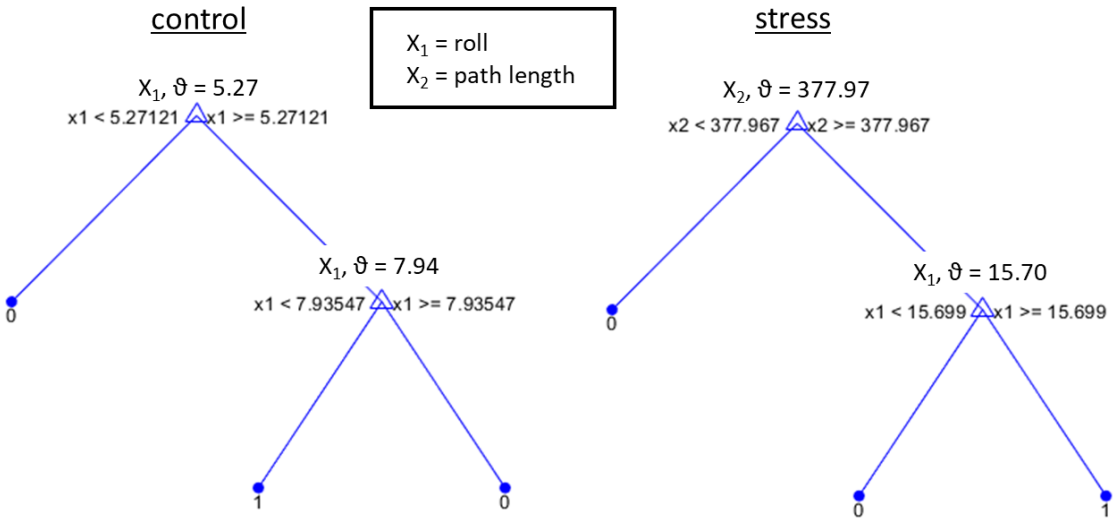

Figure 6.3: Decision trees built using two attributes, roll, x_1, and path length, x_2. Left is the tree built over the warm water treated group. Right is the tree built over the cold water (stress) treated group. Trees were only allowed to make two cuts. Thresholds for splits at split nodes are given by θ.

Viewing the performance measures of the trees for the different treatments, see Table 6.9, one can see that the performance measures for the stressed group are better than for

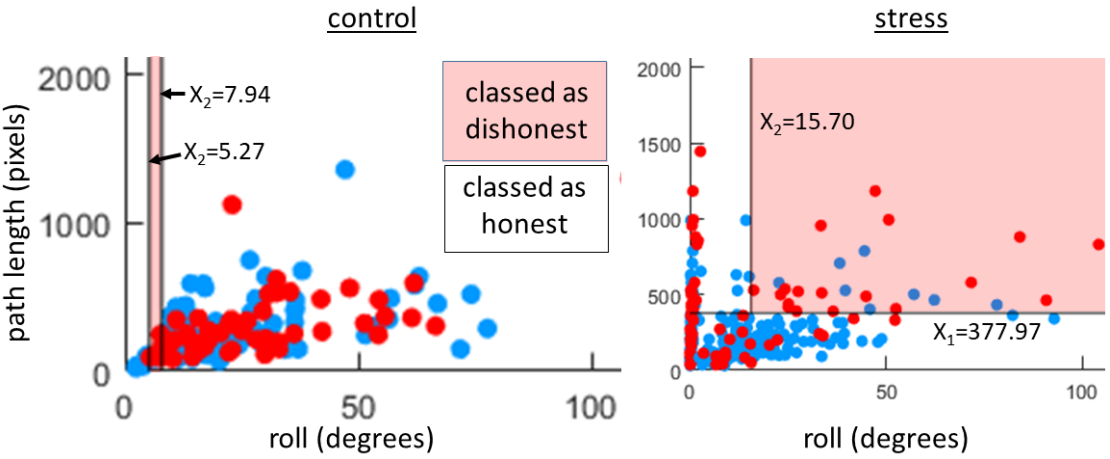

Figure 6.4: Scatter plots for control (left) and stress (right) using two attributes, roll and path length together with the decision boundaries found by the decision trees in Figure 6.3. Blue dots represent honest subjects and red dots represent dishonest subjects. Red shading indicates areas classified as dishonest by the corresponding decision tree. White areas are classified as honest. In the control group, left, the honest and dishonest groups overlap and the decision tree can only make one narrow cut. In the stressed group, right, the two groups can be better separated with a large rectangular area in the upper right hand corner being classed as dishonest.

	Warm water	Cold water
Honest/All (naive cl.)	(91/151) 0.6159	(116/157)0.7389
Classification rate	0.6225	0.7962
Precision	0.5555	0.6364
Recall	0.0862	0.5122
b_class rate	0.5216	0.7044

Table 6.9: Performance values for decision tree built over all treatments using three features, maximum yaw, maximum roll and maximum path length. Here, 'naive cl.' is the naive classification of assigning the majority class to all instances.

the control group. The classification rate of 0.7962 for the tree built on the cold water participants is better than the naive classification rate of 0.7389 obtained by just choosing the majority class. As in the poker data set, the classes are imbalanced, with nearly three quarters correctly reporting and one quarter misreporting. Given this imbalanced data, I also looked at the confusion matrix, as this shows all the information about the classifier's performance and has been suggested as useful for such cases (Pantic, 2009). This is shown in Table 6.10, where the rows represent the true class and the columns represent the class returned by the classifier.

From the perspective of lie detection, the confusion matrix reveals that, given that the

	Cl. misreporter	Cl. honest
Misreporter	21	20
Truthful	12	104

Table 6.10: Confusion matrix for the classifier built from the participants who underwent the stress treatment. Values along the diagonal (yellow) are the number of correctly classified instances. Numbers in white cells give incorrectly classified instances. Cl., classifiers.

participants underwent a stress treatment, the decision tree would detect a liar correctly about half of the time. It would correctly detect an honest person as honest about 90% of the time. About 10% of honest people would, however, be falsely classified as liars and about 50% of liars would be falsely classified as honest.

6.3 Concluding remarks

In this chapter, the CNN-BLSTM and OpenFace AU detectors were run on the dice rolling videos. It was seen that detection levels were very low for both and that there were many false positives caused by facial occlusions. This precluded the meaningful detection of facial expressions present in the dataset. On the other hand, head pose is an additional feature relevant for emotions and deception that is readily available as part of the OpenFace detectors. Using statistical tests, this modality revealed that there were significant differences in the behaviour of honest and dishonest subjects, which, however, only became apparent after participants were subjected to the stress treatment. Here, for maximum values, all four attributes, pitch, yaw, roll and path length were statistically significant with a moderate to large effect size, see Table 6.8, the most discriminative attribute being path length with t(155)=-4.1533, p = 0.000011 and Cohen's d = -0.8269. It was found that under stress the dishonest reporters displayed significantly larger head movements in comparison to the honest participants, a finding that was corroborated using decision trees. In the stressed group I was able to obtain a classifier with an accuracy of 80%. Inspecting the confusion matrix shows that it classifies a misreporter as a misreporter with 50% accuracy and a truthful person as truthful with nearly 90% accuracy. Of those classified as misreporters, 36% are actually honest, and of those classified as honest, 16% are liars. This supports the idea that people do lie differently when subjected to stress than without it.

Chapter 7

Discussion

This thesis is an investigation into affective computing and decision making using automatic action unit detectors together with detectors for other AU-related features such as head pose. The main question is, "can action unit detectors together with head pose detectors detect deception?" I examined two different deception scenarios, and the answer I found is, "yes, better than randomly choosing, if one applies similar standards to those used for polygraph deception detection". This is because, as for the polygraph, there is no proof that the behaviours and physical traits detected unambiguously indicate deception and not stress, which may accompany deception. The first scenario, poker, was based on a poker study designed at the Institute of Creative Technologies at the University of Southern California. The second scenario was a dice rolling experiment for studying the effects of stress on deceit which was designed and carried out here at the University of Nottingham as part of my PhD. In both scenarios, poker and the dice rolling experiment, action unit detectors together with head pose detectors were able to pick up signs of deceit. In both cases, I used CNN-BLSTM and OpenFace, as they were built differently and might have performance differences and also to see to what extent they corroborate each other, which they partially did. I discovered that the type of deceitful behaviour I detected in the poker dataset, which is a dyadic human-human interaction with a strong social component, was different from that which I detected in the dice rolling experiment, which was a non-dyadic human-computer interaction. In the poker dataset, participants were more likely to exhibit AU12, lip corner puller, just before they folded than when they called or raised. In the dice rolling experiment, which is much less a social interaction, participants who were deceptive exhibited significantly more head motion than

honest participants. In this scenario, AU12, which is associated with social interaction and was prominent in the poker study, nearly disappeared. The dice rolling experiment I presented in this thesis constitutes a unique behavioural economics experiment to study deception under stress. It came to a successful conclusion and showed that stress does modify deceptive behaviour — participants under stress lied less and, simultaneously, their deceptive behaviour proved to be more easily detectable by automatic means. This experiment produced a sizeable database of good quality videos, at the same resolution as the poker dataset, of 373 different individuals. These videos are very descriptively annotated to provide a rich source for further research into the effects of stress on deception.

The two deception scenarios that I have studied here both evoke spontaneous behaviour. In the first, poker, deception is socially acceptable and indeed expected to be part of the game. It was obvious, when watching the videos, that between some participants there was a lot of interaction with a plethora of different facial expressions. This made it seem that it would be difficult to disentangle signs of deception from other social signals. This dataset is most similar to that created with the automatic dyadic data recorder (ADDR) in Sen et al. (2018). In their study, however, no statistically significant differences between deceitful and honest participants were found for the AU12, which was unexpected, as AU12 is an important action unit in Duping Delight, which was a focus of their study that used OpenFace (Ekman, 1985). This outcome is in contrast to the poker study carried out here, in which significant differences were found in the amounts of AU12 and AU5 detected by CNN-BLSTM between people when they folded and when they raised and called. However, I also used OpenFace and similarly did not detect statistically significant differences with these detectors. Although CNN-BLSTM and OpenFace did not agree in this respect, correlation between the two indicated that they were both detecting AU12 correctly to some extent. Sen and colleagues attribute this absence of significant differences for AU12 to a possible failure to elicit the intended Duping Delight in their experiment. This absence might also have been due to their having used OpenFace, which I found lacked sensitivity and for which I also detected no statistically significant differences. Another reason for the failure to find signs of Duping Delight could have been their having used average AU values over longer periods of time causing them to lose important information in the process of their search. Bartlett and colleagues also collected aggregates in the form of histograms over long periods of time and could successfully distinguish between expressions of faked and real pain, but for their study participants

were required to make faked pain expressions for one full minute, which is probably not realistic and differs from a spontaneous scenario (Bartlett et al., 2014). I, however, opted to keep the values intact as time frames were short and behaviour was spontaneous. I then fused individual classification together into a single classification in a late fusion approach as was done in Valstar et al. (2007). I note that I did try aggregate values over longer periods of time (up to nine seconds), as well as decision trees built using means over different time windows. This caused detection rates to drop and significance to be lost (data not shown).

In contrast to the highly social human-human interaction in the poker dataset, the dice rolling experiment was designed to cut down on social signalling in order to focus on the effects of the stress treatment on a form of deception closer to the concept of mal-intent. Here, the participant was given the opportunity to lie to a computer in order to maximize their reward. It was not known in the beginning if the participants would lie to a computer controlling the dice. However, as this experimental software was being built, another work was published that also used computer controlled videos of dice, but in the setting of group decision making (Kocher et al., 2018). In their experiment, which did not involve computer vision or webcams, participants reported the results of dice rolling individually and then in groups. It was found that when they reported individually, they misreported just over 30% of the time (in groups they misreported more often). Participants in the study presented in this thesis were similarly dishonest just over 30% of the time. This seems to be less than reported for the control group in the Fischbacher and Föllmi-Heusi experiment, which used real dice, although in that experiment the proportion of dishonest reporters had to be inferred from the aggregate results as the ground truth was not known (Fischbacher and Föllmi-Heusi, 2013). Given the reduced social interaction in the die experiment presented here, it was also not known what kinds of facial expressions participants would exhibit beforehand. The participants did show a variety of potentially deceptive behaviours and facial expressions, including examples of what looked to me like Duping Delight, negative expressions and head ducking, possibly associated with guilt or shame. However, the level of facial expressions was much lower than in the poker dataset while the level of occlusions was much higher than in the poker dataset and the facial AU detectors, both CNN-BLSTM and OpenFace, did not detect anything meaningful. For this reason, I turned to head pose detection which is computed by OpenFace. To date, head pose has not been as much a focus of automatic human behaviour understanding as facial expressions. It has, however, been shown to be

a relevant source of information about affective states and there have been studies in the past using head together with hand motion for deception detection. Lu and colleagues presented a feasibility study in distinguishing between deceptive and honest subjects by automatically detecting head and hand motion (Lu et al., 2005). They focused on three types of motion to distinguish deceptive from non-deceptive behaviour. Natural motion, which is smooth and relaxed, was associated with honesty; agitated motion, which is abrupt and jerky, was associated with deceit; and over-controlled motion, where the subject moved little in their attempt to suppress signs of agitation, was also associated with deceit. Meservy and colleagues built a deception detector based on the work of Lu and colleagues. This study was also a feasibility study based on a small set of videos of students acting out theft (Meservy et al., 2005). These studies were done 15 years ago when head tracking was more difficult and less reliable. The current head pose estimator in OpenFace is more advanced and, in this thesis, a confirmation of the idea that dishonest people move differently was presented based on head motion alone. I thus propose to give this means of automatic detection more attention when studying deception.

This latter point is given additional weight because it is clear from this and other studies that action unit detectors focused on facial muscles alone, presently have limitations. OpenFace was not sensitive enough for the poker dataset, and much less so for the dice rolling data set. CNN-BLSTM was more sensitive, but when it was used on the dice rolling database, which has facial expressions, albeit very subtle ones, its performance was dominated by false positives caused by glasses, hands on face and head rotations. This was noted as a problem in (Jaiswal, 2018). There have been other works that investigated issues like this. Commercial detectors of basic emotions were investigated in Dupré et al. (2018) and Dupré et al. (2020). Although this approach is slightly different from AU detectors, many of the issues are the same. The authors note that the detectors did not perform as well on spontaneous or naturalistic facial expressions as they did on posed expressions. They noted that detection of some facial expressions was better than others, that the detectors had problems overcoming idiosyncrasies in a persons appearance and that there was a need for more databases containing spontaneous and naturalistic behaviour for expression-detection algorithms to learn from. Recently, Ertugrul and colleagues evaluated how well AU detectors perform on datasets outside the domain in which they were trained (Ertugrul et al., 2019). This is not information that one can glean from the performance measures different AU detectors are reported to have in the literature. That is because AU detectors are always trained and evaluated

on the same dataset, or group of datasets. Therefore, the detectors might have a different performance on a truly independent dataset and this performance on an independent dataset better reflects how they will perform on real-life applications. To carry out their study, Ertugrul and colleagues trained AU detectors on one dataset and then calculated their performance measures on an independent dataset. The two datasets they used were BP4D+ (Zhang et al., 2016), an extended version of BP4D, and GFT, which involves social interaction between groups of three (Girard et al., 2017). As in this thesis, to gain a more general understanding of AU detection, Ertugrul and colleagues compared the performances of two types of classifiers, which they termed 'deep' and 'shallow'. For their deep classifier, they built their own convolutional neural network classifier, similar to CNN-BLSTM which I use, and for their shallow classifier they used the SVM based OpenFace 2.0, as I did. Hence their work is especially relevant to my thesis as their results methodically affirm my impression about the differences between the behaviours of CNN-BLSTM and OpenFace. Their study showed that action unit detectors perform worse in domains that differ from the ones they were trained in and hence their performance values are not as high as reported elsewhere. This decreased performance is frequently below the threshold needed for behavioural research. Their comparison of the two classifier types also led them to conclude that deep methods are more reliable and generalizable than shallow methods, which could explain why CNN-BLSTM seemed to perform better on the poker dataset than OpenFace. Altogether, my experiments and the above mentioned works agree on the need for more varied databases exhibiting spontaneous behaviour to learn AU detectors and detectors of basic emotions. To create robust AU detectors there is still a demand for databases to learn from that "include a large sample of varying ethnic background, age, and sex, that includes people who have facial hair and wear jewelry or eyeglasses, and includes both normal and clinically impaired individuals" (Kanade et al., 2000).

There are possible alternatives to detecting weak, subtle or naturalistic facial signals with out-of-the-box detectors that detect canonical AUs or basic emotions, which was the approach used in this thesis. Interestingly, some of the facial expressions only became apparent to me when viewing the videos in fast forward showing that many facial expressions are hard for humans to detect. To deal with such subtle expressions that cannot be easily spotted by the human eye and also aren't picked up by current action unit detectors, Wu and colleagues created classifiers to detect deception directly from low level pixel information using the real life trial dataset of Pérez-Rosas and colleagues,

which was introduced in Section 3.5 (Wu et al., 2018). They used the hand annotated MUMIN labels that were provided by Pérez-Rosas and colleagues and combined these with Improved Density Trajectory (Wang et al., 2016) to capture motion in videos in order to train micro-expression detectors. The output of these micro-expression classifiers took on the role of action unit detectors in my studies and was likewise used as high-level input to their deception classifier. Their approach was multimodal, but they found that visual micro-expressions, which included both facial muscle actions as well as head motion, were the most effective at classifying deceit. One should, however, remember that their ground truth was the trial verdicts and so not necessarily known with absolute certainty. Additionally, the video segments were hand picked, which might introduce bias as to what represents signs of deceit, as illustrated in Figure 3.9. In another recent study using the same trial dataset, Ding and colleagues applied a deep learning approach (Ding et al., 2018). They also used the three modalities – visual, audio and transcripts – and found that facial expressions together with body motion were the most discriminative. To deal with the small size of the dataset (121 videos) they used meta learning (Santoro et al., 2017) and adversarial learning (Goodfellow et al., 2014) to augment their data set. These methods could also provide ways forward with the dice rolling dataset. The study by Ding and colleagues did not even rely on facial annotations. This points to the interesting possibility of using deep learning to study deception in the dice rolling experiment, which has a strong ground truth, and then using transfer learning to adapt the classifier to new, but related domains. Perhaps it would be possible to augment current AU detectors this way and circumvent the need for FACS coding, which is a major obstacle that limits the datasets that are available for learning general purpose detectors.

In this work an automated method was used to detect the impact of stress on lying. This, to my knowledge, is the first time such a technique was applied. I have found that stress reduces lying and makes it more detectable, which also constitute novel and possibly groundbreaking findings. There have, however, been numerous studies on the effects of stress on decision making and it has been shown that the decision making parts of the brain are sensitive to stress (Starcke and Brand, 2012; Youssef et al., 2012). Schwabe and colleagues showed that the application of stress reduced their study subjects' abilities to think in a goal directed way and caused them to rely more on habit than their unstressed counterparts (Schwabe and Wolf, 2009). Since lying is thought to be cognitively difficult, it seems likely that stress can alter lying behaviour (Gombos, 2006). Lying is also related to moral behaviour and it has been hypothesized that moral deci-

sions are based on intuitive and fast moving emotion and affect which take place quickly and automatically (Haidt, 2007). According to this theory, it requires effort to overcome initial moral reactions and these automatic affective responses are in conflict with more time-consuming reasoned considerations that take place later (Greene and Haidt, 2002). Hence, it is reasonable to think stress might further interfere with this process and there have been studies to this effect. It has been proposed that stress acts like a switch that allows intuitive responses to bypass reasoning (Yu, 2016). The effects of stress on moral decision-making are complex and depend on factors such as gender and types of stress, such as chronic or acute, but the body of evidence suggests that stress leads to a lessening of what could be perceived as amoral behaviour, that is, it leads to behaviour that is less utilitarian and more deontological (Zhang et al., 2018; Vveinhardt et al., 2020). This could explain why stress treated participants in the dice rolling experiment lied less as it might be their first intuitive reaction to be truthful (Abeler et al., 2019). Regarding observable behaviour, it has been proposed that deception causes changes in arousal as well as negative emotions associated with guilt and shame (Burgoon and Buller, 1996). Therefore, subjecting participants to additional stress might increase their need to fidget while decreasing their ability to conceal this, possibly resulting in increased head movement, as detected here, as well as increased face touching which caused more occlusions.

Furthermore, I found a strong gender effect on lying behaviour with men submitting significantly more misreports. Gender effects in lying behaviour are well-known, and the results here are in line with current experimental evidence showing that women tend to lie less than men, particularly when strategic, selfish lying is involved to maximize personal benefit (Dreber and Johannesson, 2008; Erat and Gneezy, 2012; Conrads et al., 2013; Chen et al., 2020). Of note, these findings are based on dyadic human interactions, whereby the gender makeup of the dyad did not fundamentally change the overall outcome, but influenced the amount of lying (Jung and Vranceanu, 2017). Here, in contrast, humans are interacting with a computer. Importantly, these patterns seem to be robust and persist in the presence of a human-computer interface. To my knowledge, this is the first time gender effects have been linked to deceptive behaviour in computer-human interactions.

There is more research left to do with the dice rolling dataset. Aside from the self assessment manikin, the questionnaires have not been evaluated. The mouse tracking has also not been investigated and the lying patterns need to be considered, too. There have

been many studies based on the Fischbacher and Föllmi-Heusi experiment that represent variations on their dice rolling paradigm, which study the nature of deceptive behaviour (Gächter and Schulz, 2016; Conrads et al., 2013; Bucciol and Piovesan, 2011). The previously mentioned study by Kocher and colleagues uses computer controlled videos of a die being rolled. They report in their paper that they expect less partial lying, meaning the participants are expected to be either fully honest or maximally lying. They explain that this is because the participants are aware of the experimental set-up and know that the computer controls the dice rolling and that the experimenter knows their answers. The paper does not give further detail on how the participants can know this. Partial lying does actually occur infrequently in their experiments. In the dice rolling experiment presented in this thesis a similar set-up was used. I have also explained how I intentionally tried to make the dice rolling experiment feel realistic in the hope that then the participants would lie more and partial lying took place. At present, the role the computer interface plays in the process of decision making is poorly understood. Yet, there are works that suggest that this aspect can have an important influence on the outcomes of human deliberations when interacting with computers. One such work deals with embodied conversational agents and another with immersive virtual reality systems. They posit that part of how a person reacts to a computer is dependent on their own ability to suspend disbelief (McKeown, 2015) and part of it is the ability of the system to share common ground with and include the participant while maintaining plausibility (McKeown, 2015; Slater, 2009). While the experiment presented here did not involve an embodied virtual agent or immersive virtual reality it was a human-computer interaction where engagement was important. The interface, as well as the participant's attitude towards the computer interaction, might also play a part with respect to determining why people decide to lie or not, so it might be important to consider this in the future. Concerning another aspect of the experiment, as the participants each performed a series of multiple rolls, which was not done in other dice rolling studies, interesting patterns of lying can emerge in ways they might not have before, as some patterns suggest an attempt to disguise lying while others suggest an aversion to appearing greedy. Through combining behavioural experimentation with computer vision approaches and mouse tracking data, the die rolling experiment should therefore continue to be a valuable source for further discoveries about human conduct and automatic detection of deceit.

Appendices

Appendix A

Poker study: Supplementary data

A.1 Feature selection

In Chapter 4, feature selection was performed on the decision trees. The following table gives the complete list of tuples that the algorithm iterated through in its heuristic search for the best tuple. The Column 1 gives the tuple over which the current decision trees are made, Column 2 gives the tuple's best balanced classification rate, Column 3 gives the balanced classification rate of the tuple after voting, Column 4 gives the offset and Column 5 the duration of the tree with the best balanced classification rate. In each row, the largest value between the classification rate without voting and with voting is highlighted. One can see that voting improves results in the majority of cases.

Table A.1: tuples chosen by feature selection along with their balanced classification rate before and after voting, and offset/duration parameters. The largest balanced classification rate, with or without voting, is highlighted blue.

AU tuples	**bal. class**	**with voting**	**offset**	**duration**
(1,12)	0.5869	0.5982	-5	3
(5,12)	0.5790	0.6115	-30	3
(9,15)	0.5884	0.5972	5	3
(1,25)	0.5886	0.6163	30	3
(15,25)	0.5757	0.6026	-45	5
(25,26)	0.5750	0.5721	-35	3

Continued on next page

Table A.1 – *Continued from previous page*

AU tuples	bal. class	with voting	offset	duration
(1,2,12)	0.5765	0.5971	-5	3
(2,5,12)	0.5769	0.5954	10	3
(5,6,12)	0.5814	0.6103	-25	5
(5,9,12)	0.5808	0.6173	-30	4
(5,12,25)	0.5788	0.5921	-25	3
(9,15,25)	0.5765	0.6071	-35	5
(15,25,26)	0.5771	0.5850	-35	3
(1,2,12,45)	0.5829	0.5897	-5	3
(1,2,5,12)	0.5835	0.5948	0	4
(1,5,9,12)	0.5919	0.6117	0	5
(4,5,12,25)	0.6011	0.6193	-30	3
(5,9,12,20)	0.5836	0.5979	5	3
(5,9,12,26)	0.6114	0.6081	-30	4
(5,9,12,45)	0.5936	0.6053	-30	4
(1,4,5,12,25)	0.5848	0.5808	-30	3
(1,5,6,9,12)	0.5877	0.5828	-30	3
(1,5,9,12,45)	0.5948	0.6193	0	3
(2,4,5,12,25)	0.5904	0.6098	-30	3
(2,5,9,12,45)	0.5802	0.5841	-60	4
(4,5,6,12,25)	0.5810	0.5801	-10	3
(4,5,9,12,20)	0.5997	0.5941	-30	3
(4,5,9,12,45)	0.5818	0.6021	-30	3
(4,5,12,15,25)	0.5907	0.5865	10	3
(5,9,12,25,45)	0.5861	0.5986	-25	4
(5,9,12,26,45)	0.5939	0.6131	-25	3
(1,4,5,6,12,25)	0.5910	0.5997	-25	4
(2,4,5,9,12,20)	0.5996	0.6222	-30	4
(2,4,5,12,20,25)	0.5930	0.6227	-30	3
(4,5,6,9,12,25)	0.6012	0.6109	-30	5
(4,5,6,12,25,45)	0.5869	0.5892	-30	4
(4,5,9,12,20,26)	0.6014	0.6149	-30	3
(4,5,9,12,20,45)	0.6060	0.6079	-30	3

Continued on next page

Table A.1 – *Continued from previous page*

AU tuples	**bal. class**	**with voting**	**offset**	**duration**
(4,5,12,15,20,25)	0.5993	0.5987	10	3
(5,6,9,12,25,45)	0.6023	0.6026	-25	3
(5,9,12,15,25,45)	0.5905	0.5899	-30	4
(1,4,5,9,12,20,26)	0.5852	0.5877	-30	4
(2,4,5,6,9,12,25)	0.5885	0.5914	-30	5
(2,5,6,9,12,25,45)	0.5850	0.5786	-30	5
(4,5,6,9,12,15,25)	0.5931	0.6063	-30	5
(4,5,6,9,12,25,45)	0.6005	0.6119	-30	5
(4,5,6,12,25,26,45)	0.5869	0.5933	-10	3
(4,5,9,12,15,25,45)	0.6014	0.6365	-30	4
(4,5,9,12,20,25,45)	0.5905	0.6189	-30	3
(4,5,9,12,20,26,45)	0.5878	0.5960	-30	3
(4,5,12,15,20,25,45)	0.6049	0.6068	10	3
(5,6,9,12,15,25,45)	0.5869	0.5859	-30	5
(5,9,12,15,20,25,45)	0.5885	0.5947	-30	4
(1,4,5,6,9,12,20,26)	0.5876	0.5877	-25	3
(1,4,5,9,12,15,20,26)	0.5869	0.5980	-30	4
(1,4,5,9,12,15,25,45)	0.6042	0.6299	-30	4
(1,5,6,9,12,15,25,45)	0.5866	0.5714	-30	4
(2,4,5,6,9,12,20,25)	0.5862	0.5923	-30	5
(2,4,5,6,9,12,25,45)	0.5851	0.5938	-30	5
(2,4,5,9,12,15,25,45)	0.5926	0.5985	-30	4
(2,4,5,9,12,20,25,45)	0.5886	0.6122	-30	3
(4,5,6,9,12,15,25,45)	0.5855	0.6212	-30	3
(5,9,12,15,20,25,26,45)	0.5915	0.6032	-30	4
(1,2,4,5,9,12,15,20,26)	0.5893	0.5963	-25	5
(1,2,4,5,9,12,15,25,45)	0.5920	0.6009	-30	4
(1,4,5,9,12,15,20,25,26)	0.5802	0.5668	-40	3
(1,4,5,9,12,15,20,25,45)	0.5878	0.6067	-30	4
(2,4,5,6,9,12,20,25,45)	0.5884	0.5928	-30	5
(2,4,5,9,12,15,20,25,45)	0.5826	0.5863	15	3
(4,5,6,9,12,15,20,25,45)	0.5845	0.6111	-30	3

Continued on next page

Table A.1 – *Continued from previous page*

AU tuples	**bal. class**	**with voting**	**offset**	**duration**
(1,2,4,5,6,9,12,15,25,45)	0.5821	0.6022	15	3
(1,2,4,5,9,12,15,20,25,45)	0.5817	0.5872	-30	4
(1,4,5,6,9,12,15,25,26,45)	0.5888	0.5966	15	3
(2,4,5,6,9,12,15,20,25,45)	0.5896	0.6040	15	3
(1,2,4,5,6,9,12,15,20,25,26)	0.5926	0.6004	15	3
(1,2,4,5,6,9,12,15,20,25,45)	0.5817	0.5865	15	3
(1,2,4,5,6,9,12,15,25,26,45)	0.5834	0.6037	15	3
(1,2,4,5,9,12,15,20,25,26,45)	0.5943	0.5917	15	3
(1,2,5,6,9,12,15,20,25,26,45)	0.5801	0.5895	15	3
(1,4,5,6,9,12,15,20,25,26,45)	0.5824	0.5898	15	3
(2,4,5,6,9,12,15,20,25,26,45)	0.5809	0.5876	15	3
(1,2,4,5,6,9,12,15,20,25,26,45)	0.5938	0.6079	15	3

A.2 Comparison of CNN-BLSTM and OpenFace statistics

To explore the differences between CNN-BLSTM and OpenFace detectors, I looked at their statistics. For each of the 64 players in the Poker database and for each of the 12 action units used in the poker study, I found the minimum, maximum, median, mean and mode as well as their variance/standard deviation, first and third quartiles values and interquartile range. These values were then plotted in a box plot.

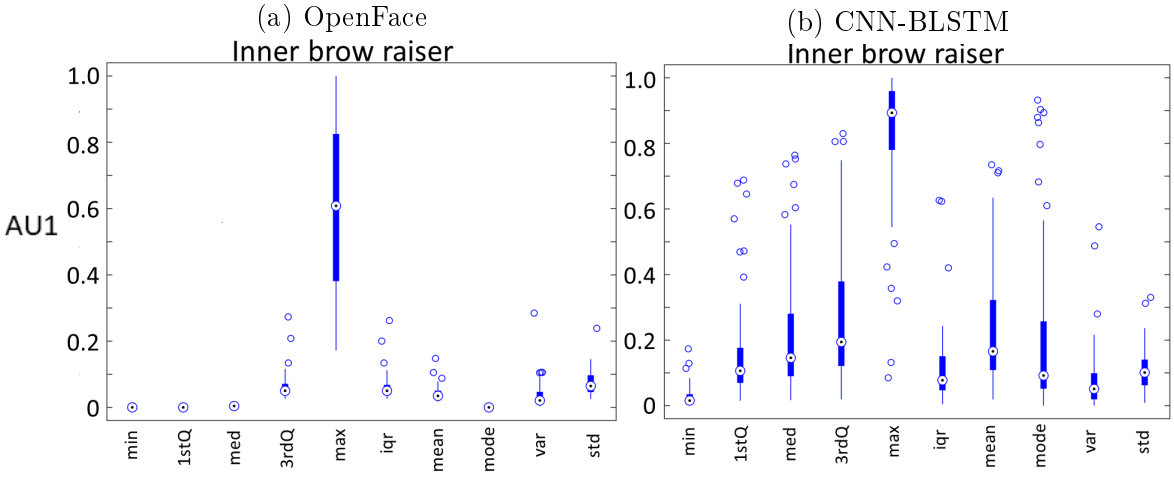

Figure A.1: Comparing statistics for OpenFace and CNN-BLSTM.

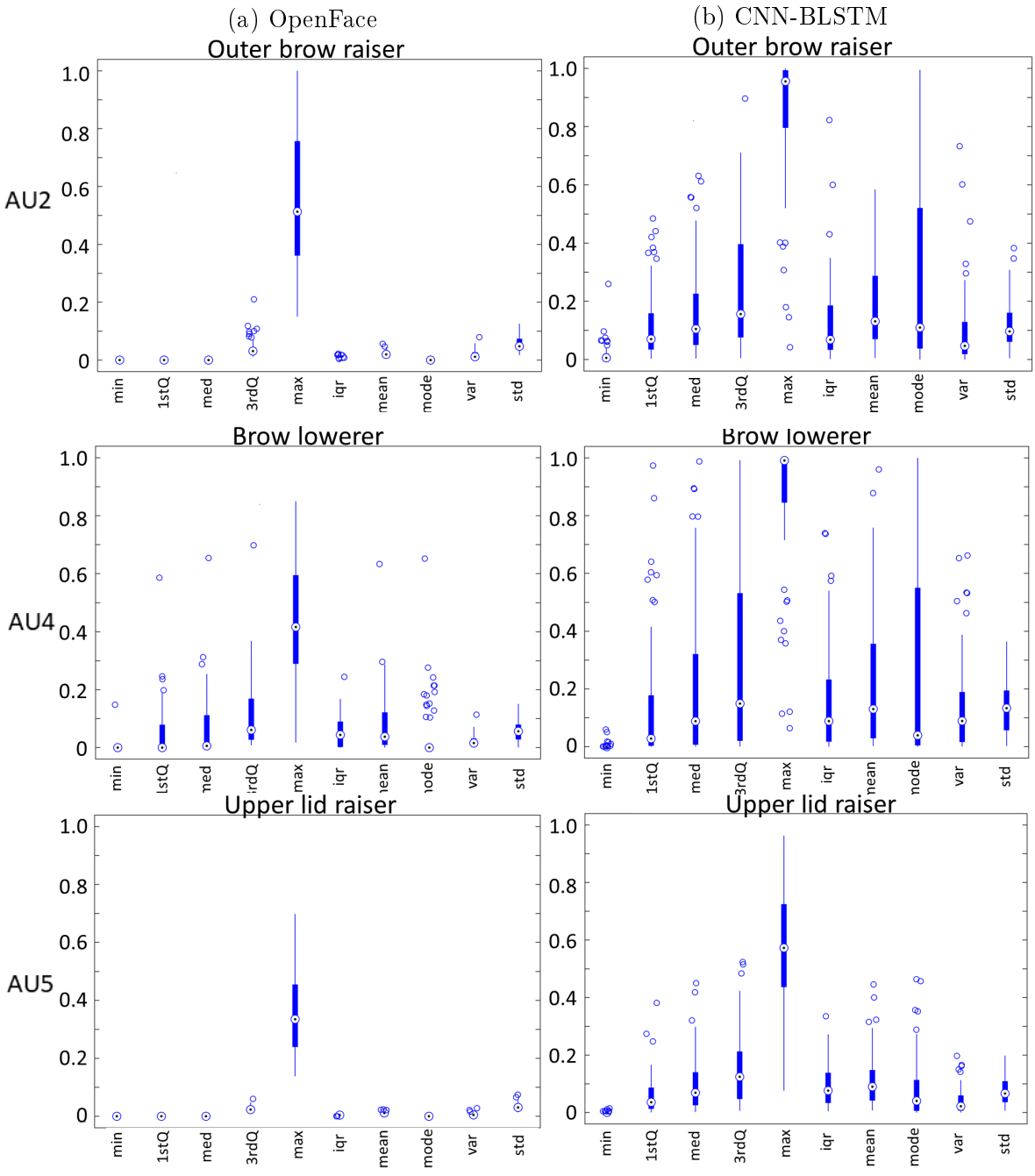

Figure A.3: Comparing statistics for OpenFace and CNN-BLSTM for AUs 2, 4 and 5.

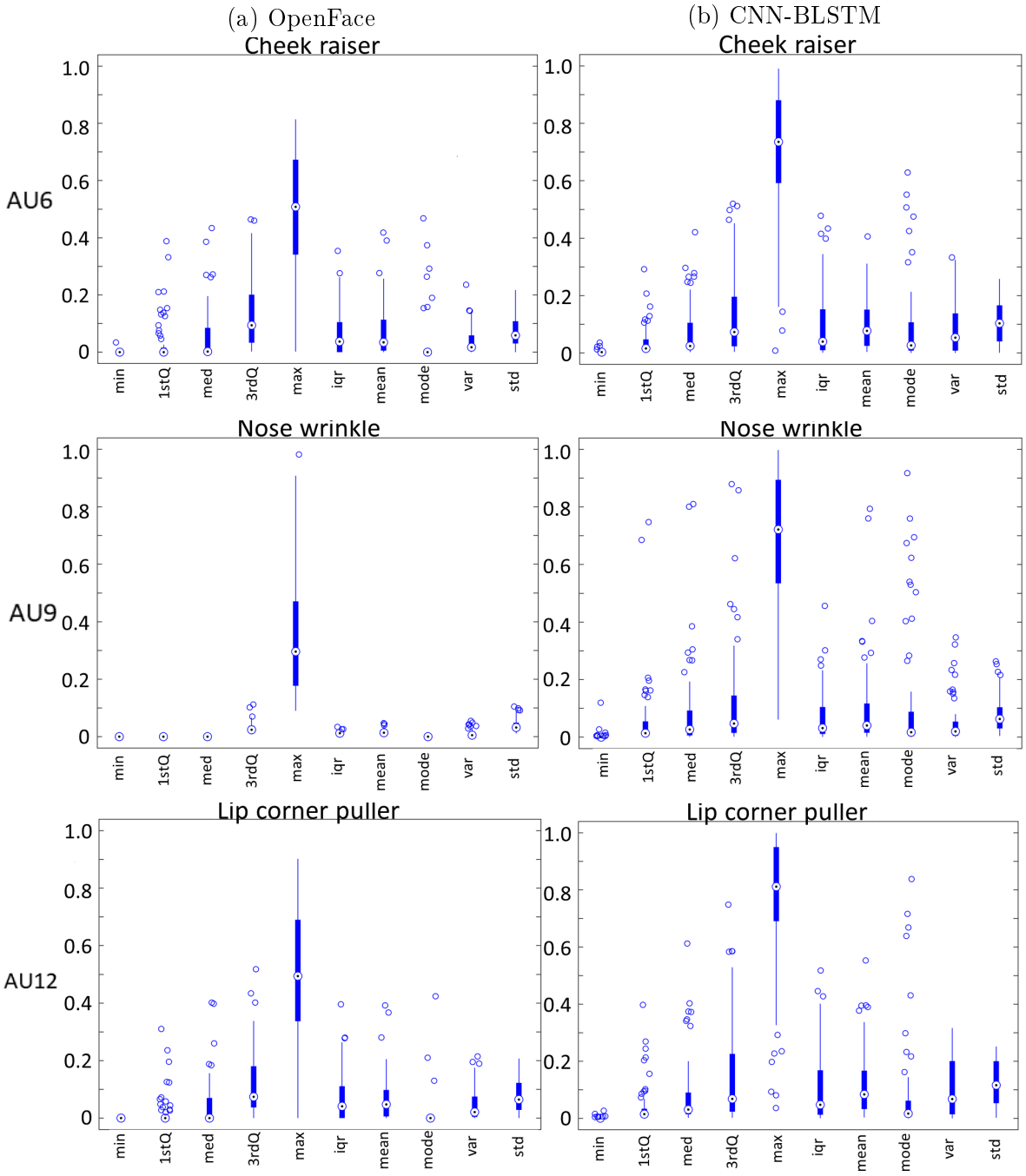

Figure A.5: Comparing statistics for OpenFace and CNN-BLSTM for AUs 6, 9, 12.

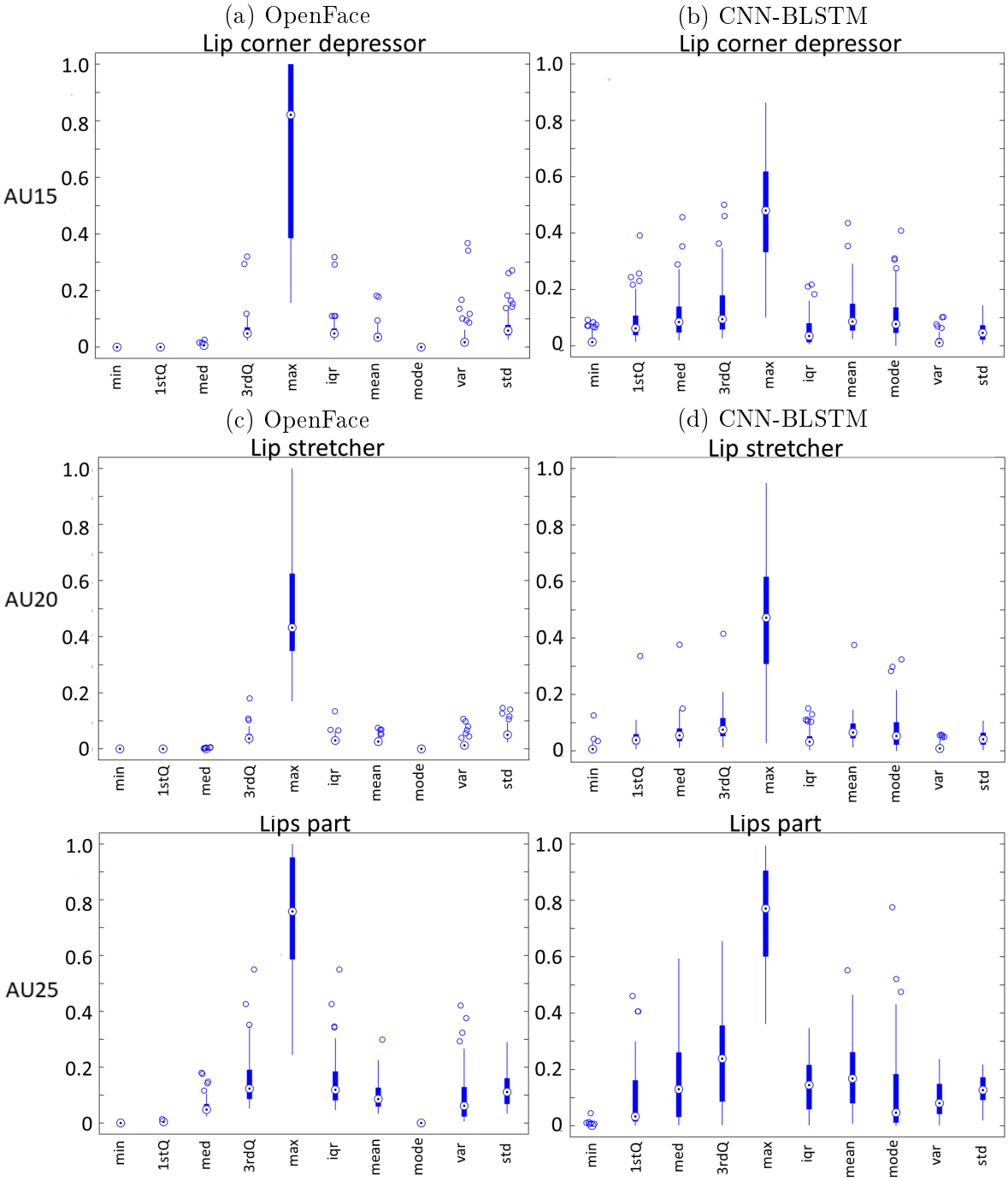

Figure A.7: Comparing statistics for OpenFace and CNN-BLSTM for AUs 15, 20 and 25.

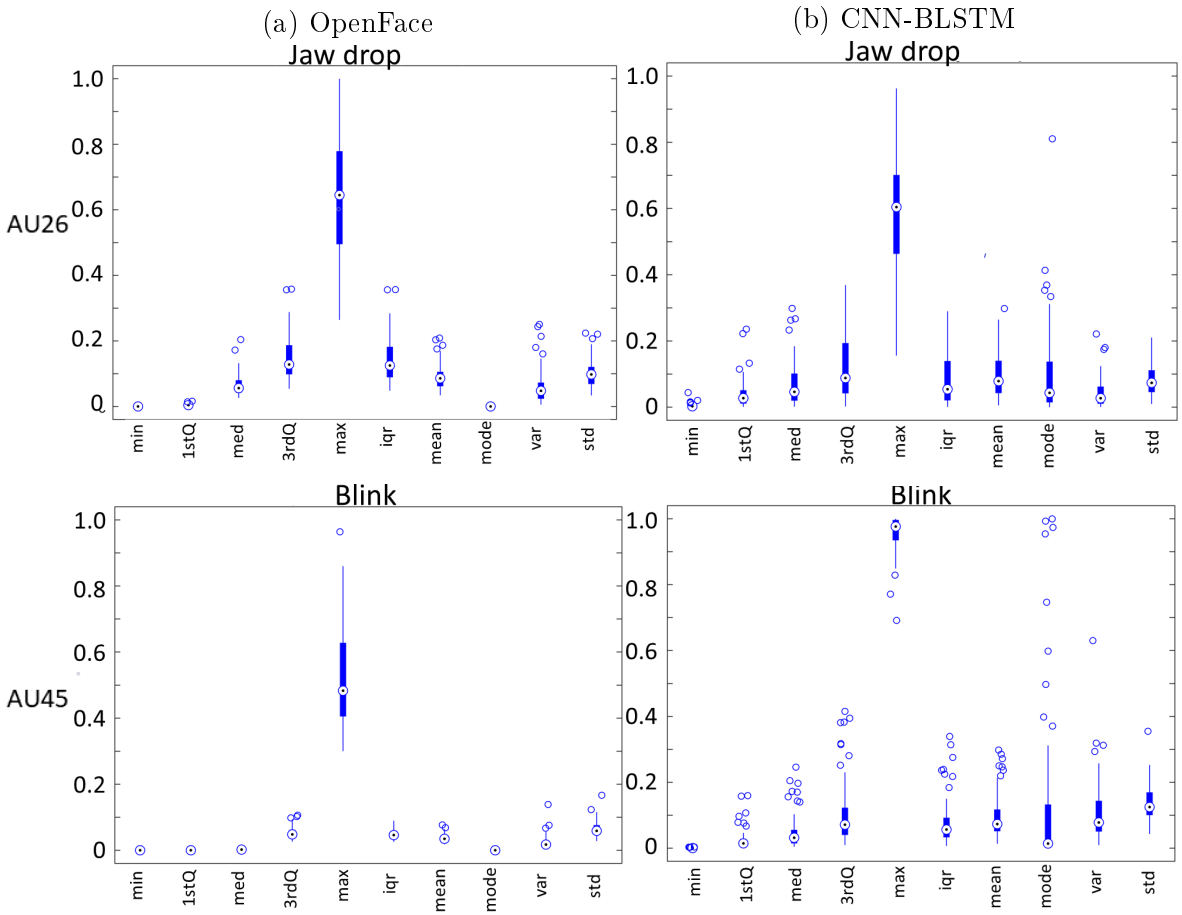

Figure A.9: Comparing statistics for OpenFace and CNN-BLSTM for AUs 26 and 45.

Appendix B

Die experiment: Supplementary materials and methods

B.1 Protocol for running the experimental sessions

1. **Set-up** for a day of sessions. There were at most four sessions on one day. Each session could accommodate up to ten participants.

   (a) **The night before experiments**, check that there is enough free memory on the computers. Make sure software is running. Check that the network computer in the anteroom is detecting the experiment computers. For the water experiments, move 20 chairs to the front of the room and tape numbers on them in pairs so each PC used (ten from PCs 2, 5, 6, 9, 12, 15, 16, 19, 22, 25, 26, 29) has a pair of chairs, one for the participant to sit on, and one, to the left, for the participant's water bucket to sit on. Get questionnaires ready for the next day, labelled by PC, and set first 10 on tables for the first sessions of ten. Get PC card numbers ready in their cotton bag. If it is a cold water treatment, fill the 14L-buckets just under half way with water and set buckets on their chairs. The water is taken from a shower down the hall and is warmer than room temperature. Doing this the night before allows the water to cool down to room temperature (18°C) overnight. Also, the buckets should not be too full, otherwise, adding a 2kg bag of ice to it directly before a session will not cool it down to 1-2°C. If it is a warm water treatment,

leave the buckets empty to be filled the next morning. Get 30-40 hand towels and kitchen towels ready for participants to dry their hands. The number of hand towels needed depends on how many sessions there will be on this day. Towels were brought from home.

(b) **At least 2hrs before the experiments**, for cold water experiments bring ice to the lab. For each participant a 2kg bag is required, bought from local grocery store on the way to the lab. As a safeguard, a few extra bags were bought. Keep them as cold as possible in cardboard boxes with bubble wrap insulation. For all treatments, turn on ten computers and have them waiting for the participants, with the dice rolling program minimized. Put an information sheet and consent form on the keyboards of the PCs. Put a large envelope with the questionnaires in the back left corner of the tables with the 10 PCs. For warm water experiments, warm up shower water to about 45°C as it takes a while to reach this temperature and fill buckets up over half way to prevent them from cooling down too quickly. Set them on their chairs and put thermometers in them and lids on them to slow their cooling down. Using an electric kettle and a cup, keep their temperatures at 39-40°C: when the temperature drops in a bucket of water, scoop out a few cups of water, boil them in the kettle and pour them back in.

(c) **Just before the experiments** the second of two experimenters arrives. For cold water experiments, a 2kg bag of ice is dumped into each bucket and stirred. Thermometers are checked to see that cold water is between 1-2°C, or warm water is 39-40°C. Right before the experiments, ice is stirred and removed with sieves for cold water (there might still be a few pieces in the water) and put in an extra container out of sight. Thermometers are removed for both treatments. Now, warm water should be 39-40°C and cold water should be <2°C. This ensures that when the treatment begins approximately 15 minutes later the water will be 37-39°C for warm water treatments and <4°C for cold water treatments.

2. **Carrying out the experiment**

(a) **Seating the participants.** Experimenter 1 opens the door of the anteroom and greets the participants. They ask to see each student's ID and compare that to the list of participants and instruct the participants to go to the door of

the lab, where Experimenter 2 is waiting. Experimenter 2 is holding a cotton bag and asks each participant to draw a card out of the bag and go to the PC at the table with the number of their drawn card. They are told they can read the consent form and instructions (on paper, stapled together) on their keyboard, but they should not do anything with their computer until they are instructed to. They are told not to do anything with the PC at their table to prevent them from starting the experiment before they undergo the treatment. At this point their computer screen has the windows desktop visible, see Figure B.1.

(b) **Giving the participants instructions.** When all of the participants have sat down, Experimenter 1 explains that

i. They will do an experiment on their computer.

ii. The computer will be recording them. The webcam is visibly mounted on their monitor.

iii. Any images made of them are for research purposes only and won't appear anywhere.

iv. They are there on a voluntary basis and can withdraw at any time.

v. They should read the instructions (Figure B.10) and consent form (Figure B.11) and sign the consent form to continue.

vi. They are told if they have any questions during the experiment, they should raise their hand.

(c) **The water treatment**

i. After 5-10 minutes when consent forms have been signed (which is 10-15 minutes after the begin of the experiment), Experimenter 1 tells the participants they need to come to the front and sit in the chair with the same number as their PC. They should wait there until they are told to put their left hand in the bucket of water to their left, unless they operate their mouse with their left hand, then they should reach over and put their right hand in the bucket to their left. They are told it is just water with nothing added and it is not unhealthy.

ii. When they are seated, Experimenter 1 asks the participants to roll their left sleeves or their right sleeve if they operated the mouse with their left hand. When Experimenter 2 has the timer ready for three minutes, the participants are told to submerge their hand including the wrist, into the

water. At the signal, Experimenter 1 tells the participants to put their hands in the water and leave them there until they are told to remove them. During cold water treatments participants frequently try and remove their hand from the cold water. Experimenter 1 asks them to keep it in the water. If they have to take it out they are instructed to put it back in as soon as they can.

iii. The participants are not informed how long they have to keep their hands in the water.

iv. During these three minutes, Experimenter 1 goes to the network controller in the anteroom and maximizes the die rolling software so it is waiting for the participants when they return.

v. When the three minutes are up, a beeper goes off. Experimenter 1 tells the participants to remove their hands from the water, dry themselves with the towel by their buckets, return to their desks and follow the instructions on their screen.

(d) **The dice rolling experiment**

i. When they return to their desks, the participants are instructed by their screen, see Figure B.2, to press 'start' to confirm that they have read their instructions and signed their consent form. The next screen, Figure B.3, prompts the participant to roll the die by pressing the 'Roll Die' button. One second after pressing the 'roll die' button, a randomly chosen video of die being thrown is shown. When this short video (2.5 seconds) is over, buttons become activated so the participant can report the number the die has rolled, see Figure B.5. These two steps, roll and repeat roll, is repeated until the participant has rolled 20 times, although they are not told how many times this is.

(e) **The questionnaires**

i. After the 20^{th} iteration of the dice rolling experiment, the final screen, see Figure B.6, appears and informs the participant of their reward, which is the sum of their individual rewards, and it instructs them to fill out the questionnaires on their table. The experimenters can tell when the dice rolling experiment is over since the clicking stops and they can hear the envelopes that contain the questionnaires open. After about ten minutes, Experimenter 1 checks to see if everyone is done with the

questionnaires.

ii. While the participants are filling out the questionnaires, Experimenter 2 puts monetary payments (reward + participation fee) in envelopes which are then sealed and marked with the corresponding PC number.

(f) **End of experiment and reward collection**

i. When everyone has finished filling out their questionnaires, Experimenter 1 tells the participants they will call out a PC number. The person sitting at that number should collect their belongings, their questionnaire, consent form and PC card and go to the front (anteroom) to collect their payment.

ii. When the participant is called, they come into the anteroom, leave their consent form, questionnaire and PC card and collect their payment. When they have left, the next participant is called to collect their payment.

3. **After session/between sessions**

(a) Two full hours are needed between sessions. Software and memory need to be checked, and software needs to be started up for the next sessions. Questionnaires and PC cards need to be prepared. In cold water treatments, excess water from melted ice needs to be emptied out to keep the volume down. For warm water treatments, water needs to be heated up in the kettle, just as in part 1.(c) **just before the experiments**.

(b) If it is the end of a day of sessions, data needs to be copied and stored and memory freed on the lab computers as they have memory limits. Everything is cleaned up and shut down.

B.2 Interface

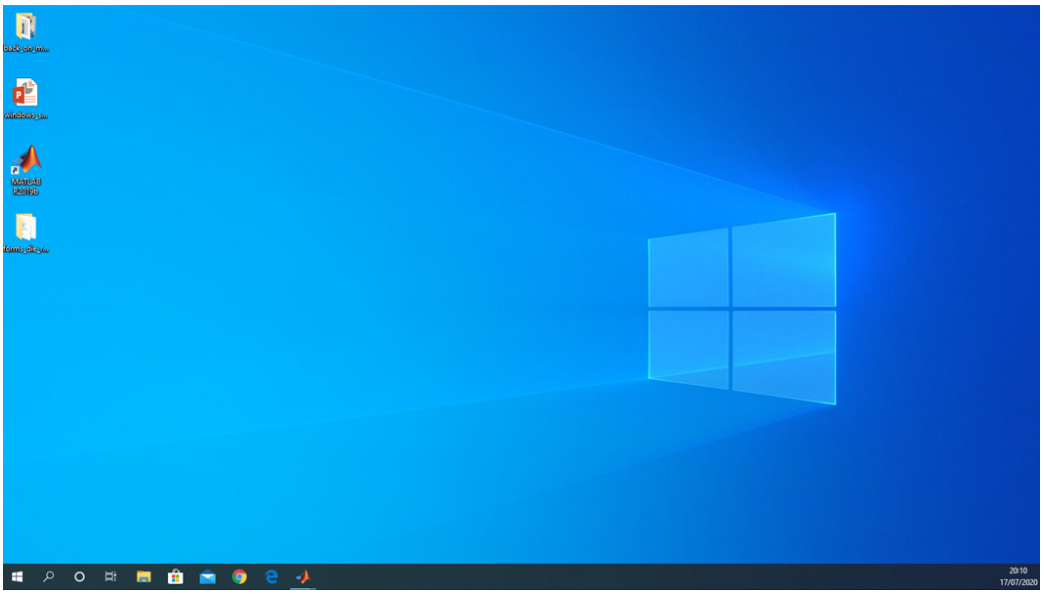

Figure B.1: The screen when the participant arrived at their PC.

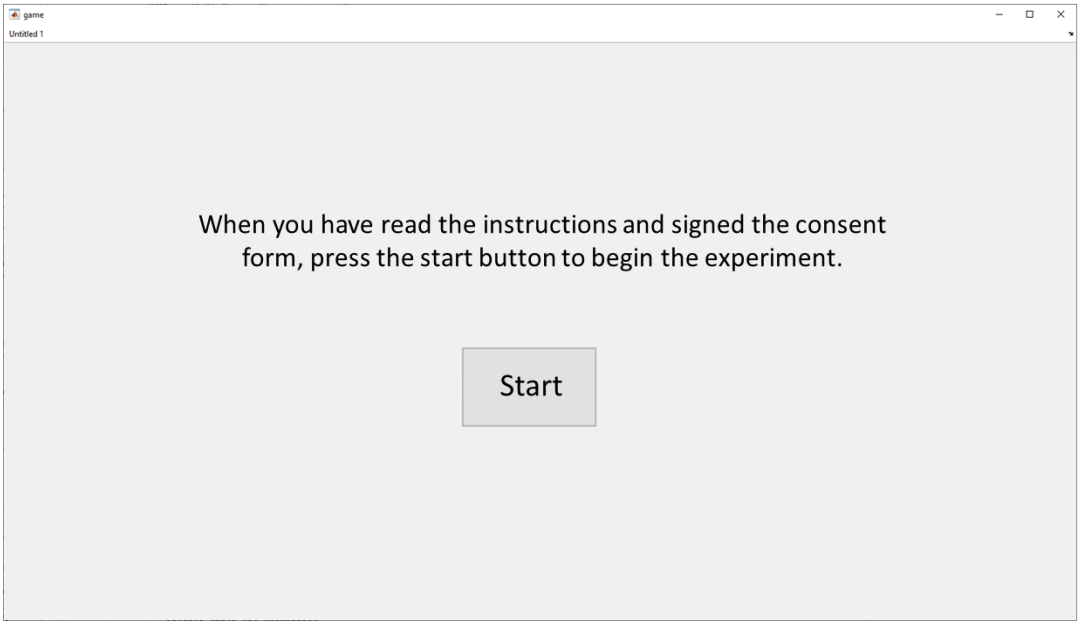

Figure B.2: When the participant returned to their PC after the water treatment, this was on their monitor prompting them to begin the experiment.

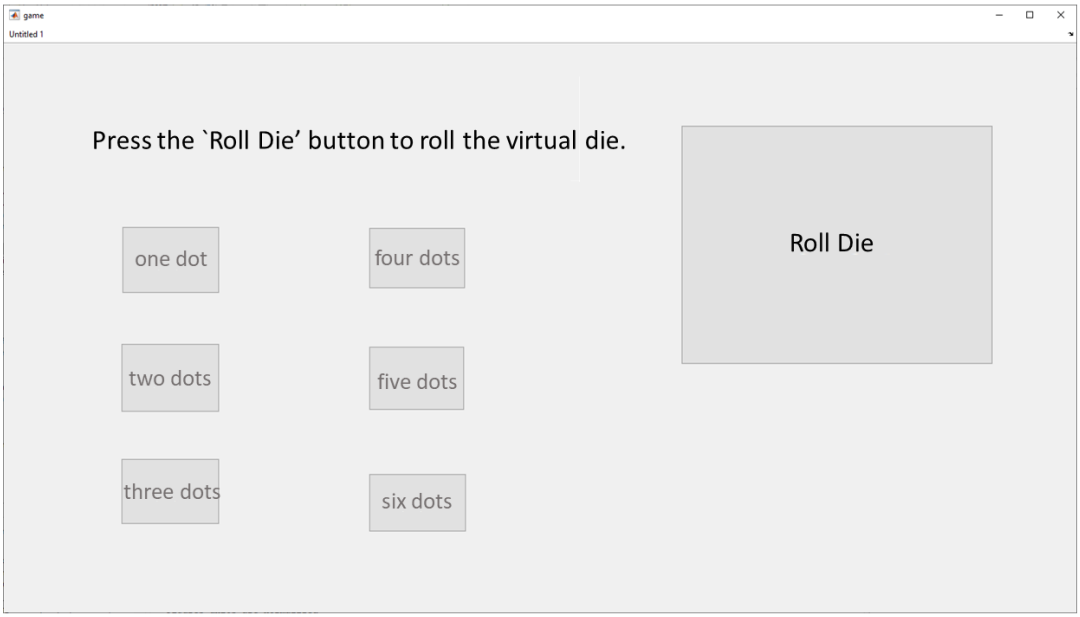

Figure B.3: After pressing start or reporting what they rolled, the participant was prompted to roll the die. This happened twenty times.

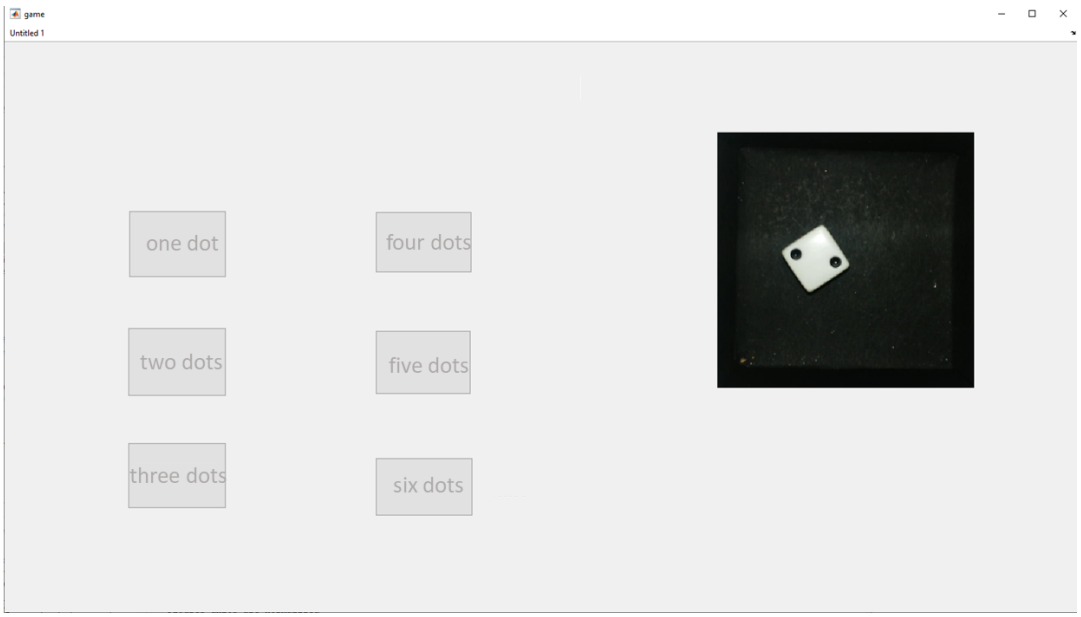

Figure B.4: After pressing the 'Roll Die' button, the participant was shown a randomly selected video of a die being rolled.

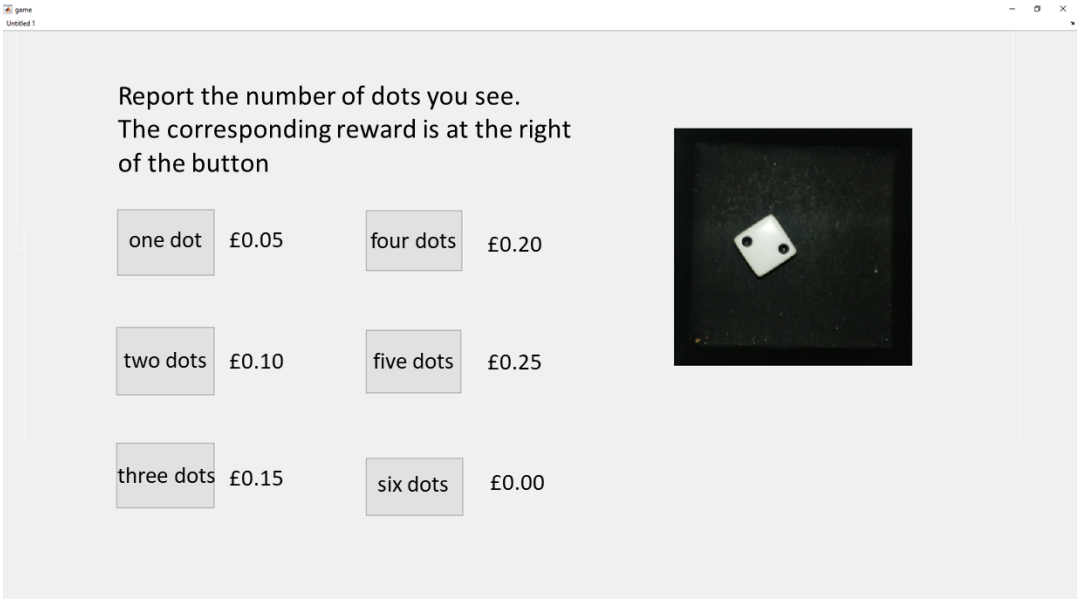

Figure B.5: After the video ended, the reward buttons were activated and the participant was prompted to report what they rolled.

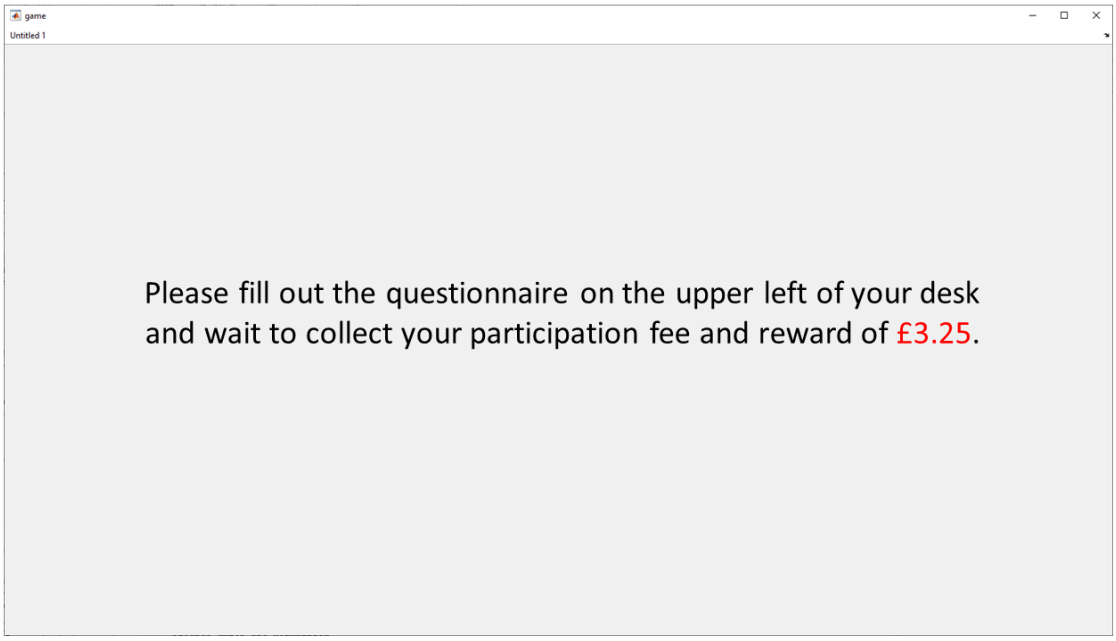

Figure B.6: After the participant had cycled through Figures B.3, B.4 and B.5 20 times, they were shown their reward and asked to fill out the questionnaire on their table.

B.3 Questionnaires

The questionnaire consisted of four different parts: Parts 1) CRT and 2) MACH-IV are as described (Frederick (2005); Exline et al. (1970)). Parts 3) (SAM), an extension of Bradley and Lang (1994) and 4) demographic questions are shown below.

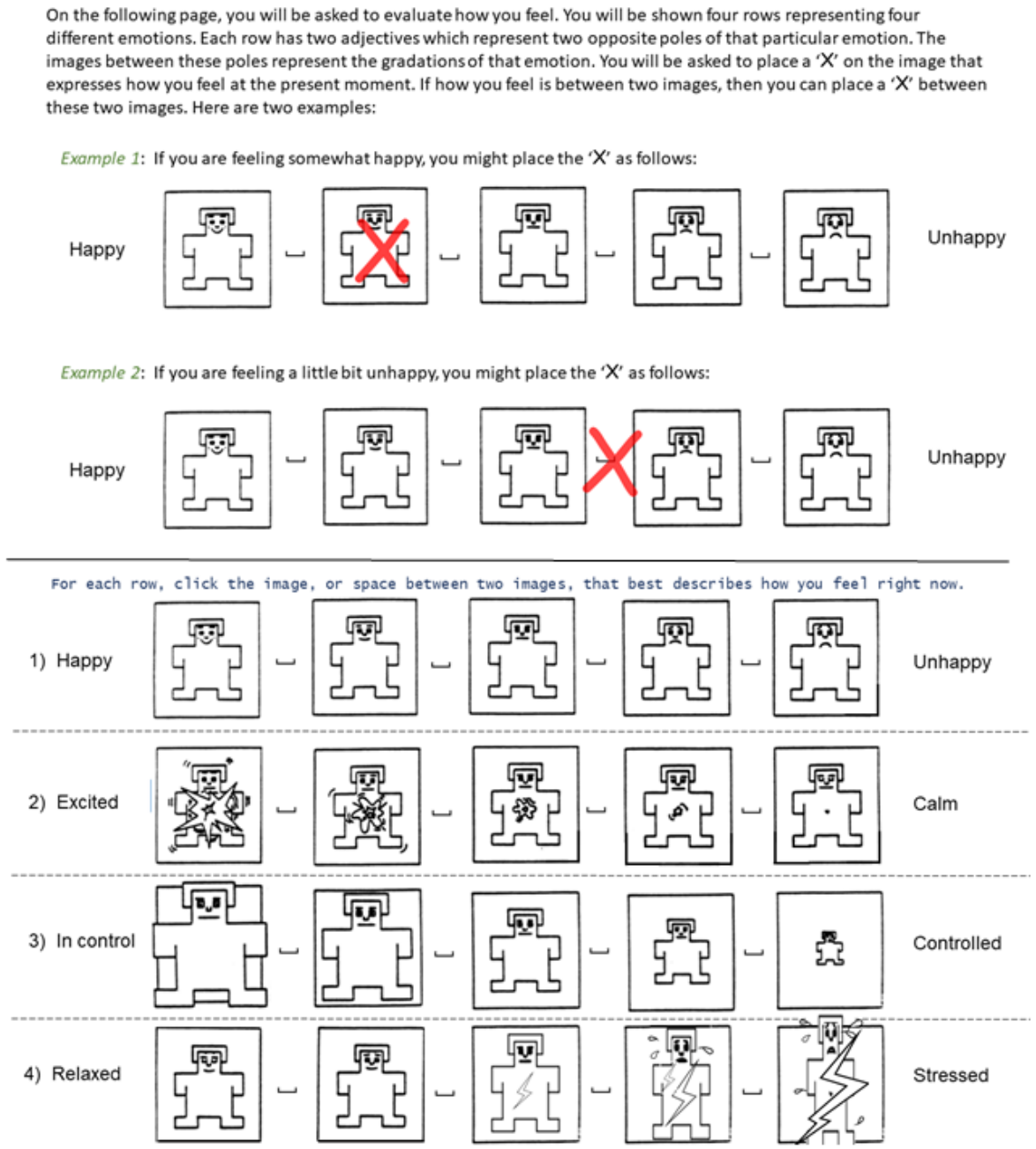

Figure B.7: SAM. Pole names for 1) - 3) are from (Lombard et al. (2000)). The fourth manikin is new. Numerical values 9–1, left to right, as in Figure 5.5.

Please answer the following demographic questions:

What is your gender?

- ☐ Male
- ☐ Female
- ☐ Other

What is your age in years? _____

What University School do you belong to? _____

How large was the community where you lived the most time of your life?

- ☐ Up to 2,000 inhabitants (1)
- ☐ Between 2,000 and 10,000 inhabitants (2)
- ☐ Between 10,000 and 100,000 inhabitants (3)
- ☐ More than 100,000 inhabitants (4)

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

Please select a number on a scale, where the value 0 means: 'not at all willing to take risks' and a value 10 means: 'very willing to take risks'.

- ☐ 0 – not at all willing to take risks (1)
- ☐ 1 (2)
- ☐ 2 (3)
- ☐ 3 (4)
- ☐ 4 (5)
- ☐ 5 (6)
- ☐ 6 (7)
- ☐ 7 (8)
- ☐ 8 (9)
- ☐ 9 (10)
- ☐ 10 – very willing to take risks (11)

Which hand do you use to operate your computer mouse?

- ☐ my right hand
- ☐ my left hand

Figure B.8: Demographic questionnaire page one of two.

☐ either hand (ambidextrous)

Which of the following religious denominations describes you better?

☐ No religion (1)
☐ Catholic (2)
☐ Protestant (3)
☐ Muslim (4)
☐ Orthodox (5)
☐ Eastern Religion (6)
☐ other denomination (7)

How religious are you??

☐ extremely non-religious (1)
☐ (2)
☐ (3)
☐ (4)
☐ (5)
☐ (6)
☐ extremely religious (7)

Would you describe yourself as politically on the "left" (e.g. a liberal) or on the "right" (e.g. a conservative)?

☐ Very liberal (1)
☐ Liberal (2)
☐ Center (3)
☐ Conservative (4)
☐ Very conservative (5)

What country are you from?

_____

Approximately how many Facebook friends do you have?

_____

Figure B.9: Demographic questionnaire, continued, page two of two. The question "What country are you from?" was added half way through the experiment.

B.4 Instructions and consent form

The Instructions and Informed Consent Form were stapled together and sitting on the keyboard of the PC when the participant arrived.

Instructions

Step 1. Consent

Before proceeding with this experiment, please read these instructions and fill out the **Informed Consent Form** on the next page.

Step 2. Water Treatment

You will be asked by the organiser to come to the front and sit in the chair labelled the same number as your PC. He will indicate when you should submerge your left hand in the container of water to your left. However, if you use your left hand to operate your mouse, then submerge your right hand instead. After submerging your hand, wait again for the organiser to signal you to remove it. Then dry your hand and return to your PC.

Step 3. Computer Experiment

You can now begin the computer part of the experiment. You will be asked on your screen a number of times to roll a virtual die by means of pressing a **'Roll Die'** button. Each time you roll the die, you will be shown a randomly selected video of a die being rolled. You will be asked to report the number of dots shown on the upward face of the die. There is a different monetary reward associated with each number of dots, as will be shown on the computer screen. The rewards are associated with the different faces of the die as follows:

- 1 dot - £0.05
- 2 dots - £0.10
- 3 dots - £0.15
- 4 dots - £0.20
- 5 dots - £0.25
- 6 dots - £0

You will be given the cumulative sum of these rewards in addition to your basic £3.50 participation fee at the conclusion of this experiment.

When you are ready to begin the experiment press **'Start'** on your computer screen.

Step 4. Questionnaire

After the computer experiment, you will be prompted to fill out a questionnaire.

Step 5. Payment

When you have completed the questionnaire, you will be asked to come to the front desk to receive your payment.

Figure B.10: Instructions sheet.

Informed Consent Form

Introduction
This study attempts to collect information about how people make decisions.

Procedures
You will take part in a decision making study. Digital images of you will be recorded during this experiment. After this study, you will answer a number of questions on a variety of topics.

Risks/Discomforts
Risks are minimal for involvement in this study. Although we do not expect any harm to come upon any participant due to electronic malfunction of the computer, it is possible though extremely rare and uncommon.

Benefits
There are no direct benefits for participants. However, it is hoped that through your participation, researchers will learn more about human decision making.

Confidentiality
All data obtained from participants will be kept confidential and will only be reported in an aggregate format (by reporting only combined results and never reporting individual ones). All questionnaires will be concealed, and no one other than then primary investigators listed below will have access to them.

Compensation
You will receive a participation fee of £3 as well as the reward you earn during the experiment.

Participation
Participation in this research study is completely voluntary. You have the right to withdraw at any time or refuse to participate entirely.

I have read and understood the above consent form.

Name: _____

Signature: _____

PI's
Doratha Vinkemeier
Thorsten Chmura
Roberto Hernan Gonzales

Figure B.11: Consent form sheet.

Appendix C

Die experiment: Supplementary dice rolling data

C.1 Digital data collected

In addition to the consent form and questionnaires, during the dice rolling experiment additional digital data is collected. For each participant, this data is as follows:

- A video beginning when the participant presses the 'Start' button on screen B.2 and ending after their report of the 20^{th} roll.
- Timestamps for each frame of the video above.
- 2D coordinates for the participant's mouse positions along with their timestamps, beginning when the participant presses the 'Start' button on screen B.2 and ending after their report of the 20^{th} roll.
- The twenty timestamps when the participant pressed the 'Roll Die' button.
- The twenty randomly chosen dice videos for that participant.
- The twenty timestamps when the videos ended.
- The twenty timestamps when the participant reported what they rolled.
- What the participant reported that they rolled.

- The date and time of the experiment.
- The computer used.
- The sequential number of the participant at that computer.

C.2 Die rolling data

In this section, I present the die rolling data graphically represented in a compact form. It is split into three groups by treatment: warm water, cold water and no water treatment. This makes it easy to see what types of lying are taking place. The participant's unique ID (in column 1, part.) is represented by a pair of numbers on a blue background. Each participant takes up two rows, the one with their ID and the one below, which begins with a '∼'. The numbers on a dark grey background are what they really rolled. The numbers directly below on a light grey background are what they reported. When what was rolled does not match what was reported, this pair of numbers, what was rolled located directly above what was reported, is highlighted in red if it lead to a profit or yellow if it lead to a loss so that lies can be easily spotted. In the column 'L' is the number of lies that participant made. A '*' directly below this, means that this participant is possibly an example of *homo economicus*, that is, they maximize their reward by always choosing 5.

C.2.1 Warm water treatment

Table C.1: Dice rolling data for the warm water treatment.

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(12,10)	1	3	3	1	4	4	4	6	6	5	1	1	1	3	2	4	3	5	3	2	0
∼	1	3	3	1	4	4	4	6	6	5	1	1	1	3	2	4	3	5	3	2	
(12,15)	6	6	6	1	3	5	2	2	2	1	5	3	4	2	6	4	6	2	1	2	17
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	4	4	4	4	4	5	5	
(12,16)	6	2	4	2	5	6	4	5	2	2	5	3	6	5	6	6	3	5	1	5	0
∼	6	2	4	2	5	6	4	5	2	2	5	3	6	5	6	6	3	5	1	5	
(12,17)	1	5	1	6	2	1	2	2	6	6	5	4	1	4	1	6	2	6	1	3	11

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	1	5	5	5	3	1	4	2	6	5	5	4	5	4	3	5	5	5	5	3	
(12,24)	1	6	6	6	2	3	3	2	1	1	4	1	4	4	5	6	6	4	1	5	0
∼	1	6	6	6	2	3	3	2	1	1	4	1	4	4	5	6	6	4	1	5	
(12,25)	1	3	1	5	5	2	2	3	6	2	1	1	3	5	2	2	1	5	5	2	0
∼	1	3	1	5	5	2	2	3	6	2	1	1	3	5	2	2	1	5	5	2	
(12,26)	1	1	2	5	3	5	3	1	3	1	2	3	2	6	4	4	2	2	1	3	0
∼	1	1	2	5	3	5	3	1	3	1	2	3	2	6	4	4	2	2	1	3	
(12,27)	2	6	6	3	3	3	2	5	3	6	5	1	4	2	2	6	1	3	3	5	7
∼	2	6	5	5	3	3	2	5	5	5	5	1	4	5	2	5	1	3	5	5	
(12,28)	5	4	1	4	5	3	1	3	1	4	6	5	2	5	6	1	2	3	1	2	5
∼	5	4	1	4	5	3	2	3	3	4	6	5	2	5	6	2	3	3	2	2	
(12,29)	3	2	4	1	2	2	3	5	6	5	5	6	4	1	4	4	6	3	4	5	0
∼	3	2	4	1	2	2	3	5	6	5	5	6	4	1	4	4	6	3	4	5	
(12,30)	5	6	4	3	6	5	5	3	6	2	1	3	1	1	2	3	6	4	1	6	0
∼	5	6	4	3	6	5	5	3	6	2	1	3	1	1	2	3	6	4	1	6	
(12,31)	2	2	4	1	4	6	4	1	3	1	1	2	4	3	2	3	2	1	3	6	0
∼	2	2	4	1	4	6	4	1	3	1	1	2	4	3	2	3	2	1	3	6	
(12,32)	5	2	5	4	4	5	5	2	3	5	4	3	5	6	6	4	3	3	3	3	6
∼	5	5	5	4	4	5	5	4	2	5	4	3	5	2	2	4	3	3	4	3	
(12,8)	6	5	4	6	2	1	1	2	5	6	5	4	5	2	3	4	2	3	6	1	0
∼	6	5	4	6	2	1	1	2	5	6	5	4	5	2	3	4	2	3	6	1	
(12,9)	6	6	4	2	4	5	3	1	6	5	1	5	5	5	1	4	5	5	2	6	13
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(15,10)	6	4	3	5	4	6	1	1	1	2	4	2	4	2	2	2	6	2	2	5	0
∼	6	4	3	5	4	6	1	1	1	2	4	2	4	2	2	2	6	2	2	5	
(15,11)	1	4	2	2	1	6	6	5	5	6	2	3	2	6	2	4	4	6	2	1	1
∼	1	4	2	2	1	6	6	5	5	6	2	3	2	6	2	5	4	6	2	1	
(15,15)	5	4	5	1	6	1	1	2	6	1	2	3	2	4	6	3	1	6	6	4	18
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(15,16)	1	6	4	2	3	2	6	3	4	4	3	3	4	2	4	1	4	2	2	3	9
∼	1	5	5	2	5	5	3	3	4	4	5	3	4	5	4	5	5	2	2	3	

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(15,17)	3	2	2	1	1	1	3	2	1	2	4	5	3	1	1	5	4	2	4	3	0
∼	3	2	2	1	1	1	3	2	1	2	4	5	3	1	1	5	4	2	4	3	
(15,7)	2	6	3	3	2	4	5	4	1	1	1	2	6	6	2	3	2	4	4	2	2
∼	2	6	3	5	2	4	5	4	1	1	1	2	6	6	2	3	4	4	4	2	
(15,9)	6	3	1	3	1	6	1	4	6	6	3	4	2	5	3	6	4	2	1	3	0
∼	6	3	1	3	1	6	1	4	6	6	3	4	2	5	3	6	4	2	1	3	
(16,10)	1	6	6	1	3	2	2	5	2	5	5	3	5	5	1	6	6	5	6	6	0
∼	1	6	6	1	3	2	2	5	2	5	5	3	5	5	1	6	6	5	6	6	
(16,11)	1	1	4	3	6	4	1	5	2	1	1	3	4	4	1	5	1	3	6	3	0
∼	1	1	4	3	6	4	1	5	2	1	1	3	4	4	1	5	1	3	6	3	
(16,16)	1	3	1	4	6	1	2	5	2	4	2	4	3	6	2	6	6	3	1	4	19
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(16,17)	2	2	1	2	3	5	4	6	2	1	3	1	2	5	6	5	4	5	2	3	1
∼	2	2	5	2	3	5	4	6	2	1	3	1	2	5	6	5	4	5	2	3	
(16,18)	2	4	3	5	1	3	5	1	1	2	2	4	1	3	3	5	2	5	6	1	0
∼	2	4	3	5	1	3	5	1	1	2	2	4	1	3	3	5	2	5	6	1	
(16,25)	4	5	1	5	5	3	3	3	4	4	2	5	1	6	1	6	2	6	2	5	0
∼	4	5	1	5	5	3	3	3	4	4	2	5	1	6	1	6	2	6	2	5	
(16,26)	3	4	1	2	4	5	6	5	4	2	1	2	4	5	6	2	4	5	2	2	0
∼	3	4	1	2	4	5	6	5	4	2	1	2	4	5	6	2	4	5	2	2	
(16,27)	2	3	6	1	6	4	6	4	5	1	5	2	3	6	6	5	6	3	2	5	0
∼	2	3	6	1	6	4	6	4	5	1	5	2	3	6	6	5	6	3	2	5	
(16,28)	1	6	1	1	6	6	6	3	5	3	3	3	1	5	4	2	5	2	3	4	0
∼	1	6	1	1	6	6	6	3	5	3	3	3	1	5	4	2	5	2	3	4	
(16,29)	5	2	1	1	3	3	2	4	1	5	5	6	5	3	2	3	6	5	1	2	5
∼	5	2	5	5	3	3	2	4	2	5	5	5	5	3	2	3	6	5	4	2	
(16,30)	4	1	6	5	6	2	6	1	2	5	6	4	2	6	4	2	3	3	1	2	0
∼	4	1	6	5	6	2	6	1	2	5	6	4	2	6	4	2	3	3	1	2	
(16,31)	5	1	2	5	2	1	3	4	3	2	2	3	2	6	4	2	1	2	5	5	1
∼	5	1	2	5	2	1	5	4	3	2	2	3	2	6	4	2	1	2	5	5	
(16,32)	5	3	4	2	4	2	1	3	1	3	6	2	1	5	4	2	2	3	5	2	0

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	5	3	4	2	4	2	1	3	1	3	6	2	1	5	4	2	2	3	5	2	
(16,7)	3	2	3	4	4	4	3	4	6	5	5	3	3	6	6	3	2	2	6	6	0
∼	3	2	3	4	4	4	3	4	6	5	5	3	3	6	6	3	2	2	6	6	
(16,9)	5	6	5	6	5	2	3	4	2	2	2	6	6	6	4	3	2	4	4	4	2
∼	5	6	5	6	5	2	3	4	2	2	2	5	6	5	4	3	2	4	4	4	
(19,13)	2	5	1	5	3	4	3	1	4	2	2	3	2	5	4	3	3	6	4	2	0
∼	2	5	1	5	3	4	3	1	4	2	2	3	2	5	4	3	3	6	4	2	
(19,14)	4	3	5	3	5	4	4	4	3	5	6	6	1	3	1	6	5	6	2	6	1
∼	4	3	5	3	5	4	4	4	3	5	6	6	1	3	1	6	5	5	2	6	
(19,15)	4	2	4	6	3	2	5	4	2	3	3	6	6	5	4	3	1	4	2	3	0
∼	4	2	4	6	3	2	5	4	2	3	3	6	6	5	4	3	1	4	2	3	
(19,22)	5	4	5	2	4	5	3	6	1	4	6	5	3	3	4	4	5	6	4	3	1
∼	5	4	5	2	4	5	3	6	1	4	6	6	3	3	4	4	5	6	4	3	
(19,23)	3	3	1	3	3	1	3	2	6	5	3	5	3	4	3	3	5	5	1	4	0
∼	3	3	1	3	3	1	3	2	6	5	3	5	3	4	3	3	5	5	1	4	
(19,24)	6	2	6	1	1	4	4	1	1	6	4	2	1	4	6	1	1	1	1	6	20
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(19,25)	5	2	2	6	6	3	3	1	3	6	3	5	3	4	5	4	3	3	2	6	0
∼	5	2	2	6	6	3	3	1	3	6	3	5	3	4	5	4	3	3	2	6	
(19,26)	4	6	6	6	1	5	6	5	3	4	6	4	3	6	3	2	5	6	6	1	0
∼	4	6	6	6	1	5	6	5	3	4	6	4	3	6	3	2	5	6	6	1	
(19,27)	4	1	6	6	4	5	3	5	2	4	3	6	3	6	4	4	4	2	6	1	0
∼	4	1	6	6	4	5	3	5	2	4	3	6	3	6	4	4	4	2	6	1	
(19,28)	4	2	3	3	5	6	4	4	3	3	2	6	5	5	1	1	5	2	5	5	0
∼	4	2	3	3	5	6	4	4	3	3	2	6	5	5	1	1	5	2	5	5	
(19,29)	2	6	4	6	6	4	6	2	1	3	4	3	5	6	1	3	6	5	6	2	0
∼	2	6	4	6	6	4	6	2	1	3	4	3	5	6	1	3	6	5	6	2	
(19,6)	1	1	6	4	5	3	3	6	2	5	4	2	2	1	6	1	3	2	3	2	7
∼	1	1	5	4	5	3	4	5	2	5	4	2	3	2	5	1	3	2	4	2	
(19,7)	2	1	4	4	5	3	4	2	1	4	1	1	1	2	1	4	1	1	4	1	2
∼	2	1	4	4	5	3	4	2	1	4	5	1	5	2	1	4	1	1	4	1	

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(19,8)	1	1	2	2	4	3	4	1	4	1	3	6	2	3	2	4	2	2	6	6	0
∼	1	1	2	2	4	3	4	1	4	1	3	6	2	3	2	4	2	2	6	6	
(19,9)	5	2	4	2	6	3	6	3	5	3	6	2	2	6	2	4	6	6	6	4	0
∼	5	2	4	2	6	3	6	3	5	3	6	2	2	6	2	4	6	6	6	4	
(22,12)	5	3	5	2	1	5	4	3	1	5	6	1	6	6	1	6	2	4	1	1	0
∼	5	3	5	2	1	5	4	3	1	5	6	1	6	6	1	6	2	4	1	1	
(22,13)	1	1	6	3	6	4	6	3	5	5	2	1	5	5	5	2	1	6	5	5	0
∼	1	1	6	3	6	4	6	3	5	5	2	1	5	5	5	2	1	6	5	5	
(22,14)	4	6	5	4	1	2	6	6	2	2	6	4	2	1	1	4	1	1	4	2	14
∼	4	5	5	4	5	5	5	5	4	4	5	4	4	4	5	4	4	5	4	5	
(22,20)	6	5	1	1	3	3	3	3	6	1	4	4	6	6	6	2	4	4	2	6	0
∼	6	5	1	1	3	3	3	3	6	1	4	4	6	6	6	2	4	4	2	6	
(22,21)	5	5	1	1	4	2	4	5	2	6	2	4	4	5	3	3	6	3	4	1	0
∼	5	5	1	1	4	2	4	5	2	6	2	4	4	5	3	3	6	3	4	1	
(22,22)	2	4	5	3	6	2	5	3	1	3	3	2	5	6	3	1	3	4	5	5	0
∼	2	4	5	3	6	2	5	3	1	3	3	2	5	6	3	1	3	4	5	5	
(22,23)	5	6	2	1	1	6	4	2	3	2	4	6	6	3	3	6	3	4	4	2	2
∼	5	6	2	1	1	6	4	2	3	2	4	6	5	3	3	5	3	4	4	2	
(22,24)	5	5	4	3	3	2	6	4	2	4	4	2	6	1	3	5	5	3	6	3	1
∼	5	5	4	3	3	2	6	4	2	4	4	2	6	1	3	5	5	3	5	3	
(22,25)	3	2	3	4	4	6	3	3	4	2	1	6	2	6	5	6	3	1	2	3	0
∼	3	2	3	4	4	6	3	3	4	2	1	6	2	6	5	6	3	1	2	3	
(22,26)	4	3	5	4	4	3	3	4	2	2	1	2	5	1	4	1	6	5	6	5	0
∼	4	3	5	4	4	3	3	4	2	2	1	2	5	1	4	1	6	5	6	5	
(22,27)	3	5	5	6	5	5	4	4	6	3	6	2	4	4	6	5	5	6	3	5	1
∼	3	5	5	6	5	5	4	4	6	3	5	2	4	4	6	5	5	6	3	5	
(22,7)	2	1	4	2	2	1	4	4	6	1	4	6	5	2	6	2	3	6	4	5	0
∼	2	1	4	2	2	1	4	4	6	1	4	6	5	2	6	2	3	6	4	5	
(22,8)	3	3	2	2	6	4	3	2	2	6	4	4	1	3	4	2	1	1	5	5	0
∼	3	3	2	2	6	4	3	2	2	6	4	4	1	3	4	2	1	1	5	5	
(22,9)	2	6	5	5	3	5	4	5	2	3	6	6	6	1	3	6	6	4	3	4	0

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	2	6	5	5	3	5	4	5	2	3	6	6	6	1	3	6	6	4	3	4	
(25,13)	3	4	3	2	4	2	1	5	1	4	1	6	1	2	4	1	1	4	2	1	0
∼	3	4	3	2	4	2	1	5	1	4	1	6	1	2	4	1	1	4	2	1	
(25,14)	1	5	5	6	3	5	1	1	3	5	3	2	6	5	4	1	6	5	5	5	12
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(25,21)	3	3	4	4	6	3	2	1	2	4	1	6	5	5	3	1	4	2	6	4	0
∼	3	3	4	4	6	3	2	1	2	4	1	6	5	5	3	1	4	2	6	4	
(25,22)	6	5	6	2	5	3	1	3	6	5	1	2	4	2	3	4	2	4	6	1	0
∼	6	5	6	2	5	3	1	3	6	5	1	2	4	2	3	4	2	4	6	1	
(25,23)	3	6	5	6	3	5	5	5	4	6	3	6	3	2	2	4	6	5	2	5	0
∼	3	6	5	6	3	5	5	5	4	6	3	6	3	2	2	4	6	5	2	5	
(25,24)	4	4	2	5	5	2	5	1	3	3	2	6	5	4	2	1	4	2	5	2	0
∼	4	4	2	5	5	2	5	1	3	3	2	6	5	4	2	1	4	2	5	2	
(25,25)	3	2	1	4	5	3	2	2	1	2	4	4	6	4	3	3	3	2	2	6	18
∼	3	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	
(25,26)	6	6	3	5	3	1	4	5	2	1	6	2	2	3	5	1	1	4	4	3	0
∼	6	6	3	5	3	1	4	5	2	1	6	2	2	3	5	1	1	4	4	3	
(25,28)	4	3	3	6	3	3	1	4	5	2	3	3	4	4	4	1	5	3	1	4	1
∼	4	3	3	6	3	5	1	4	5	2	3	3	4	4	4	1	5	3	1	4	
(25,4)	4	6	2	3	4	2	4	4	6	1	3	3	2	2	5	2	2	3	4	1	1
∼	4	6	2	3	4	2	4	4	5	1	3	3	2	2	5	2	2	3	4	1	
(25,6)	2	2	1	6	6	2	3	5	2	2	3	4	6	2	3	1	6	5	3	5	0
∼	2	2	1	6	6	2	3	5	2	2	3	4	6	2	3	1	6	5	3	5	
(25,7)	3	6	1	2	6	2	6	5	4	1	3	6	4	2	2	2	5	1	5	3	0
∼	3	6	1	2	6	2	6	5	4	1	3	6	4	2	2	2	5	1	5	3	
(25,8)	1	3	6	4	5	6	2	1	3	1	3	1	2	6	4	3	2	3	3	6	19
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(26,14)	4	3	2	3	6	5	6	4	4	6	3	6	2	5	6	2	4	3	4	1	0
∼	4	3	2	3	6	5	6	4	4	6	3	6	2	5	6	2	4	3	4	1	
(26,15)	2	1	6	3	1	3	6	4	5	1	1	4	3	3	4	3	5	5	5	3	2
∼	2	1	6	3	1	3	5	4	5	5	1	4	3	3	4	3	5	5	5	3	

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(26,16)	5	4	4	5	1	2	6	4	2	3	4	5	5	6	1	2	2	4	1	5	0
∼	5	4	4	5	1	2	6	4	2	3	4	5	5	6	1	2	2	4	1	5	
(26,17)	2	2	3	3	6	6	2	4	4	6	6	4	2	6	6	1	6	6	2	1	10
∼	2	2	3	3	6	6	5	5	5	5	6	4	5	2	6	5	6	5	3	2	
(26,18)	2	4	6	3	3	6	3	3	4	3	2	1	4	6	2	6	4	4	1	4	12
∼	2	4	5	5	4	5	3	4	4	3	4	5	5	4	4	5	4	4	5	4	
(26,19)	6	3	1	3	4	4	3	4	4	6	2	4	1	2	4	6	4	4	4	5	0
∼	6	3	1	3	4	4	3	4	4	6	2	4	1	2	4	6	4	4	4	5	
(26,20)	6	5	5	1	2	1	2	5	6	6	6	3	6	1	2	2	1	4	1	2	0
∼	6	5	5	1	2	1	2	5	6	6	6	3	6	1	2	2	1	4	1	2	
(26,21)	4	6	4	1	2	2	1	1	3	4	3	2	2	4	6	6	3	4	6	6	0
∼	4	6	4	1	2	2	1	1	3	4	3	2	2	4	6	6	3	4	6	6	
(26,22)	3	2	6	6	4	1	1	4	5	5	2	5	2	5	4	3	4	3	3	1	0
∼	3	2	6	6	4	1	1	4	5	5	2	5	2	5	4	3	4	3	3	1	
(26,23)	4	1	1	4	1	3	3	5	6	3	4	2	3	5	6	3	6	5	5	1	0
∼	4	1	1	4	1	3	3	5	6	3	4	2	3	5	6	3	6	5	5	1	
(26,6)	5	6	3	1	4	4	2	6	5	6	4	6	1	5	2	3	3	4	4	5	0
∼	5	6	3	1	4	4	2	6	5	6	4	6	1	5	2	3	3	4	4	5	
(29,12)	2	4	4	5	1	2	1	6	2	1	2	6	4	2	3	6	4	4	6	2	0
∼	2	4	4	5	1	2	1	6	2	1	2	6	4	2	3	6	4	4	6	2	
(29,13)	5	4	6	5	6	1	4	1	5	5	4	6	3	6	2	1	1	2	3	3	0
∼	5	4	6	5	6	1	4	1	5	5	4	6	3	6	2	1	1	2	3	3	
(29,14)	1	4	1	2	2	3	5	4	4	6	6	1	1	5	6	5	5	2	3	3	14
∼	4	4	5	3	5	4	5	5	4	5	5	4	5	5	5	5	5	3	4	4	
(29,21)	5	2	1	2	3	4	5	3	6	5	3	5	2	3	2	5	1	4	1	5	0
∼	5	2	1	2	3	4	5	3	6	5	3	5	2	3	2	5	1	4	1	5	
(29,22)	1	1	6	4	4	3	6	3	6	5	5	5	3	5	6	3	2	4	6	5	15
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(29,23)	6	6	4	3	6	6	1	4	5	6	1	5	3	6	4	1	4	3	1	5	0
∼	6	6	4	3	6	6	1	4	5	6	1	5	3	6	4	1	4	3	1	5	
(29,24)	5	2	1	6	6	4	6	2	6	1	1	5	3	1	2	1	2	6	6	6	16

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	5	2	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	
(29,25)	6	1	4	6	5	6	6	5	4	5	6	5	4	2	1	2	4	1	1	6	0
∼	6	1	4	6	5	6	6	5	4	5	6	5	4	2	1	2	4	1	1	6	
(29,26)	3	5	3	1	5	1	5	4	2	3	6	6	4	6	3	6	4	1	1	6	0
∼	3	5	3	1	5	1	5	4	2	3	6	6	4	6	3	6	4	1	1	6	
(29,27)	4	6	4	3	5	6	3	3	5	6	2	1	4	4	6	1	1	2	4	5	7
∼	4	5	4	3	5	5	3	3	5	5	2	5	4	4	5	5	4	2	4	5	
(29,28)	5	3	6	2	6	3	2	2	5	1	1	5	1	5	4	2	2	6	1	2	0
∼	5	3	6	2	6	3	2	2	5	1	1	5	1	5	4	2	2	6	1	2	
(29,3)	2	2	4	2	4	5	5	3	6	2	6	6	4	3	3	2	1	4	6	4	18
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(29,5)	5	1	5	6	5	1	3	6	4	5	4	3	5	5	6	6	5	5	2	2	12
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(29,6)	3	3	3	1	1	5	5	5	4	2	5	5	6	2	5	5	5	2	3	4	9
∼	3	3	5	5	4	5	5	5	4	4	5	5	5	4	5	5	5	4	4	5	
(29,7)	4	6	2	1	1	1	1	4	6	6	1	3	5	1	3	2	1	5	1	2	17
∼	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	
(5,14)	3	5	3	5	6	6	4	3	3	5	4	4	5	4	6	2	3	4	4	6	0
∼	3	5	3	5	6	6	4	3	3	5	4	4	5	4	6	2	3	4	4	6	
(5,15)	1	6	3	4	4	2	6	1	5	4	2	4	5	1	1	1	3	3	6	5	1
∼	1	6	3	4	4	2	6	1	3	4	2	4	5	1	1	1	3	3	6	5	
(5,16)	6	2	5	4	3	2	6	5	1	4	6	6	2	2	6	5	6	3	6	4	0
∼	6	2	5	4	3	2	6	5	1	4	6	6	2	2	6	5	6	3	6	4	
(5,23)	2	2	3	4	1	5	4	5	3	2	3	3	2	6	3	5	1	3	3	5	0
∼	2	2	3	4	1	5	4	5	3	2	3	3	2	6	3	5	1	3	3	5	
(5,24)	2	6	2	3	1	3	2	2	3	5	4	2	3	1	2	2	2	6	2	6	3
∼	2	5	2	3	1	3	2	2	3	5	4	2	3	1	2	2	2	5	2	5	
(5,25)	2	1	6	5	2	1	2	1	2	2	2	4	6	4	6	6	6	5	4	5	9
∼	2	5	5	5	2	5	2	5	2	2	5	4	5	4	5	5	5	5	4	5	
(5,26)	2	5	1	1	4	2	5	2	2	3	2	6	4	6	2	4	3	4	4	1	0
∼	2	5	1	1	4	2	5	2	2	3	2	6	4	6	2	4	3	4	4	1	

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(5,31)	2	2	4	2	6	4	4	3	2	3	6	3	1	5	4	3	3	1	3	6	0
∼	2	2	4	2	6	4	4	3	2	3	6	3	1	5	4	3	3	1	3	6	
(5,32)	1	3	5	3	1	4	6	2	6	2	3	3	2	4	3	1	6	6	1	5	15
∼	1	5	5	5	5	4	5	5	5	5	5	5	5	4	5	5	5	5	5	5	
(5,33)	2	3	1	6	5	4	4	6	2	1	4	6	1	2	5	1	5	1	1	5	10
∼	4	3	5	5	5	4	4	4	5	5	4	6	5	2	5	5	5	4	5	5	
(5,4)	2	6	5	6	4	1	2	3	6	4	2	2	6	1	1	2	3	6	6	2	18
∼	5	5	5	5	5	5	5	5	5	5	4	4	5	4	5	4	3	5	4	4	
(5,7)	4	4	1	4	3	5	4	2	3	4	1	4	6	3	4	2	3	4	5	2	0
∼	4	4	1	4	3	5	4	2	3	4	1	4	6	3	4	2	3	4	5	2	
(5,8)	5	1	4	2	4	2	1	5	5	6	2	3	4	2	6	2	1	6	6	3	0
∼	5	1	4	2	4	2	1	5	5	6	2	3	4	2	6	2	1	6	6	3	
(5,9)	4	3	4	3	4	2	4	2	2	1	4	4	3	2	3	6	1	5	6	4	11
∼	4	3	4	5	4	5	4	5	4	4	4	4	5	4	5	5	4	5	5	4	
(6,13)	5	4	4	6	2	2	2	4	2	4	6	6	6	6	6	3	3	4	5	6	14
∼	5	4	5	4	5	3	5	4	5	4	3	3	5	5	4	3	4	5	5	3	
(6,14)	4	2	2	5	5	1	4	1	1	6	5	1	3	5	3	3	3	3	4	5	9
∼	4	3	5	5	5	3	4	3	4	5	5	4	3	5	3	3	4	4	4	5	
(6,15)	4	1	6	2	5	3	1	5	4	5	6	6	6	4	5	3	1	6	5	6	10
∼	4	5	5	2	5	3	5	5	5	5	5	5	6	4	5	5	5	5	5	5	
(6,22)	1	5	6	2	3	2	2	6	4	3	4	2	2	2	3	5	3	2	6	4	0
∼	1	5	6	2	3	2	2	6	4	3	4	2	2	2	3	5	3	2	6	4	
(6,23)	5	3	4	2	2	1	5	2	1	4	5	2	5	2	6	1	2	6	1	2	0
∼	5	3	4	2	2	1	5	2	1	4	5	2	5	2	6	1	2	6	1	2	
(6,24)	5	1	5	5	5	3	5	3	5	3	5	4	4	3	2	2	6	2	1	4	0
∼	5	1	5	5	5	3	5	3	5	3	5	4	4	3	2	2	6	2	1	4	
(6,25)	5	2	1	5	6	4	5	2	2	5	4	2	3	6	5	5	5	1	3	1	0
∼	5	2	1	5	6	4	5	2	2	5	4	2	3	6	5	5	5	1	3	1	
(6,26)	5	5	6	5	4	1	3	2	4	2	2	3	5	2	6	1	4	3	4	2	0
∼	5	5	6	5	4	1	3	2	4	2	2	3	5	2	6	1	4	3	4	2	
(6,27)	4	4	6	1	4	1	3	3	6	3	1	5	3	6	6	3	1	4	4	1	0

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	4	4	6	1	4	1	3	3	6	3	1	5	3	6	6	3	1	4	4	1	
(6,28)	1	4	6	2	3	6	1	1	6	6	5	6	1	3	2	4	4	1	1	3	0
∼	1	4	6	2	3	6	1	1	6	6	5	6	1	3	2	4	4	1	1	3	
(6,29)	6	1	2	2	5	1	3	6	6	6	2	6	5	6	6	1	5	5	1	1	0
∼	6	1	2	2	5	1	3	6	6	6	2	6	5	6	6	1	5	5	1	1	
(6,32)	5	4	1	3	3	1	5	3	4	2	1	2	4	4	6	2	6	5	3	3	0
∼	5	4	1	3	3	1	5	3	4	2	1	2	4	4	6	2	6	5	3	3	
(6,6)	1	3	2	5	2	6	6	5	1	6	6	6	3	2	2	4	2	2	1	2	0
∼	1	3	2	5	2	6	6	5	1	6	6	6	3	2	2	4	2	2	1	2	
(6,7)	4	4	2	3	6	6	5	2	1	1	4	4	4	2	1	2	4	4	1	6	0
∼	4	4	2	3	6	6	5	2	1	1	4	4	4	2	1	2	4	4	1	6	
(6,8)	4	2	3	1	4	4	3	1	2	6	3	1	5	4	5	6	2	2	1	1	0
∼	4	2	3	1	4	4	3	1	2	6	3	1	5	4	5	6	2	2	1	1	
(9,10)	1	6	3	6	6	6	6	6	4	3	6	5	6	1	6	5	5	1	3	3	1
∼	1	6	3	6	6	6	6	6	4	3	5	5	6	1	6	5	5	1	3	3	
(9,15)	1	1	3	4	2	4	1	4	1	1	4	3	4	6	6	6	3	3	3	5	0
∼	1	1	3	4	2	4	1	4	1	1	4	3	4	6	6	6	3	3	3	5	
(9,16)	5	3	6	3	3	6	6	3	3	3	6	5	4	5	6	2	4	5	2	3	0
∼	5	3	6	3	3	6	6	3	3	3	6	5	4	5	6	2	4	5	2	3	
(9,17)	4	6	3	5	5	5	4	2	1	1	4	2	4	1	6	4	6	6	5	3	0
∼	4	6	3	5	5	5	4	2	1	1	4	2	4	1	6	4	6	6	5	3	
(9,23)	5	6	5	1	4	5	2	2	1	4	5	3	1	4	6	3	1	3	5	4	0
∼	5	6	5	1	4	5	2	2	1	4	5	3	1	4	6	3	1	3	5	4	
(9,24)	5	5	4	6	5	1	4	2	5	2	1	3	2	2	3	6	4	1	3	5	0
∼	5	5	4	6	5	1	4	2	5	2	1	3	2	2	3	6	4	1	3	5	
(9,25)	3	6	3	4	6	2	6	4	2	5	5	4	3	3	1	4	2	3	6	1	18
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(9,26)	6	3	3	6	1	6	6	6	4	2	4	2	6	5	4	4	4	1	3	3	6
∼	5	3	4	6	2	5	5	6	4	2	4	3	6	5	4	4	4	1	3	3	
(9,27)	5	2	4	6	1	1	1	4	4	2	5	5	2	3	1	3	3	3	3	1	0
∼	5	2	4	6	1	1	1	4	4	2	5	5	2	3	1	3	3	3	3	1	

Continued on next page

Table C.1 – *Warm water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(9,28)	6	4	1	2	5	2	5	1	3	3	2	6	1	6	1	5	1	4	6	2	14
∼	5	4	5	5	5	5	5	3	4	5	2	5	4	5	5	5	4	4	5	3	
(9,29)	2	2	6	4	2	1	6	4	4	5	4	6	4	2	3	2	4	1	5	5	0
∼	2	2	6	4	2	1	6	4	4	5	4	6	4	2	3	2	4	1	5	5	
(9,30)	4	2	3	6	6	4	1	2	2	1	3	1	3	5	3	2	3	1	1	2	0
∼	4	2	3	6	6	4	1	2	2	1	3	1	3	5	3	2	3	1	1	2	
(9,32)	4	3	1	5	6	3	6	1	4	6	2	3	4	4	6	6	1	3	4	5	13
∼	4	4	4	5	5	4	5	4	4	5	4	4	4	4	5	5	4	4	4	5	
(9,33)	6	1	6	4	3	4	4	5	4	2	3	5	1	6	6	6	1	6	4	1	9
∼	6	1	5	4	4	4	5	5	4	2	4	5	5	4	5	6	5	4	4	1	
(9,6)	1	6	5	2	3	2	5	4	1	2	1	6	4	2	2	4	6	1	1	3	12
∼	1	6	5	5	4	5	5	4	4	2	4	5	4	3	4	4	5	2	4	5	
(9,8)	3	6	5	4	4	3	3	5	2	2	2	6	3	2	6	2	6	5	3	3	0
∼	3	6	5	4	4	3	3	5	2	2	2	6	3	2	6	2	6	5	3	3	
(9,9)	2	1	4	3	2	4	3	2	1	4	1	4	1	1	3	2	1	5	3	6	11
∼	2	4	4	3	5	4	5	2	4	4	1	4	5	3	5	4	2	5	5	5	

C.2.2 Cold water treatment

Table C.2: Dice rolling data for the cold water treatment..

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(12,11)	6	2	4	1	2	2	1	2	3	6	1	5	5	4	4	5	5	2	5	5	0
∼	6	2	4	1	2	2	1	2	3	6	1	5	5	4	4	5	5	2	5	5	
(12,12)	1	5	1	6	6	3	4	3	2	5	2	3	3	3	3	2	3	6	6	2	18
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(12,13)	2	1	4	6	5	5	5	4	3	6	3	2	3	4	2	6	3	6	2	4	0
∼	2	1	4	6	5	5	5	4	3	6	3	2	3	4	2	6	3	6	2	4	
(12,14)	4	4	5	3	2	1	3	3	4	1	1	2	4	2	6	3	4	1	6	1	0
∼	4	4	5	3	2	1	3	3	4	1	1	2	4	2	6	3	4	1	6	1	
(12,21)	5	1	5	1	2	2	6	3	2	3	3	6	2	5	1	6	5	1	4	3	10

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	5	5	5	1	2	5	5	3	2	5	5	5	2	5	5	5	5	5	4	5	
(12,22)	1	6	5	1	2	2	4	2	5	6	4	5	3	6	4	3	2	3	2	4	4
∼	1	5	5	3	2	2	4	2	5	5	4	5	3	5	4	3	2	3	2	4	
(12,23)	5	2	6	3	4	5	5	3	3	6	6	2	5	2	5	2	6	2	2	6	14
∼	5	5	5	5	4	5	5	4	4	5	5	3	5	3	5	3	5	3	3	5	
(12,33)	5	4	4	2	1	4	3	6	3	6	4	5	5	6	4	3	3	5	5	5	0
∼	5	4	4	2	1	4	3	6	3	6	4	5	5	6	4	3	3	5	5	5	
(12,34)	2	4	3	3	4	2	1	3	1	2	3	3	1	2	2	1	5	5	6	3	0
∼	2	4	3	3	4	2	1	3	1	2	3	3	1	2	2	1	5	5	6	3	
(12,35)	3	3	5	3	3	2	2	6	2	4	6	1	4	4	3	1	2	3	5	5	0
∼	3	3	5	3	3	2	2	6	2	4	6	1	4	4	3	1	2	3	5	5	
(12,37)	4	4	1	6	2	3	3	2	1	4	2	1	6	6	3	3	4	1	5	1	0
∼	4	4	1	6	2	3	3	2	1	4	2	1	6	6	3	3	4	1	5	1	
(12,38)	6	4	2	4	6	1	1	4	2	1	6	4	5	6	4	5	4	1	3	6	0
∼	6	4	2	4	6	1	1	4	2	1	6	4	5	6	4	5	4	1	3	6	
(12,40)	5	6	2	1	1	5	6	4	5	6	3	6	5	2	2	2	1	6	5	3	0
∼	5	6	2	1	1	5	6	4	5	6	3	6	5	2	2	2	1	6	5	3	
(15,12)	6	4	5	5	6	5	3	4	5	1	1	5	4	4	3	5	5	2	1	5	0
∼	6	4	5	5	6	5	3	4	5	1	1	5	4	4	3	5	5	2	1	5	
(15,13)	6	4	3	1	2	1	1	3	6	3	1	3	1	2	2	2	2	4	2	4	0
∼	6	4	3	1	2	1	1	3	6	3	1	3	1	2	2	2	2	4	2	4	
(15,14)	3	5	6	3	4	1	1	2	2	2	5	1	6	1	6	1	2	4	5	1	0
∼	3	5	6	3	4	1	1	2	2	2	5	1	6	1	6	1	2	4	5	1	
(15,21)	4	3	2	2	2	6	6	2	1	1	2	1	6	5	6	4	5	2	4	6	0
∼	4	3	2	2	2	6	6	2	1	1	2	1	6	5	6	4	5	2	4	6	
(15,22)	1	3	3	1	6	2	2	4	3	4	4	2	1	5	2	3	2	5	4	5	0
∼	1	3	3	1	6	2	2	4	3	4	4	2	1	5	2	3	2	5	4	5	
(15,23)	6	3	1	4	5	3	2	5	2	4	1	4	2	2	5	4	5	2	1	3	0
∼	6	3	1	4	5	3	2	5	2	4	1	4	2	2	5	4	5	2	1	3	
(15,25)	5	4	2	6	3	3	3	5	2	4	4	2	2	4	4	5	4	3	1	6	0
∼	5	4	2	6	3	3	3	5	2	4	4	2	2	4	4	5	4	3	1	6	

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(15,26)	5	5	1	2	6	1	3	5	6	6	4	2	4	1	3	3	5	3	2	3	0
∼	5	5	1	2	6	1	3	5	6	6	4	2	4	1	3	3	5	3	2	3	
(15,27)	5	6	5	3	1	4	4	6	2	6	2	5	5	3	4	4	6	4	4	5	15
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(15,28)	3	6	3	3	4	6	6	5	6	1	6	4	4	4	2	2	4	4	5	5	0
∼	3	6	3	3	4	6	6	5	6	1	6	4	4	4	2	2	4	4	5	5	
(15,29)	4	2	4	4	5	3	4	2	2	2	1	1	6	4	4	5	1	5	5	5	0
∼	4	2	4	4	5	3	4	2	2	2	1	1	6	4	4	5	1	5	5	5	
(15,30)	2	1	3	2	2	5	2	6	5	5	5	5	1	3	1	3	2	6	4	4	0
∼	2	1	3	2	2	5	2	6	5	5	5	5	1	3	1	3	2	6	4	4	
(16,12)	6	4	2	4	2	3	5	2	3	6	5	1	4	6	6	5	1	2	2	2	0
∼	6	4	2	4	2	3	5	2	3	6	5	1	4	6	6	5	1	2	2	2	
(16,13)	4	1	3	4	1	1	6	4	4	3	4	5	6	6	3	2	2	1	1	3	0
∼	4	1	3	4	1	1	6	4	4	3	4	5	6	6	3	2	2	1	1	3	
(16,14)	1	3	2	4	3	2	1	5	6	1	2	5	2	1	2	4	5	5	6	5	0
∼	1	3	2	4	3	2	1	5	6	1	2	5	2	1	2	4	5	5	6	5	
(16,15)	5	1	4	2	2	4	5	1	5	2	2	4	6	1	5	5	2	4	1	6	0
∼	5	1	4	2	2	4	5	1	5	2	2	4	6	1	5	5	2	4	1	6	
(16,22)	3	6	4	3	3	2	4	3	5	4	3	2	4	5	5	3	4	5	4	2	0
∼	3	6	4	3	3	2	4	3	5	4	3	2	4	5	5	3	4	5	4	2	
(16,23)	1	5	3	2	3	2	1	3	2	6	2	1	2	5	3	1	4	1	5	6	1
∼	1	5	3	2	3	2	1	3	2	6	2	5	2	5	3	1	4	1	5	6	
(16,24)	6	3	2	3	6	5	2	2	5	4	5	6	5	4	2	2	1	6	6	4	0
∼	6	3	2	3	6	5	2	2	5	4	5	6	5	4	2	2	1	6	6	4	
(16,33)	1	6	5	5	6	3	5	3	3	2	4	1	1	2	1	3	4	5	5	4	0
∼	1	6	5	5	6	3	5	3	3	2	4	1	1	2	1	3	4	5	5	4	
(16,34)	4	2	4	6	5	6	3	1	3	4	2	5	6	6	5	6	5	5	6	4	1
∼	4	2	4	6	5	6	3	1	3	4	2	5	6	6	5	6	5	5	5	4	
(16,35)	1	4	4	4	4	4	4	1	4	5	5	6	6	4	3	5	4	4	3	5	0
∼	1	4	4	4	4	4	4	1	4	5	5	6	6	4	3	5	4	4	3	5	
(16,37)	3	5	3	5	5	3	4	3	4	2	6	5	4	3	4	1	1	3	6	3	0

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	3	5	3	5	5	3	4	3	4	2	6	5	4	3	4	1	1	3	6	3	
(16,38)	3	3	5	5	2	6	3	6	6	5	5	6	2	1	2	5	5	3	2	2	0
∼	3	3	5	5	2	6	3	6	6	5	5	6	2	1	2	5	5	3	2	2	
(16,39)	4	4	3	5	1	5	6	1	3	5	2	4	2	6	3	1	5	2	5	2	1
∼	4	4	3	5	1	5	6	1	2	5	2	4	2	6	3	1	5	2	5	2	
(16,40)	5	1	3	6	5	6	1	3	5	5	2	6	6	5	3	1	2	6	6	3	0
∼	5	1	3	6	5	6	1	3	5	5	2	6	6	5	3	1	2	6	6	3	
(16,41)	2	3	4	1	5	6	3	3	6	1	3	6	4	4	6	6	2	1	1	4	0
∼	2	3	4	1	5	6	3	3	6	1	3	6	4	4	6	6	2	1	1	4	
(19,10)	5	4	2	6	6	4	3	6	2	4	3	6	5	6	5	5	2	5	2	5	0
∼	5	4	2	6	6	4	3	6	2	4	3	6	5	6	5	5	2	5	2	5	
(19,11)	6	2	4	3	4	1	1	6	1	2	3	1	4	3	2	1	4	1	6	5	19
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(19,12)	4	5	2	3	2	4	6	4	3	3	1	1	4	2	6	6	5	6	4	6	7
∼	4	5	2	3	2	4	5	4	3	3	5	5	4	2	5	5	5	5	4	5	
(19,19)	1	1	3	4	4	2	6	6	1	5	4	4	6	5	1	5	6	1	5	1	0
∼	1	1	3	4	4	2	6	6	1	5	4	4	6	5	1	5	6	1	5	1	
(19,20)	4	1	3	3	4	1	3	5	5	2	1	6	6	5	4	1	2	2	4	6	17
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(19,21)	3	3	6	4	1	3	6	4	2	5	4	3	6	3	3	3	2	4	4	5	0
∼	3	3	6	4	1	3	6	4	2	5	4	3	6	3	3	3	2	4	4	5	
(19,30)	5	2	5	6	5	3	6	2	2	1	4	6	1	1	1	6	5	1	6	4	0
∼	5	2	5	6	5	3	6	2	2	1	4	6	1	1	1	6	5	1	6	4	
(19,31)	2	6	6	3	1	6	2	2	2	5	1	4	1	4	6	2	6	5	2	1	0
∼	2	6	6	3	1	6	2	2	2	5	1	4	1	4	6	2	6	5	2	1	
(19,32)	5	3	4	2	2	3	4	5	2	1	2	1	4	1	3	6	6	2	4	1	0
∼	5	3	4	2	2	3	4	5	2	1	2	1	4	1	3	6	6	2	4	1	
(19,34)	5	2	4	2	3	5	4	6	4	1	3	4	1	3	4	2	1	5	4	5	0
∼	5	2	4	2	3	5	4	6	4	1	3	4	1	3	4	2	1	5	4	5	
(19,35)	4	6	4	3	5	3	5	4	5	4	1	2	5	5	3	5	2	1	5	3	0
∼	4	6	4	3	5	3	5	4	5	4	1	2	5	5	3	5	2	1	5	3	

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(19,36)	4	2	1	6	2	6	4	1	2	5	3	3	6	5	2	5	2	5	4	6	0
∼	4	2	1	6	2	6	4	1	2	5	3	3	6	5	2	5	2	5	4	6	
(19,37)	3	6	1	1	5	6	3	6	3	6	5	2	2	1	6	6	5	4	6	2	0
∼	3	6	1	1	5	6	3	6	3	6	5	2	2	1	6	6	5	4	6	2	
(19,38)	2	2	4	3	1	3	1	1	2	4	4	4	6	3	1	3	4	4	2	3	7
∼	5	5	4	3	5	3	1	1	2	4	4	4	1	4	4	3	4	4	2	4	
(19,39)	4	6	4	3	2	3	4	5	6	3	3	2	4	2	5	2	4	5	3	4	0
∼	4	6	4	3	2	3	4	5	6	3	3	2	4	2	5	2	4	5	3	4	
(22,10)	5	4	3	4	2	1	5	2	5	3	1	2	4	2	2	3	6	4	4	6	2
∼	5	4	3	4	2	1	5	2	5	3	2	2	5	2	2	3	6	4	4	6	
(22,11)	2	5	2	5	6	5	2	3	6	3	4	6	1	3	1	1	5	5	2	1	15
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(22,17)	2	3	3	1	4	6	4	2	1	2	4	3	2	2	6	6	2	4	2	2	0
∼	2	3	3	1	4	6	4	2	1	2	4	3	2	2	6	6	2	4	2	2	
(22,18)	1	6	1	6	5	6	5	2	4	4	2	1	5	3	6	3	6	1	6	2	6
∼	4	5	1	6	5	5	5	2	4	5	5	1	5	4	6	3	6	1	6	2	
(22,19)	5	5	2	2	6	6	3	3	6	3	3	6	3	2	6	1	3	5	4	6	0
∼	5	5	2	2	6	6	3	3	6	3	3	6	3	2	6	1	3	5	4	6	
(22,28)	4	1	2	6	6	6	6	2	1	4	6	5	2	2	3	3	6	6	5	5	1
∼	4	1	4	6	6	6	6	2	1	4	6	5	2	2	3	3	6	6	5	5	
(22,29)	5	4	5	1	5	2	4	1	2	5	6	5	3	6	4	6	3	6	3	5	0
∼	5	4	5	1	5	2	4	1	2	5	6	5	3	6	4	6	3	6	3	5	
(22,30)	1	3	3	2	3	3	6	2	5	6	5	6	3	3	3	6	1	2	4	5	0
∼	1	3	3	2	3	3	6	2	5	6	5	6	3	3	3	6	1	2	4	5	
(22,32)	6	4	6	4	4	4	3	6	4	6	6	5	5	4	5	2	1	1	3	2	0
∼	6	4	6	4	4	4	3	6	4	6	6	5	5	4	5	2	1	1	3	2	
(22,33)	5	5	4	4	5	1	6	5	2	5	2	1	5	1	5	2	5	4	1	5	7
∼	5	5	4	4	5	5	5	5	5	5	2	5	5	5	5	5	5	4	5	5	
(22,34)	5	4	4	5	4	4	1	1	2	6	1	1	5	2	1	2	2	6	1	2	0
∼	5	4	4	5	4	4	1	1	2	6	1	1	5	2	1	2	2	6	1	2	
(22,35)	1	1	2	6	1	3	1	5	1	4	6	3	5	3	2	4	1	3	5	5	0

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	1	1	2	6	1	3	1	5	1	4	6	3	5	3	2	4	1	3	5	5	
(22,36)	2	5	3	6	3	5	2	5	6	5	6	6	6	4	5	5	2	3	4	6	5
∼	2	5	3	5	3	5	2	5	5	5	6	5	5	4	5	5	2	3	4	5	
(22,37)	6	1	3	3	4	2	6	2	5	4	6	6	2	1	6	1	6	2	1	4	4
∼	6	5	3	3	4	2	5	2	5	4	6	6	2	1	5	1	5	2	1	4	
(25,10)	3	5	4	2	4	3	3	2	3	2	5	3	1	2	3	6	4	6	1	6	2
∼	3	5	4	2	4	3	3	5	3	5	5	3	1	2	3	6	4	6	1	6	
(25,11)	4	5	5	2	2	4	3	5	5	4	1	2	4	4	1	1	4	4	2	3	4
∼	5	5	5	5	2	5	3	5	5	5	1	2	4	4	1	1	4	4	2	3	
(25,12)	3	1	2	3	6	1	4	3	4	1	2	3	2	1	3	3	3	6	2	1	5
∼	3	5	2	3	5	1	4	3	4	1	4	3	2	1	3	3	5	5	2	1	
(25,18)	4	4	5	5	3	5	6	1	6	5	3	5	3	1	4	1	5	4	2	1	0
∼	4	4	5	5	3	5	6	1	6	5	3	5	3	1	4	1	5	4	2	1	
(25,19)	2	3	2	5	3	6	5	1	4	2	3	4	6	4	2	1	2	4	5	2	0
∼	2	3	2	5	3	6	5	1	4	2	3	4	6	4	2	1	2	4	5	2	
(25,20)	5	2	6	6	5	2	2	6	5	3	6	5	2	2	1	1	4	1	5	6	9
∼	6	5	5	6	3	3	2	6	5	5	6	4	3	2	2	1	4	1	5	6	
(25,29)	1	3	2	1	4	2	3	3	6	3	1	5	6	3	3	4	6	4	1	4	0
∼	1	3	2	1	4	2	3	3	6	3	1	5	6	3	3	4	6	4	1	4	
(25,30)	5	4	4	1	6	4	2	4	3	5	3	5	5	5	4	1	1	1	6	2	0
∼	5	4	4	1	6	4	2	4	3	5	3	5	5	5	4	1	1	1	6	2	
(25,31)	3	4	4	4	2	4	4	1	5	5	3	2	2	3	5	1	1	4	1	3	0
∼	3	4	4	4	2	4	4	1	5	5	3	2	2	3	5	1	1	4	1	3	
(25,33)	2	6	2	3	5	2	1	4	1	6	2	3	2	6	4	4	4	3	4	5	0
∼	2	6	2	3	5	2	1	4	1	6	2	3	2	6	4	4	4	3	4	5	
(25,34)	6	4	6	3	3	4	6	1	1	6	3	4	2	5	5	1	4	2	5	6	1
∼	6	4	5	3	3	4	6	1	1	6	3	4	2	5	5	1	4	2	5	6	
(25,35)	5	5	6	5	1	1	6	3	1	3	6	3	6	4	5	6	2	6	4	1	0
∼	5	5	6	5	1	1	6	3	1	3	6	3	6	4	5	6	2	6	4	1	
(25,36)	3	4	5	2	3	1	4	4	5	3	4	3	4	2	2	6	5	3	5	3	0
∼	3	4	5	2	3	1	4	4	5	3	4	3	4	2	2	6	5	3	5	3	

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(25,37)	2	6	4	6	3	4	4	5	2	5	4	2	2	3	6	1	3	1	1	1	0
∼	2	6	4	6	3	4	4	5	2	5	4	2	2	3	6	1	3	1	1	1	
(25,9)	5	1	4	1	2	3	6	3	3	6	1	1	2	2	6	6	2	2	3	5	0
∼	5	1	4	1	2	3	6	3	3	6	1	1	2	2	6	6	2	2	3	5	
(26,11)	5	3	6	2	2	1	2	3	3	6	6	5	4	4	5	4	6	6	6	4	0
∼	5	3	6	2	2	1	2	3	3	6	6	5	4	4	5	4	6	6	6	4	
(26,12)	2	4	2	2	3	2	5	3	6	6	3	3	2	4	4	6	4	1	6	4	0
∼	2	4	2	2	3	2	5	3	6	6	3	3	2	4	4	6	4	1	6	4	
(26,13)	2	2	3	5	2	4	1	6	5	6	1	6	2	3	1	4	3	2	1	3	0
∼	2	2	3	5	2	4	1	6	5	6	1	6	2	3	1	4	3	2	1	3	
(26,24)	4	6	4	5	5	3	2	1	3	1	3	5	5	4	4	3	1	5	2	1	0
∼	4	6	4	5	5	3	2	1	3	1	3	5	5	4	4	3	1	5	2	1	
(26,25)	6	3	1	1	6	4	6	3	5	1	1	3	3	2	2	1	4	4	1	1	0
∼	6	3	1	1	6	4	6	3	5	1	1	3	3	2	2	1	4	4	1	1	
(26,26)	1	4	1	5	6	3	2	4	5	3	4	3	6	4	3	3	1	2	4	4	0
∼	1	4	1	5	6	3	2	4	5	3	4	3	6	4	3	3	1	2	4	4	
(26,28)	4	4	1	6	2	5	5	1	4	4	1	2	3	5	6	1	2	6	6	4	0
∼	4	4	1	6	2	5	5	1	4	4	1	2	3	5	6	1	2	6	6	4	
(26,29)	5	3	3	3	1	6	1	3	5	2	3	6	5	4	6	5	5	1	6	6	0
∼	5	3	3	3	1	6	1	3	5	2	3	6	5	4	6	5	5	1	6	6	
(26,30)	2	1	1	2	3	3	4	1	2	5	4	6	6	3	5	5	5	2	2	4	0
∼	2	1	1	2	3	3	4	1	2	5	4	6	6	3	5	5	5	2	2	4	
(26,31)	4	6	4	1	3	4	5	2	3	2	4	6	4	6	3	4	6	2	3	3	0
∼	4	6	4	1	3	4	5	2	3	2	4	6	4	6	3	4	6	2	3	3	
(26,32)	5	2	6	6	4	4	6	2	3	4	6	1	1	6	3	6	2	2	1	4	0
∼	5	2	6	6	4	4	6	2	3	4	6	1	1	6	3	6	2	2	1	4	
(26,33)	4	3	1	5	1	2	4	1	6	3	1	6	4	4	3	4	1	3	1	2	0
∼	4	3	1	5	1	2	4	1	6	3	1	6	4	4	3	4	1	3	1	2	
(26,8)	2	4	3	6	5	1	3	6	6	3	3	4	4	2	1	3	1	4	3	6	0
∼	2	4	3	6	5	1	3	6	6	3	3	4	4	2	1	3	1	4	3	6	
(29,10)	5	4	4	5	2	4	6	4	2	3	6	2	5	1	4	6	1	6	5	4	2

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	5	4	4	5	2	4	6	4	5	3	6	4	5	1	4	6	1	6	5	4	
(29,11)	5	4	2	6	2	1	1	4	4	1	6	3	4	2	6	5	1	2	6	1	1
∼	5	4	2	6	2	1	1	4	4	1	6	3	4	2	6	6	1	2	6	1	
(29,18)	2	3	2	5	5	4	2	1	3	3	5	2	4	6	2	1	4	3	3	2	0
∼	2	3	2	5	5	4	2	1	3	3	5	2	4	6	2	1	4	3	3	2	
(29,19)	5	4	5	5	1	2	3	5	5	3	1	1	5	6	5	5	1	5	4	4	0
∼	5	4	5	5	1	2	3	5	5	3	1	1	5	6	5	5	1	5	4	4	
(29,20)	6	2	3	2	6	5	1	3	4	3	3	5	5	3	3	1	4	4	1	4	0
∼	6	2	3	2	6	5	1	3	4	3	3	5	5	3	3	1	4	4	1	4	
(29,29)	3	1	6	5	5	2	1	2	5	3	4	1	1	4	1	1	6	1	4	6	0
∼	3	1	6	5	5	2	1	2	5	3	4	1	1	4	1	1	6	1	4	6	
(29,30)	5	2	6	6	6	5	5	3	3	6	4	1	1	1	3	6	3	2	5	5	0
∼	5	2	6	6	6	5	5	3	3	6	4	1	1	1	3	6	3	2	5	5	
(29,31)	3	2	5	3	6	4	5	1	2	3	2	2	1	4	1	3	4	6	5	4	0
∼	3	2	5	3	6	4	5	1	2	3	2	2	1	4	1	3	4	6	5	4	
(29,33)	4	2	6	6	4	3	2	2	3	3	5	6	2	1	2	6	4	6	1	5	5
∼	4	2	5	6	4	3	2	2	5	5	5	6	2	1	2	5	4	5	1	5	
(29,34)	3	4	4	1	2	2	2	1	1	3	4	3	3	4	3	4	4	1	4	2	0
∼	3	4	4	1	2	2	2	1	1	3	4	3	3	4	3	4	4	1	4	2	
(29,35)	4	1	4	5	4	6	3	1	3	6	6	2	4	3	6	4	2	2	4	5	0
∼	4	1	4	5	4	6	3	1	3	6	6	2	4	3	6	4	2	2	4	5	
(29,36)	1	2	3	2	5	1	2	3	5	2	5	6	5	5	3	4	4	6	5	3	1
∼	1	4	3	2	5	1	2	3	5	2	5	6	5	5	3	4	4	6	5	3	
(29,37)	4	5	1	6	4	2	1	5	5	5	4	4	2	1	2	4	1	1	4	3	0
∼	4	5	1	6	4	2	1	5	5	5	4	4	2	1	2	4	1	1	4	3	
(29,8)	5	3	6	3	3	1	4	4	5	2	3	2	2	5	6	1	3	4	5	6	0
∼	5	3	6	3	3	1	4	4	5	2	3	2	2	5	6	1	3	4	5	6	
(29,9)	6	4	1	2	6	6	6	4	6	5	6	2	4	5	3	6	3	2	5	2	12
∼	5	4	5	4	3	5	3	4	5	5	4	4	4	5	5	4	3	2	5	4	
(5,10)	2	5	1	1	4	4	1	1	3	4	4	3	4	5	2	1	6	3	3	5	0
∼	2	5	1	1	4	4	1	1	3	4	4	3	4	5	2	1	6	3	3	5	

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(5,12)	1	5	2	6	3	5	5	6	3	3	3	6	3	2	2	5	2	3	2	4	16
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(5,13)	5	1	5	1	1	5	2	1	1	5	3	6	2	5	3	1	1	5	3	6	0
∼	5	1	5	1	1	5	2	1	1	5	3	6	2	5	3	1	1	5	3	6	
(5,20)	4	6	4	3	2	5	3	5	6	4	6	6	2	6	3	3	4	2	3	5	14
∼	5	5	5	5	4	5	3	5	5	4	5	5	3	5	4	5	4	5	5	5	
(5,21)	4	2	4	4	3	2	3	4	4	6	4	2	2	3	6	4	3	3	1	6	0
∼	4	2	4	4	3	2	3	4	4	6	4	2	2	3	6	4	3	3	1	6	
(5,22)	4	5	3	2	2	5	5	4	4	1	2	1	4	1	1	5	3	2	1	5	0
∼	4	5	3	2	2	5	5	4	4	1	2	1	4	1	1	5	3	2	1	5	
(5,34)	5	2	5	3	5	2	2	6	1	2	5	1	5	3	1	6	2	3	3	6	0
∼	5	2	5	3	5	2	2	6	1	2	5	1	5	3	1	6	2	3	3	6	
(5,35)	5	2	2	2	5	5	5	6	6	6	3	4	3	4	2	6	5	6	1	5	0
∼	5	2	2	2	5	5	5	6	6	6	3	4	3	4	2	6	5	6	1	5	
(5,36)	1	5	5	4	1	1	5	4	4	4	6	4	2	1	1	3	3	5	3	4	15
∼	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	
(5,38)	1	5	5	6	1	4	4	2	5	4	2	6	4	5	1	6	2	2	4	3	0
∼	1	5	5	6	1	4	4	2	5	4	2	6	4	5	1	6	2	2	4	3	
(5,39)	4	1	6	4	6	4	3	3	6	4	1	4	4	5	2	5	4	1	6	4	0
∼	4	1	6	4	6	4	3	3	6	4	1	4	4	5	2	5	4	1	6	4	
(5,40)	4	4	4	6	6	5	4	4	3	2	5	5	3	6	2	2	6	1	3	4	0
∼	4	4	4	6	6	5	4	4	3	2	5	5	3	6	2	2	6	1	3	4	
(5,41)	6	1	2	4	6	5	4	5	6	6	6	6	6	1	2	5	2	1	6	5	0
∼	6	1	2	4	6	5	4	5	6	6	6	6	6	1	2	5	2	1	6	5	
(5,42)	2	1	3	1	6	2	2	2	4	2	6	6	2	1	4	6	6	4	3	6	0
∼	2	1	3	1	6	2	2	2	4	2	6	6	2	1	4	6	6	4	3	6	
(5,43)	5	3	3	6	5	1	5	2	5	2	4	1	6	4	5	2	6	5	2	1	0
∼	5	3	3	6	5	1	5	2	5	2	4	1	6	4	5	2	6	5	2	1	
(6,10)	3	6	1	5	6	2	3	6	2	5	1	5	1	2	5	1	4	2	4	5	0
∼	3	6	1	5	6	2	3	6	2	5	1	5	1	2	5	1	4	2	4	5	
(6,11)	4	4	2	1	1	6	1	2	4	5	5	2	3	5	3	4	2	1	3	5	0

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	4	4	2	1	1	6	1	2	4	5	5	2	3	5	3	4	2	1	3	5	
(6,12)	3	1	6	1	3	6	1	3	2	3	6	3	3	1	3	2	1	1	1	1	4
∼	3	1	6	5	5	5	1	3	2	3	5	3	3	1	3	2	1	1	1	1	
(6,19)	3	4	2	5	1	5	4	5	2	2	4	3	4	5	2	5	6	6	4	3	0
∼	3	4	2	5	1	5	4	5	2	2	4	3	4	5	2	5	6	6	4	3	
(6,20)	6	2	3	6	1	2	1	5	3	4	1	4	6	3	4	5	4	4	3	6	2
∼	6	2	3	6	1	2	1	5	3	4	1	4	6	3	4	5	4	4	4	5	
(6,21)	5	2	4	1	4	3	3	2	5	1	3	3	3	3	1	3	4	5	1	1	0
∼	5	2	4	1	4	3	3	2	5	1	3	3	3	3	1	3	4	5	1	1	
(6,34)	2	5	6	6	3	4	3	3	1	6	1	6	6	2	5	5	2	4	4	1	0
∼	2	5	6	6	3	4	3	3	1	6	1	6	6	2	5	5	2	4	4	1	
(6,35)	5	4	3	1	2	3	6	5	2	5	4	2	5	4	1	3	2	3	3	2	0
∼	5	4	3	1	2	3	6	5	2	5	4	2	5	4	1	3	2	3	3	2	
(6,36)	4	2	6	6	6	1	2	4	3	3	3	3	2	6	3	4	5	6	4	1	0
∼	4	2	6	6	6	1	2	4	3	3	3	3	2	6	3	4	5	6	4	1	
(6,38)	6	5	5	4	3	2	6	2	1	3	4	2	1	4	2	2	6	6	2	4	0
∼	6	5	5	4	3	2	6	2	1	3	4	2	1	4	2	2	6	6	2	4	
(6,39)	3	3	1	6	5	4	6	2	4	1	1	6	3	3	2	6	3	6	1	4	18
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	
(6,40)	2	3	2	4	5	6	4	3	5	1	5	5	6	6	1	5	4	3	5	6	6
∼	2	3	5	4	5	5	4	3	5	5	5	5	5	5	1	5	4	3	5	5	
(6,41)	3	6	3	2	4	5	6	2	6	5	1	4	2	2	5	4	2	2	6	1	1
∼	3	6	3	2	4	5	6	2	5	5	1	4	2	2	5	4	2	2	6	1	
(6,42)	5	6	5	6	2	4	4	2	1	6	2	2	3	5	3	3	1	2	3	6	0
∼	5	6	5	6	2	4	4	2	1	6	2	2	3	5	3	3	1	2	3	6	
(6,43)	5	3	6	4	4	5	2	2	5	6	4	3	6	3	5	3	4	3	5	5	8
∼	5	3	5	5	4	5	5	5	5	5	4	3	5	5	5	5	4	3	5	5	
(6,9)	3	2	2	3	1	3	5	5	2	1	2	4	4	3	4	3	2	1	1	3	0
∼	3	2	2	3	1	3	5	5	2	1	2	4	4	3	4	3	2	1	1	3	
(9,11)	3	2	2	2	5	6	1	1	6	1	2	6	4	2	6	5	6	5	1	5	3
∼	3	2	5	2	5	5	1	2	6	1	2	6	4	2	6	5	6	5	1	5	

Continued on next page

Table C.2 – *Cold water treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(9,12)	2	5	4	3	1	1	5	4	1	2	4	1	4	2	5	1	3	3	4	3	0
∼	2	5	4	3	1	1	5	4	1	2	4	1	4	2	5	1	3	3	4	3	
(9,13)	6	2	1	1	2	4	4	3	6	3	6	4	1	1	5	3	3	2	1	1	0
∼	6	2	1	1	2	4	4	3	6	3	6	4	1	1	5	3	3	2	1	1	
(9,14)	6	4	5	6	4	1	1	1	4	3	3	6	4	3	5	2	2	6	5	3	0
∼	6	4	5	6	4	1	1	1	4	3	3	6	4	3	5	2	2	6	5	3	
(9,20)	4	6	2	2	2	5	1	5	5	1	5	1	3	3	3	2	3	2	1	4	0
∼	4	6	2	2	2	5	1	5	5	1	5	1	3	3	3	2	3	2	1	4	
(9,21)	5	1	1	6	3	1	3	5	6	3	1	1	4	6	4	6	3	5	6	5	1
∼	5	1	1	6	3	1	3	5	6	3	1	1	4	6	4	5	3	5	6	5	
(9,22)	3	6	2	2	4	6	6	5	5	3	4	5	4	1	1	2	6	2	6	6	0
∼	3	6	2	2	4	6	6	5	5	3	4	5	4	1	1	2	6	2	6	6	
(9,34)	6	5	3	5	6	1	2	2	6	1	5	1	3	4	5	3	2	3	4	6	0
∼	6	5	3	5	6	1	2	2	6	1	5	1	3	4	5	3	2	3	4	6	
(9,35)	2	5	4	2	5	6	1	3	2	4	2	6	1	3	6	2	4	1	1	1	4
∼	2	5	4	2	5	6	1	5	5	4	2	5	1	3	5	2	4	1	1	1	
(9,36)	2	6	2	3	5	4	2	2	3	3	6	6	3	3	6	4	3	1	3	5	0
∼	2	6	2	3	5	4	2	2	3	3	6	6	3	3	6	4	3	1	3	5	
(9,38)	6	2	2	5	2	2	5	6	5	2	2	1	6	5	5	2	4	5	6	2	0
∼	6	2	2	5	2	2	5	6	5	2	2	1	6	5	5	2	4	5	6	2	
(9,39)	5	5	3	1	1	5	3	4	2	1	6	6	2	5	1	4	5	5	1	4	0
∼	5	5	3	1	1	5	3	4	2	1	6	6	2	5	1	4	5	5	1	4	
(9,40)	3	2	5	4	3	3	1	5	3	2	3	3	4	2	1	6	6	5	6	1	0
∼	3	2	5	4	3	3	1	5	3	2	3	3	4	2	1	6	6	5	6	1	
(9,41)	3	3	1	5	1	3	2	6	5	5	4	1	6	1	6	5	6	5	6	6	0
∼	3	3	1	5	1	3	2	6	5	5	4	1	6	1	6	5	6	5	6	6	
(9,42)	4	4	6	6	2	5	6	1	3	5	3	1	4	2	5	3	4	6	4	4	0
∼	4	4	6	6	2	5	6	1	3	5	3	1	4	2	5	3	4	6	4	4	

C.2.3 No water treatment

Table C.3: Dice rolling data for no treatment.

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(12,18)	4	3	4	5	6	6	4	4	1	1	5	3	5	2	2	2	1	2	5	6	0
∼	4	3	4	5	6	6	4	4	1	1	5	3	5	2	2	2	1	2	5	6	
(12,19)	4	6	4	4	6	2	5	1	4	4	6	3	5	1	1	6	6	5	2	5	1
∼	4	6	4	4	6	2	5	1	4	4	6	3	5	1	1	6	6	5	5	5	
(12,2)	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	14
∼	5	4	5	5	5	5	5	5	5	4	5	4	5	4	5	5	5	5	4	5	
(12,20)	3	5	4	5	1	4	1	4	5	5	4	1	5	4	4	2	1	5	5	4	0
∼	3	5	4	5	1	4	1	4	5	5	4	1	5	4	4	2	1	5	5	4	
(12,4)	5	4	6	1	1	2	4	5	6	4	1	4	4	3	4	4	2	5	4	3	0
∼	5	4	6	1	1	2	4	5	6	4	1	4	4	3	4	4	2	5	4	3	
(12,5)	2	2	5	2	2	3	5	5	2	6	3	3	5	6	4	3	3	2	2	1	0
∼	2	2	5	2	2	3	5	5	2	6	3	3	5	6	4	3	3	2	2	1	
(12,7)	1	2	5	1	5	1	5	5	2	1	5	5	6	6	5	4	2	6	1	3	0
∼	1	2	5	1	5	1	5	5	2	1	5	5	6	6	5	4	2	6	1	3	
(15,18)	6	1	6	2	2	3	4	1	2	6	5	2	5	1	6	1	3	2	6	4	1
∼	6	1	6	2	2	3	4	1	2	6	5	2	5	1	6	1	3	2	5	4	
(15,19)	5	1	2	1	3	3	3	2	5	3	5	5	1	1	5	2	6	4	3	4	0
∼	5	1	2	1	3	3	3	2	5	3	5	5	1	1	5	2	6	4	3	4	
(15,2)	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	0
∼	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	
(15,20)	6	1	6	2	3	4	1	2	5	5	3	4	1	5	1	4	6	3	2	5	0
∼	6	1	6	2	3	4	1	2	5	5	3	4	1	5	1	4	6	3	2	5	
(15,4)	6	2	2	3	1	2	5	3	4	5	4	1	6	1	4	6	1	3	6	2	0
∼	6	2	2	3	1	2	5	3	4	5	4	1	6	1	4	6	1	3	6	2	
(15,5)	4	2	4	2	2	5	3	6	4	4	6	3	4	3	4	1	1	1	2	2	0
∼	4	2	4	2	2	5	3	6	4	4	6	3	4	3	4	1	1	1	2	2	
(15,8)	4	4	1	1	6	6	4	6	5	4	5	2	3	3	1	4	1	4	2	3	0
∼	4	4	1	1	6	6	4	6	5	4	5	2	3	3	1	4	1	4	2	3	
(16,19)	5	5	6	4	1	2	4	6	1	4	1	6	5	3	3	4	5	6	6	4	0
∼	5	5	6	4	1	2	4	6	1	4	1	6	5	3	3	4	5	6	6	4	

Continued on next page

Table C.3 – *No treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(16,2)	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	16
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(16,20)	3	6	3	6	2	6	6	2	4	1	5	6	5	6	4	5	6	4	1	2	0
∼	3	6	3	6	2	6	6	2	4	1	5	6	5	6	4	5	6	4	1	2	
(16,21)	2	6	4	4	6	1	1	5	1	6	2	6	2	5	2	2	4	3	3	5	6
∼	2	6	4	4	6	1	1	5	1	6	2	5	3	5	3	3	4	4	4	5	
(16,4)	4	3	3	1	3	6	3	5	4	3	1	3	1	3	1	1	4	3	5	6	0
∼	4	3	3	1	3	6	3	5	4	3	1	3	1	3	1	1	4	3	5	6	
(16,5)	4	6	4	5	3	4	4	3	4	6	3	3	2	2	5	6	2	5	6	6	0
∼	4	6	4	5	3	4	4	3	4	6	3	3	2	2	5	6	2	5	6	6	
(16,8)	2	2	3	6	1	4	2	2	3	5	2	6	5	2	5	1	4	3	5	3	6
∼	2	5	3	5	1	4	5	2	3	5	5	5	5	2	5	5	4	3	5	3	
(19,16)	2	5	2	3	2	5	2	6	4	1	6	1	5	2	5	3	3	2	5	2	6
∼	2	5	2	3	2	5	2	5	4	2	5	3	5	2	5	3	5	2	5	3	
(19,17)	1	4	4	4	2	6	2	1	6	1	5	2	2	1	2	2	6	6	1	4	0
∼	1	4	4	4	2	6	2	1	6	1	5	2	2	1	2	2	6	6	1	4	
(19,18)	3	2	3	4	3	3	5	5	1	3	6	6	5	5	2	2	6	5	2	5	0
∼	3	2	3	4	3	3	5	5	1	3	6	6	5	5	2	2	6	5	2	5	
(19,2)	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	13
∼	5	4	5	4	5	5	4	5	5	5	5	5	5	5	5	4	5	5	4	4	
(19,4)	5	6	5	4	5	1	5	1	5	2	5	1	4	3	5	5	1	2	3	6	12
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(22,15)	5	2	5	3	2	6	3	4	6	3	1	1	3	1	2	4	3	6	4	2	0
∼	5	2	5	3	2	6	3	4	6	3	1	1	3	1	2	4	3	6	4	2	
(22,16)	5	6	6	6	6	1	3	1	6	1	5	3	1	5	1	4	2	2	1	2	0
∼	5	6	6	6	6	1	3	1	6	1	5	3	1	5	1	4	2	2	1	2	
(22,2)	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	5
∼	6	4	4	2	5	5	2	3	6	5	5	2	5	2	5	3	3	3	4	4	
(22,4)	3	5	5	3	5	5	1	3	3	4	4	4	6	5	2	5	4	2	4	2	0
∼	3	5	5	3	5	5	1	3	3	4	4	4	6	5	2	5	4	2	4	2	
(22,6)	6	4	2	5	5	3	2	2	1	4	5	3	2	6	3	6	4	4	4	5	0

Continued on next page

Table C.3 – *No treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	6	4	2	5	5	3	2	2	1	4	5	3	2	6	3	6	4	4	4	5	
(25,15)	3	3	2	4	1	1	5	2	4	5	4	1	1	6	1	2	4	4	3	5	0
∼	3	3	2	4	1	1	5	2	4	5	4	1	1	6	1	2	4	4	3	5	
(25,16)	4	3	5	6	5	5	2	6	3	2	3	3	5	5	4	4	6	2	6	1	0
∼	4	3	5	6	5	5	2	6	3	2	3	3	5	5	4	4	6	2	6	1	
(25,17)	1	2	6	6	5	1	5	3	4	1	4	6	6	4	2	3	5	6	1	3	0
∼	1	2	6	6	5	1	5	3	4	1	4	6	6	4	2	3	5	6	1	3	
(25,2)	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	0
∼	6	4	3	2	6	5	2	1	6	6	5	2	5	2	5	3	3	2	4	4	
(25,5)	2	1	4	5	1	1	4	4	4	4	3	1	3	2	1	5	3	5	6	2	14
∼	5	5	4	5	5	5	5	4	4	5	5	5	5	4	5	5	5	5	5	5	
(26,10)	3	2	3	2	3	6	6	2	6	1	1	4	3	4	5	3	2	6	2	3	0
∼	3	2	3	2	3	6	6	2	6	1	1	4	3	4	5	3	2	6	2	3	
(26,4)	4	6	3	4	6	1	4	1	5	6	1	3	2	3	3	3	3	5	2	3	0
∼	4	6	3	4	6	1	4	1	5	6	1	3	2	3	3	3	3	5	2	3	
(26,7)	6	5	6	2	5	3	5	5	2	2	6	3	6	6	4	3	5	6	6	2	4
∼	6	5	5	2	5	3	5	5	2	4	6	3	6	5	4	3	5	6	5	2	
(26,9)	2	5	3	4	6	1	5	6	4	6	3	5	1	4	1	3	1	4	3	3	2
∼	2	5	3	4	6	1	6	6	4	6	3	6	1	4	1	3	1	4	3	3	
(27,2)	5	5	1	3	5	6	3	4	1	6	3	6	6	6	4	4	1	2	2	2	0
∼	5	5	1	3	5	6	3	4	1	6	3	6	6	6	4	4	1	2	2	2	
(29,15)	3	1	5	3	3	2	1	3	3	6	6	2	1	4	5	5	2	1	5	4	8
∼	3	5	5	4	3	5	1	5	5	4	5	2	1	4	5	5	2	5	5	4	
(29,16)	4	6	6	6	2	2	2	5	4	3	6	4	6	6	6	3	3	3	5	5	13
∼	4	6	6	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	
(29,17)	2	2	2	4	2	5	6	4	6	2	3	4	5	2	2	3	5	3	5	5	2
∼	2	2	2	4	2	5	5	4	6	2	3	4	5	2	4	3	5	3	5	5	
(29,4)	1	6	3	6	2	2	6	4	4	3	5	1	6	5	4	4	4	4	4	2	0
∼	1	6	3	6	2	2	6	4	4	3	5	1	6	5	4	4	4	4	4	2	
(2,10)	4	6	5	1	2	5	5	6	2	2	6	4	6	2	1	6	1	3	1	3	0
∼	4	6	5	1	2	5	5	6	2	2	6	4	6	2	1	6	1	3	1	3	

Continued on next page

Table C.3 – *No treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
(2,4)	1	4	5	2	6	4	4	5	2	5	4	2	5	1	3	2	2	5	6	1	12
∼	5	5	5	4	5	4	4	5	4	5	4	5	5	5	4	3	5	5	5	5	
(2,5)	4	2	6	3	6	2	3	2	1	1	1	5	3	5	4	6	3	5	6	5	0
∼	4	2	6	3	6	2	3	2	1	1	1	5	3	5	4	6	3	5	6	5	
(2,7)	3	2	4	3	6	3	4	5	6	5	6	2	6	6	6	5	5	5	6	3	0
∼	3	2	4	3	6	3	4	5	6	5	6	2	6	6	6	5	5	5	6	3	
(2,9)	2	6	1	2	5	1	5	2	3	3	5	1	1	4	2	2	3	2	3	2	0
∼	2	6	1	2	5	1	5	2	3	3	5	1	1	4	2	2	3	2	3	2	
(5,17)	2	5	6	5	5	1	4	2	1	1	1	6	2	1	5	3	1	3	6	6	2
∼	2	5	6	5	5	1	4	2	1	1	5	5	2	1	5	3	1	3	6	6	
(5,18)	3	2	1	5	6	4	1	5	4	6	4	6	5	6	6	6	3	6	2	2	0
∼	3	2	1	5	6	4	1	5	4	6	4	6	5	6	6	6	3	6	2	2	
(5,19)	2	5	5	2	4	6	4	3	3	5	5	2	6	2	6	4	1	1	6	5	0
∼	2	5	5	2	4	6	4	3	3	5	5	2	6	2	6	4	1	1	6	5	
(5,5)	6	4	4	6	3	6	3	3	4	3	6	2	2	6	3	6	1	4	4	5	7
∼	6	4	4	5	3	5	5	3	4	3	5	2	4	5	3	3	1	4	4	5	
(6,16)	4	4	5	4	3	5	4	6	2	4	3	5	4	4	1	3	3	3	5	4	12
∼	5	4	5	5	4	5	5	5	4	4	5	5	4	5	2	5	5	5	5	4	
(6,17)	6	3	3	1	2	6	3	2	1	3	4	1	1	4	5	2	6	4	3	3	0
∼	6	3	3	1	2	6	3	2	1	3	4	1	1	4	5	2	6	4	3	3	
(6,18)	5	5	6	5	2	5	5	4	5	3	3	5	2	3	6	4	6	6	6	2	0
∼	5	5	6	5	2	5	5	4	5	3	3	5	2	3	6	4	6	6	6	2	
(6,3)	2	6	1	6	5	5	4	1	1	5	4	1	1	2	1	1	1	1	2	2	0
∼	2	6	1	6	5	5	4	1	1	5	4	1	1	2	1	1	1	1	2	2	
(6,5)	1	3	3	3	4	2	3	2	6	2	2	2	4	1	4	1	2	6	6	4	0
∼	1	3	3	3	4	2	3	2	6	2	2	2	4	1	4	1	2	6	6	4	
(9,1)	1	3	3	6	4	4	2	4	3	2	2	3	6	5	1	3	5	1	4	3	16
∼	1	4	4	5	5	4	3	5	4	4	3	5	5	5	5	4	5	2	5	4	
(9,18)	6	1	6	4	1	3	4	1	1	3	2	5	2	4	4	3	1	6	2	6	0
∼	6	1	6	4	1	3	4	1	1	3	2	5	2	4	4	3	1	6	2	6	
(9,19)	5	3	5	6	3	1	4	2	2	2	1	5	2	5	2	5	3	3	5	6	14

Continued on next page

Table C.3 – *No treatment, continued from previous page*

part.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	L
∼	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	*
(9,3)	4	3	6	1	6	1	2	3	5	2	2	5	4	3	5	3	1	2	4	2	1
∼	4	3	6	2	6	1	2	3	5	2	2	5	4	3	5	3	1	2	4	2	
(9,4)	3	4	6	1	2	1	2	5	2	4	3	5	5	6	5	4	5	4	3	5	3
∼	3	4	5	5	2	1	2	5	2	4	3	5	5	5	5	4	5	4	3	5	
(9,7)	2	5	4	3	1	5	1	5	4	3	4	1	5	6	1	4	4	2	3	1	5
∼	2	5	4	3	5	5	1	5	4	3	4	4	5	5	1	4	4	2	5	5	

Chapter 8

Bibliography

Johannes Abeler, Daniele Nosenzo, and Collin Raymond. Preferences for truth-telling. *Econometrica*, 87:1115–1153, 2019.

Andra Adams, Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson. Decoupling facial expressions and head motions in complex emotions. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII*, pages 274–280, 2015.

Shivani Agarwal, Aatif Awan, and Dan Roth. Learning to detect objects in images via a sparse, part-based representation. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1475–1490, 2004.

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Workshops*, page 8, 2019.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, 41:273–287, 2007.

Zara Ambadar, Jonathan W. Schooler, and Jeffrey F. Cohn. Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, 16:403–410, 2005.

APA Editorial. The truth about lie detectors (aka polygraph tests). *American Psychological Association*, 5 August 2004. www.apa.org/research/action/polygraph.

Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, pages 1–6, 2015.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, pages 1–10, 2016.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 59–66, 2018.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marcella, Alcix M. Martinez, and Seth D. Pollack. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movement. *Psychological Science in the Public Interest*, 20:1–68, 2019.

Marian Stewart Bartlett, Gwen Littlewort, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *2004 IEEE International Conference on Systems, Man and Cybernetics*, volume 1, pages 592–597, 2004.

Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, pages 568–573, 2005.

Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, and Kang Lee. Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, 24:738–743, 2014.

John N. Bassili. Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance*, 4:373–379, 1978.

Vaughan Bell. The truth about lie detectors. *The Guardian, England*, 21 April 2012.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, New York, 1st (corrected at 8th printing, 2009) edition, 2006.

Jake Bittle. Lie detectors have always been suspect. AI has made the problem worse. *MIT Technology Review*, March 2020.

Owen Bowcott. Lie detectors should be used to monitor sex offenders, UK study says. *The Guardian, England*, 18 March 2020.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347:145–149, 2015.

Tim Bradford. What if lie detectors were brought into politics? *The New European*, 21 January 2020.

Margaret Bradley and Peter Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25:49–59, 1994.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Routledge, New York, 1st edition, 1984.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359:418–424, 2018.

Penelope Brown and Stephen C. Levinson. *Politeness: Some universals in language usage (Vol. 4)*. Cambridge University Press, Cambridge, UK, 1987.

Alessandro Bucciol and Marco Piovesan. Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology*, 32:73–78, 2011.

Judee K. Burgoon and David Buller. Interpersonal deception theory: Reflections on the nature of theory building and the theoretical status of interpersonal deception theory. *Communication Theory*, 6:311–328, 1996.

Mei Chen, Tingyu Zhang, Ruqian Zhang, Ning Wang, Qing Yin, Yangzhuo Li, Jieqiong Liu, Tao Liu, and Xianchun Li. Neural alignment during face-to-face spontaneous deception: Does gender make a difference? *Human Brain Mapping*, doi: 10.1002/hbm.25173, 2020.

Thorsten Chmura, Christoph Engel, and Markus Englerth. At the mercy of a prisoner three dictator experiments. *Applied Economics Letters*, 24:774–778, 2017.

Cindy K. Chung and James W. Pennebaker. *Applied Natural Language Processing: Identification, Investigation and Resolution*, chapter Linguistic Inquiry and Word Count (LIWC): pronounced "Luke",... and other useful facts, pages 206–229. IGI Global, Hershey, PA, 2012.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Jeffrey F. Cohn and Fernando De La Torre. Automated face analysis for affective computing. In Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas, editors, *Oxford library of psychology. The Oxford handbook of affective computing*, pages 131–150. Oxford University Press, 2015.

Jeffrey F. Cohn and Karen L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multi-resolution & Information Processing*, 2:121–132, 2004.

Julian Conrads, Bernd Irlenbusch, Rainer Michael Rilke, and Gari Walkowitz. Lying and team incentives. *Journal of Economic Psychology*, 34:1–7, 2013.

Brice Corgnet and Roberto Hernán-González. Revisiting the trade-off between risk and incentives: The shocking effect of random shocks? *Management Science*, 65:1096–1114, 2019.

Brice Corgnet, Antonio M. Espin, Roberto Hernán-González, Praveen Kujal, and Stephen Rassenti. To trust, or not to trust: Cognitive reflection in trust games. *Journal of Behavioral and Experimental Economics*, 64:20–27, 2016.

Carlos Crivelli and Alan J. Fridlund. Facial displays are tools for social influence. *Trends in Cognitive Sciences*, 22:388–399, 2018.

Rachel Croson and Uri Gneezy. Gender differences in preferences. *Journal of Economic Literature*, 47:448–474, 2009.

Rick Dale and Nicholas D. Duran. The cognitive dynamics of negated sentence verification. *Cognitive Science*, 35:983–996, 2011.

Charles Darwin. *The Expression of the Emotions in Man and Animals, Anniversary Edition; edited by Paul Ekman*. Oxford University Press, U.S.A., 4th edition, 2009. first published 1872.

Paul K. Davis, Walter L. Perry, Ryan Andrew Brown, Douglas Yeung, Parisa Roshan, and Phoenix Voorhies. Using behavioral indicators to help detect potential violent acts: A review of the science base. Technical Report RR-215-NAVY, RAND Corporation, Santa Monica, CA, 2013.

David Demirdjian and Sybor Wang. Recognition of temporal events using multiscale bags of features. In *2009 IEEE Workshop on Computational Intelligence for Visual Intelligence*, pages 8–13, 2009.

Bell M. DePaulo, Deborah A. Kashy, Susan E. Kirkendol, Melissa M. Wyer, and Jennifer A. Epstein. Lying in everyday life. *Journal of Personality and Social Psychology*, 70: 979–995, 1996.

Bella DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological Bulletin*, 129:74–118, 2003.

Anthony Dickinson. Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 308:67–78, 1985.

Mingyu Ding, An Zhao, Zhiwu Lu, Tao Xiang, and Ji-Rong Wen. Face-focused cross-stream network for deception detection in videos. In *2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 2018.

Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. Are your participants gaming the system? Screening mechanical turk workers. In *SIGCHI Conference on Human Factors in Computing Systems, CHI '10*, pages 2399–2402, 2010.

Anna Dreber and Magnus Johannesson. Gender differences in deception. *Economics Letters*, 99:197–199, 2008.

Shichuan Du, Yong Tao, and Aleix M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111:E1454–E1462, 2014.

Guillaume-Benjamin Duchenne de Boulogne. *The Mechanism of Human Facial Expression; edited and translated by R. Andrew Cuthbertson*. Cambridge University Press, Cambridge, UK, 1990.

Damien Dupré, Nicole Andelic, Gawain Morrison, and Gary J. McKeown. Accuracy of three commercial automatic emotion recognition systems across different individuals and their facial expressions. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 627–632, 2018.

Damien Dupré, Eva G. Krumhuber, Cennis Küster, and Gary J. McKeown. A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLoS ONE*, 15:e0231968, 2020.

Ute Eberle. Phrenologie. *GEOkompakt*, 15:56–60, 2008.

Joy Egede, Michel Valstar, and Brais Martinez. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. In *12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017*, pages 689–696, 2017.

Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W. W. Norton & Company, New York, 1985.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

Paul Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000:205–221, 2003.

Paul Ekman. Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364:3449–3451, 2009.

Paul Ekman and Wallace V. Friesen. Detecting deception from the body or face. *Journal of Personality and Social Psychology*, 29:288–298, 1974.

Paul Ekman and Wallace V. Friesen. Facial action coding system: A technique for the measurement of facial movement. 1978.

Paul Ekman and Wallace V. Friesen. *Unmasking the face: A guilde to recognizing emotions from facial expressions*. Malor Books, Institute for the Study of Human Knowledge, Palo Alto, CA, 2003.

Paul Ekman and Maureen O'Sullivan. Who can catch a liar? *The American Psychologist*, 46:913–920, 1991.

Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. *Facial Action Coding System - The Manual on CD-ROM*, 2nd edition, 2002.

Sanjiv Erat and Uri Gneezy. White lies. *Management Science*, 58:723–733, 2012.

Iltir Onal Ertugrul, Jeffrey F. Cohn, László A. Jeni, Zheng Zhang, Lijung Yin, and Qiang Ji. Cross-domain au detection: Domains, learning approaches, and measures. In *14th IEEE International Conference on Automatic Face Gesture Recognition, FG 2019*, pages 1–8, 2019.

Eugene G. Ewaschuck. Hearsay evidence. *Osgoode Hall Law Journal*, 16:407–443, 1978.

Ralph V. Exline, John Thibaut, Carole B. Hickey, and Peter Gumpert. Visual interaction in relation to machiavellianism and an unethical act. In Richard Christie and Florence L. Geis, editors, *Studies in Machiavellianism*, pages 53–75. Academic Press, 1970.

Urs Fischbacher and Franziska Föllmi-Heusi. Lies in disguise–An experimental study on cheating. *Journal of the European Economic Association*, 11, 2013.

Johnny R. J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

Shane Frederick. Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19:25–42, 2005.

Jonathan Freeman, Kristin Pauker, and Diana T. Sanchez. A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, 27:502–517, 2016.

Wallace V. Friesen and Paul Ekman. Nonverbal leakage and clues to deception. *Psychiatry*, 32:88–106, 1969.

Simon Gächter and Jonathan F. Schulz. Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531:496–499, 2016.

Jeffrey M. Girard, Jeffrey F. Cohn, Mohammad H. Mahoor, Seyedmohammad Mavadati, and Dean P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 1–8, 2013.

Jeffrey M. Girard, Wen-Sheng Chu, László A. Jeni, Jeffrey F. Cohn, Fernando De la Torre, and Michael A. Sayette. Sayette group formation task (gft) spontaneous facial expression database. In *12th IEEE International Conference on Automatic Face Gesture Recognition, FG 2017*, pages 581–588, 2017.

Erving Goffman. *The Presentation of Self in Everyday Life*. Penguin, Harmondworth, UK, 1990.

Victor A. Gombos. The cognition of deception: The role of executive processes in producing lies. *Genetic, Social, and General Psychology Monographs*, 132:197–214, 2006.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2014.

Joshua Greene and Jonathan Haidt. How (and where) does moral judgement work? *Trends in Cognitive Sciences*, 6:517–523, 2002.

Jamie Grierson. Domestic abusers may face lie-detector test on release from prison. *The Guardian, England*, 21 January 2019.

Jamie Grierson. Lie-detector tests planned for convicted terrorists freed on licence. *The Guardian, England*, 21 January 2020.

Hatice Gunes and Maja Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1:68–99, 2010a.

Hatice Gunes and Maja Pantic. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In Jan Allbeck, Norman Badler, Timothy Bickmore, Catherine Pelachaud, and Alla Safonova, editors, *Intelligent Virtual Agents. IVA 2010. Lecture Notes in Computer Science, vol 6356*, pages 371–377. Springer Verlag, Berlin, Heidelberg, 2010b.

Hatice Gunes, Caifeng Chung, Shizhi Chen, and Yingli Tian. *Emotion Recognition: A Pattern Analysis Approach*, chapter Bodily Expression for Automatic Affect Recognition, pages 343–377. John Wiley & Sons, Inc., 2015.

Jonathan Haidt. The New Synthesis in Moral Psychology. *Science*, 316:998–1001, 2007.

Mark Harris. The lie generator: Inside the black mirror world of polygraph job screenings. *Wired*, 18 October 2018.

Christian L. Hart. Do lie detector tests really work? *Psychology Today*, 14 January 2020.

Md Kamrul Hasan, Taylan Sen, Raiyan Abdul Baten, Kurtis Glenn Haut, and Mohammed Ehsan Hoque. Liwc into the eyes: Using facial features to contextualize linguistic analysis in multimodal communication. In *8th International Conference on Affective Computing and Intelligent Interaction, ACII*, pages 1–7, 2019.

Edgar A. Hines and George E. Brown. A standard stimulus for measuring vasomotor reactions: Its application in the study of hypertension. *Staff Meetings of the Mayo Clinic*, 7:332–335, 1932.

Carl-Herman Hjortsjö. *Man's Face and Mimic Language; translated by W. Francis Salisbury*. Studentlitteratur, Lund, Sweden, 1970.

Rens Hoegen, Giota Stratou, and Jonathan Gratch. Incorporating emotion perception into opponent modeling for social dilemmas. In *16th Conference on Autonomous Agents and Multiagent Systems*, pages 801–809, 2017.

Walter S. Hulin and Daniel Katz. The Frois-Wittmann pictures of facial expression. *Journal of Experimental Psychology*, 18:482–490, 1935.

Ferris Jabr. 2 of a kind: Studies reveal new insights into the psychology of gambling. *Scientific American*, 12 August 2010.

Shashank Jaiswal. *Dynamic Deep Learning for Automatic Facial Expression Recognition and its Application in Diagnosis of ADHD & ASD*. PhD thesis, The University of Nottingham, Faculty of Science, School of Computer Science, 2018.

Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, pages 1–8, 2016.

Shashank Jaiswal, Michel F. Valstar, Alinda Gillott, and David Daley. Automatic detection of ADHD and ASD from expressive behaviour in RGBD data. In *12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017*, pages 762–769, 2017.

Seeun Jung and Radu Vranceanu. Experimental evidence on gender differences in lying behaviour. *Revue économique*, 68:859–873, 2017.

Takeo Kanade, Jeffrey F. Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000.

Dolores Kennedy. Exonerations before 1989: Floyed Fay. *The National Registry of Exonerations; a Project of the University of California Irvine Newkirk Center for Science & Society, University of Michigan Law School & Michigan State University College of Law*, 2020. www.law.umich.edu/special/exoneration/Pages/casedetailpre1989.aspx?caseid=94. Accessed 18 June 2020.

Keith Kirkpatrick. It's not the algorithm, it's the data. *Communications of the ACM*, 60: 21–23, 2017.

Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28:76–81, 1993.

Martin G. Kocher, Simeon Schudy, and Lisa Spantig. I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, 64:3995–4008, 2018.

Martin Krzywinski and Naomi Altman. Classification and regression trees. *Nature Methods*, 14:757–758, 2017.

Mark E. Laidre. How Often Do Animals Lie About Their Intentions? An Experimental Test. *The American Naturalist*, 173:337–346, 2009.

Mark E. Laidre and Rufus E. Johnstone. Animal signals. *Current Biology*, 23:R829–R833, 2013.

Peter J. Lang. The emotion probe: Studies of motivation and attention. *American Psychologist*, 50:372–385, 1995.

Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

Peter T. Leeson. Ordeals. *Journal of Law and Economics*, 55:691–714, 2012.

Roy J. Lewicki. Lying and deception: A behavioral model. In Max H. Bazerman and Roy J. Lewicki, editors, *Negotiation in Organizations*, pages 68–90. Sage Publications, Beverly Hills, CA, 1984.

Tobias Loetscher, Christopher Bockisch, Michael E.R. Nicholls, and Peter Brugger. Eye position predicts what number you have in mind. *Current Biology*, 20:R264–R265, 2010.

Matthew Lombard, Robert D. Reich, Maria Elizabeth Grabe, Cheryl Campanella Bracken, and Theresa Bolmarcich Ditton. Presence and television. *Human Communication Research*, 26:75–98, 2000.

Ryan Lu and Sarvesh Pantage. Emotion modelling in poker. Institute for Creative Technologies, University of Southern California, 2015.

Shan Lu, Gabriel Tsechpenakis, Dimitris N. Metaxas, Matthew L. Jensen, and John Kruse. Blob analysis of the head and hands: A method for deception detection. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 20c–20c, 2005.

David T. Lykken. The GSR in the detection of guilt. *Journal of Applied Psychology*, 43: 385–388, 1959.

Niccolo Machiavelli. *The Prince: The Original Classic; with an introduction by Tom Butler-Boden*. Capstone Publishing Ltd., Chichester, UK, 2010, reprinted August 2018.

Allison Marsh. A brief history of the lie detector. *IEEE Spectrum*, 31 July 2019.

Braise Martinez, Michel Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 10:325–347, 2019.

S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4:151–160, 2013.

Gary McKeown. Turing's menagerie: Talking lions, virtual bats, electric sheep and analogical peacocks: Common ground and common interest are necessary components of engagement. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII*, pages 950–955, 2015.

Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*, pages 1079–1084, 2010.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3:5–17, 2012.

Thomas O. Meservy, Matthew L. Jensen, John Kruse, Judee K. Burgoon, Jay F. Nunamaker, Douglas P. Twitchell, Gabriel Tsechpenakis, and Dimitris N. Metaxas. Deception detection through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems*, 20(5):36–43, 2005.

Tom M. Mitchell. *Machine Learning*. Mc Graw Hill, 1997.

Merylin Monaro, Luciano Gamberini, and Giuseppe Sartori. The detection of faked identity using unexpected questions and mouse dynamics. *PLoS ONE*, 12:e0177851, 2017.

David Moore, George McCabe, and Bruce Craig. *Introduction to the Practice of Statistics*. W. H. Freeman, 9th edition, 2017.

Matej Moravcik, Martin Schmid, Neil Burch, Viliam Lisy, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356:508–513, 2017.

Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79:299–318, 2008.

Maja Pantic. Machine analysis of facial behaviour: naturalistic and dynamic behaviour. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364:3505–3513, 2009.

Maja Pantic and Stewart Marian Bartlett. Machine analysis of facial expressions. In Kresimir Delac and Mislav Grgic, editors, *Face Recognition*. IntechOpen, Rijeka, 2007.

Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pages 317–321, 2005.

Sunghyun Park, Jonathan Gratch, and Louis-Philippe Morency. I already know your answer: Using nonverbal behaviors to predict immediate outcomes in a dyadic negotiation. In *14th ACM International Conference on Multimodal Interaction, ICMI '12*, pages 19–22, 2012.

Sunghyun Park, Stefan Scherer, Jonathan Gratch, Peter J. Carnevale, and Louis-Philippe Morency. I can already guess your answer: Predicting respondent reactions during dyadic negotiation. *IEEE Transactions on Affective Computing*, 6:86–96, 2015.

Brian Parkinson. Do facial movements express emotions or communicate motives? *Personality and Social Psychology Review*, 9:278–311, 2005.

Alex (Sandy) Pentland. *Honest Signals: How They Shape Our World; with contributions from Tracy Heilbeck*. MIT Press, Boston, MA, 2008.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. Deception detection using real-life trial data. In *2015 ACM International Conference on Multimodal Interaction*, pages 59–66, 2015.

Joseph Persky. Retrospectives: The ethology of homo economicus. *The Journal of Economic Perspectives*, 9:221–331, 1995.

Rosalind W. Picard. *Affective Computing*. MIT Press, Boston, MA, 2000.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–73. MIT Press, Boston, MA, 2000.

John A. Podlesny and David C. Raskin. Physiological measures and the detection of deception. *Psychological Bulletin*, 84:782–799, 1977.

Geovany A. Ramirez, Tadas Baltrušaitis, and Louis-Phillipe Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction. ACII 2011. Lecture Notes in Computer Science, vol 6975*, pages 396–406. Springer Verlag, Berlin, Heidelberg, 2011.

William E. Rinn. The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, 95:52–77, 1984.

David Roochnik. *Plato's Republic*. The Teaching Company, Chantilly, VA, 2005.

Erika L. Rosenberg, Paul Ekman, and James A. Blumenthal. Facial expression and the affective component of cynical hostility in male coronary heart disease patients. *Health Psychology*, 17:376–380, 1998.

Janet Rothwell, Zuhair Bandar, James Dominic O'Shea, and David McLean. Silent talker: A new computer-based system for the analysis of facial cues to deception. *Applied Cognitive Psychology*, 20:757–777, 2006.

Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W. Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3:eaao6760, 2018.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

James A. Russell. Core affect and the psychological construction of emotion. *Psychological Review*, 110:145–172, 2003.

Enrique Sánchez-Lozano, Georgios Tzimiropoulos, Brais Martinez, Fernando De la Torre, and Michel Valstar. A functional regression approach to facial landmark tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2037–2050, 2018.

Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for

relational reasoning. In *31st Conference on Neural Information Processing Systems, NIPS 2017*, pages 4967–4976, 2017.

Robert E. Schapire. A brief introduction to boosting. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*, pages 1401–1406, 1999.

Stefan Scherer, Giota Stratou, Gale Lucas, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert (Skip) Rizzo, and Louis-Philippe Morency. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32:648–658, 2014.

Erik J. Schlicht, Shinsuke Shimojo, Colin F. Camerer, Peter Battaglia, and Ken Nakayama. Human wagering behavior depends on opponents' faces. *PLoS ONE*, 5:e11663, 2010.

Harold Schlosberg. A scale for the judgement of facial expressions. *Journal of Experimental Psychology*, 29:497–510, 1941.

Harold Schlosberg. The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology*, 44:229–237, 1952.

Lars Schwabe and Hartmut Schächinger. Ten years of research with the socially evaluated cold pressor test: Data from the past and guidelines for the future. *Psychoneuroendocrinology*, 92:155–161, 2018.

Lars Schwabe and Oliver T. Wolf. Stress prompts habit behavior in humans. *The Journal of Neuroscience*, 29:7191–7198, 2009.

William A. Searcy and Stephen Nowicki. *The Evolution of Animal Communication: Reliability and Deception in Signaling Systems*. Princeton University Press, Princeton, N.J., 2005.

Taylan Sen, Md Kamrul Hasan, Zach Teicher, and Mohammed Ehsan Hoque. Automated dyadic data recorder (ADDR) framework and analysis of facial cues in deceptive communication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1, 2018.

R. Mark Sirkin. *Statistics for the social sciences*. Sage Publications, Inc., USA, 2006.

Mel Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 364:3549–3557, 2009.

Michael L. Slepian, Steven G. Young, Abraham M. Rutchik, and Nalini Ambady. Quality of professional players' poker hands is perceived accurately from arm motions. *Psychological Science*, 24:2335–2338, 2013.

David Smith. Craigslist's Craig Newmark: 'Outrage is profitable. Most online outrage is faked for profit'. *The Guardian, England*, 14 July 2019.

Alexander P. Spence. *Basic Human Anatomy*. The Benjamin-Cummings Publishing Company, San Francisco, CA, 2nd edition, 1990.

Katrin Starcke and Matthias Brand. Decision making under stress: A selective review. *Neuroscience and Biobehavioral Reviews*, 36:1228–1248, 2012.

Mariëlle Stel and Eric van Dijk. When do we see that others misrepresent how they feel? Detecting deception from emotional faces with direct and indirect measures. *Social Influence*, 13:137–149, 2018.

Giota Stratou, Rens Hoegen, Gale Lucas, and Jonathan Gratch. Emotional signaling in a social dilemma: An automatic analysis. In *2015 International Conference on Affective Computing and Intelligent Interaction, ACII*, pages 180–186, 2015.

Giota Stratou, Rens Hoegen, Gale Lucas, and Jonathan Gratch. Investigating gender differences in temporal dynamics during an iterated social dilemma: An automatic analysis using networks. In *2017 International Conference on Affective Computing and Intelligent Interaction, ACII*, pages 531–536, 2017a.

Giota Stratou, Job Van Der Schalk, Rens Hoegen, and Jonathan Gratch. Refactoring facial expressions: An automatic analysis of natural occurring facial expressions in iterative social dilemma. In *2017 International Conference on Affective Computing and Intelligent Interaction, ACII*, pages 427–433, 2017b.

Birgit Tanner. Die Wahrheit über die Lüge (TV documentary). *Arte*, 29 May 2019.

Leanne ten Brinke, Jooa Julia Lee, and Dana R. Carney. The physiology of (dis)honesty: Does it impact health? *Current Opinion in Psychology*, 6:177–182, 2015.

Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 9, 2018.

The Editors of Encyclopaedia Britannica. Phrenology. In *Encyclopaedia Britannica*, 1 May 2018.

The Editors of The Economist. Poker: A big deal. *The Economist*, 19 December 2007.

The Editors of Wikipedia. Polygraph. In *Wikipedia*, 10 August 2020.

Silvan S. Tomkins. *Affect, Imagery, Consciousness, Vol. 1. The Positive Affects*. Springer Verlag, New York, 2008. first piblished 1962.

Jessica L. Tracy and David Matsumoto. The spontaneous expression of pride and shame: Evidence for biologically innate nonverbal displays. *Proceedings of the National Academy of Sciences U.S.A.*, 105:11655–11660, 2008.

UK Test. Ukliedetectortest.co.uk. UK Lie Detector Test, 152-160 City Road, London, EC1V 2NX; accessed August 2020.

U.S. Government Accountability Office. Aviation security: TSA should limit future funding for behavior detection activities. Technical Report GAO-14-159, U.S. Government, 8 November 2013.

Vivian V. Valentin, Anthony Dickinson, and John P. O'Doherty. Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, 27: 4019–4026, 2007.

Michel Valstar. Automatic behaviour understanding in medicine. In *2014 Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges*, pages 57–60, 2014.

Michel Valstar and Maja Pantic. Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop, CVPRW '06*, page 149, 2006.

Michel Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F. Cohn. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *8th International Conference on Multimodal Interfaces*, pages 162–170, 2006.

Michel Valstar, Hatice Gunes, and Maja Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *9th International Conference on Multimodal Interfaces*, pages 38–45, 2007.

Michel Valstar, Timur Almaev, Jeffrey M. Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F. Cohn. FERA 2015 - second Facial Expression Recognition and Analysis Challenge. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, pages 1–8, 2015.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, Berlin, Heidelberg, 1995.

Louis Allen Vaught. *Vaught's Practical Character Reader*. Kessinger Publishing LLC, Whitefish, MT, 2010 (first published 1902).

Martina Vicianova. Historical techniques of lie detection. *Europe's Journal of Psychology*, 11:522–534, 2015.

Doratha Vinkemeier, Michel Valstar, and Jonathan Gratch. Predicting folds in poker using action unit detectors and decision trees. In *13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 504–511, 2018.

Aldert Vrij, Samantha A. Mann, Ronald P. Fisher, Sharon Leal, Rebecca Milne, and Ray Bull. Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, 32:253–265, 2008.

Jolita Vveinhardt, Daiva Majauskiene, and Dovile Valanciene. Does perceived stress and workplace bullying alter employees' moral decision-making? Gender-related differences. *Transformations in Business & Economics*, 19:323–342, 2020.

Heng Wang, Dan Oneata, Jakob Verbeek, and Cordelia Schmid. A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119:219–238, 2016.

Sharon Weinberger. Airport security: Intent to deceive? *Nature*, 465:412–415, 2010.

Robert S. Woodworth. *Experimental Psychology*. Holt, New York, 1938.

Zhe Wu, Bharat Singh, Larry S. Davis, and V. S. Subrahmanian. Deception detection in videos. In *The Thirty-Second AAAI Conference on Artificial Intelligence, AAAI-18*, pages 274–280, 2018.

Farid F. Youssef, Karine Dookeeram, Vasant Basdeo, Emmanuel Francis, Mekaeel Doman, Danielle Mamed, Stefan Maloo, Joel Degannes, Linda Dobo, Phatsimo Ditshotlo, and George Legall. Stress alters personal moral decision making. *Psychoneuroendocrinology*, 37:491–498, 2012.

Rongjun Yu. Stress potentiates decision biases: A stress induced deliberation-to-intuition (sidi) model. *Neurobiology of Stress*, 3:83–95, 2016.

Xiang Yu, Shaoting Zhang, Zhennan Yan, Fei Yang, Junzhou Huang, Norah E. Dunbar, Matthew L. Jensen, Judee K. Burgoon, and Dimitris N. Metaxas. Is interactional dissynchrony a clue to deception? Insights from automated analysis of nonverbal visual cues. *IEEE Transactions on Cybernetics*, 45:492–506, 2015.

Amir Zadeh, Yao Chong Lim Lim, Tadas Baltrušaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for 3d facial landmark detection. In *2017 IEEE International Conference on Computer Vision Workshops, ICCVW*, pages 2519–2528, 2017.

Lihi Zelnik-Manor and Michal Irani. Event-based analysis of video. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001*, volume 2, page 123, 2001.

Jane S. Zembaty. Aristotle on lying. *Journal of the History of Philosophy*, 31:7–29, 1993.

Lisong Zhang, Ming Kong, Zhongquan Li, Xia Zhao, and Liuping Gao. Chronic stress and moral decision-making: An exploration with the cni model. *Frontiers in Psychology*, 9:1702, 2018.

Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing*, 32:692–706, 2014.

Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andrew Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3438–3446, 2016.

Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. Verbal and nonverbal communication of deception. In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 14, pages 1–59. Academic Press, 1981.