Hiding in Plain Site: A Turing Test on Fake Persona Spotting

Grace McKenzie

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Lancaster University

Department of Psychology

March 2024

Abstract

This research investigated the ability of humans to accurately detect fake Facebook profiles. The prevalence of fake profiles on social media provides a consistent threat to users from malicious actors and computer-generated identities. This study wanted to move away from software and algorithm focused attempts to counter these threats and instead put the emphasis on the individual and their ability to detect a fake profile. Participants were shown a series of fake Facebook profiles (created specifically for this research) and real Facebook profiles and tasked with judging the authenticity of said profiles. Participants were also asked to identify the areas of the profile they had used to make their decision using heatmap software. Across the six studies within this research, new experimental manipulations were introduced each time in the form of time pressure, cross-cultural profiles, and training interventions. This approach allowed for investigation into the conditions that have the greatest influence over participants judgement accuracy, and also aligned the studies more closely with real world aspects of decision making in the online world. Participants were more accurate at correctly identifying real profiles as real than they were at correctly identifying fake profiles as fake, and fake profiles with a higher number of manipulated characteristics (*4 Fakes)* were judged more accurately than those with fewer (*2 Fakes)* or zero characteristics (*0 Fakes)*. However, their judgment accuracy was often no better than that of chance. Additionally, participants in all studies relied heavily upon the visual stimuli of the profiles (manipulated characteristic *Photo Type)* to inform their authenticity judgements. The insights gathered from this study add to the literature around online deception, establish a foundation for future research with a person-centric approach, and inform the design of future studies exploring similar themes.

Contents

Acknowledgements

Thank you to my supervisors Professor Paul Taylor and Professor Stacey Conchie for both their wisdom and guidance over the years. I’m so very grateful for your supervision.

I would also like to thank my family, friends, and work colleagues, old and new, who have offered me their unwavering support over the past four years. I appreciate each and every one of you.

Mr Pomerantz – your passion for both teaching and Psychology led me to develop my own passion for the subject from the age of 16. I don’t think either of us would have ever thought I would have achieved a PhD in Psychology after the disappointment on A-levels results day and the stress of getting a place on an Undergraduate degree. But here I am – I did it! (Exams really are not my forte).

Abdelbaqi – I never anticipated forming such a close friendship during this journey, especially while learning a new language in my limited spare time. You might be surprised (or not) to know that your lessons offered me a much-needed “brain-break” and something to eagerly anticipate each week. I am deeply thankful for your support, the Arabic lessons, and most of all, the laughter and your friendship. شكراً لك، صديقي العزيز

Mum – you are my inspiration, and I know I could never have accomplished this feat without you instilling the confidence in me that “I can do anything I set my mind to”.

K – there are truly no words to express how grateful I am for you and all your support. I adore you.

Declaration

I declare that this thesis is entirely my own work completed under the supervision of Professor Paul J Taylor and Professor Stacey M Conchie. No parts of this thesis have been submitted elsewhere in support of the application for the award of a higher degree.

Word count: 79,585

Signed:

Grace McKenzie

Date: 29.03.2024

Hiding in Plain Sight: A Turing Test on Fake Persona Spotting

“Oh, what a tangled web we weave when first we practice to deceive”
(Sir Walter Scott, 1808)

Deception and falsehood have always been rife within society. Deception is an inherent part of the human condition; a reason why for example, the Biblical creation story would not be complete without the deceiving of Eve by the snake in the Garden of Eden and her consequent deception of Adam. The nature of what deception is has been documented by philosophers such as Aristotle, Socrates, and Plato as far back as the Ancient Greek era. For example, Aristotle believed deception to be “mean and culpable” (Rubin, 2017), whereas Socrates coined the term ‘noble lie’, believing that lying is politically useful (The Mindless Philosopher, 2020). In early modern times, Niccolo Machiavelli believed deception is so commonplace as part of human nature that it gave rise to one of his more memorable maxims – “People are so simple, and so subject to present necessities, that anyone who seeks to deceive will always find someone who will allow himself to be deceived.” (Machiavelli, 1532, Chapter XVIII). More recently, deception had also been of interest to psychologists. Defined as “a successful or deliberate attempt to create in another a belief which the communicator considers to be untrue” (Vrij, 2000, p.15), there has been a particular emphasis on how deception can be accurately detected. Research has moved more away from the philosophical and into the scientific field.

Human deception detection

Pre-1980 there was limited literature on defining deception. There are now two main strands of research on deception detection: verbal and nonverbal behaviours (Zuckerman DePaulo & Rosenthal, 1981). Each strand has identified a variety of ways in which deception can be detected. Verbally, deception can be detected using linguistic analysis of speech content (Louth et al., 1998; Krackow, 2010; McKenzie, 2016), measurement of voice clarity (mumbling, pauses, stuttering etc.), and voice quality (pitch, speech rate, volume etc.) (Walters, 2000). Nonverbally, deception can be detected by analysing facial expressions and micro-expressions, body language, physiological reactions and illustrators (hand movements that reflect speech) and displacement activities (Ekman, 2009; Troisi, 2002). Certain traits of face-to-face deception like averting of the eyes have also been found to be strong cross-culturally (The Global Deception Research Team, 2006). Nevertheless, humans still struggle to have a high success rate in face-to-face deception detection (Vrij, 2008).

Recently, there has been a contextual shift in deception detection research as studies consider deception in the online space. Researchers are considering the types of deception which occur online and on social media (Tsikerdekis & Zeadally 2014; Kumar & Shah, 2018; van der Walt & Eloff, 2019). It is this online deception that is the focus of this thesis.

The Growth of the Online World

The consistent and continuous technological advances in our society have sparked what has been termed the 'digital revolution' (Aiken, 2016). Technology is everywhere. It is a pivotal part of daily life, and we as a human race have become dependent on technology. The widespread accessibility and portability of technological devices means that the internet is always within arm's reach. However, each new opportunity is a double-edged sword bringing with it the potential for a false or deceptive practice.

In recognition of this and as the internet continues to grow and become more accessible, some countries are developing online regulation laws in an attempt to protect internet users from illegal or disturbing content. The United Kingdom has introduced the Online Safety Bill into Parliament as of March 2022. The Bill aims to protect freedom of speech online whilst also reducing users' exposure to illegal and harmful content, specifically protecting children from pornography (U.K. Department for Digital, Culture, Media & Sport, 2022). The bill states that social media platforms will finally be held to account. The platforms will be required to uphold their terms and conditions and tackle exposure to harmful activity such as self-harming or exposure to eating disorders. The bill also explicitly states that the broadcasting regulator Ofcom (The Office of Communications) will have the ability to fine companies who do not comply with the new laws; block companies/websites who are non-compliant; and demand information regarding the algorithms used by the technology companies to understand how they use such algorithms to protect their users from potential harm.

The introduction of this bill marks a potential step toward the regulation of the internet within the UK but is limited in its reach. The scale of the problem that the online world can bring is recognised but it is not yet contained in a practical or workable sense. The burden of safety still lies upon the user. In the pre-internet world Mark Twain (1907) said "there are three kinds of lie: Lies, damned lies and statistics". One might wonder quite how many he would categorise now, with scam emails from Nigerian princes looking to extort money, or what definition would be applied to Snapchat filters allowing users to present themselves as anime characters. The online world has been a metaphorical Pandora's box in opening up a range of deception possibilities from the bizarre to the virtually impossible to detect. The scale of this new and dynamic online realm is also vast.

Social Media and its Popularity

As of January 2020, the majority of internet traffic is found on social media sites; of 4.54 billion internet users, 3.80 billion are active social media users (Chaffey, 2020), whether that be for work purposes, advertising, news or socialising, across a wealth of different platforms. As of January 2024, there are now 5.35 billion internet users, and 5.04 billion (62.3% of the world's population) are active social media users (Statista, 2024a). In the space of four years that is a growth of 32.6% in social media users.

The development of social media has provided a wider forum for deceivers to operate in and is particularly rife for identity play – lax age restrictions actually actively encourage the creation of a fake profile in order to access the platforms (Aiken, 2016). Most of the activity on our social media profiles is harmless and irrelevant, no more deceitful than we may exhibit as we make our way through a normal day. However, in keeping with the online pattern, behind each opportunity can lie a risk. The internet, and particularly social media, has provided an easily accessible platform for someone to deceive through the guise of anonymity. There can be minimal, if any, security checks or identity validation of the persons who use social media platforms, or create websites and online businesses, or create videos that distribute misinformation (Aiken, 2016). The capacity of companies to monitor newly created accounts is difficult when one considers the top 3 most popular social media companies are averaging over 8 billion monthly active users (Statista 2024b). This creates a context where deception can go unchecked and a prime example of this on social media platforms is the spread of fake profiles.

Fake profiles

Fake identity and fake persona are terms that tend to be used interchangeably within the literature and can be defined simply as 'a person pretending to be what they are not' (Morton, 2016). A fake profile is one which follows the terms and conditions of use via the relevant social media channel, without spreading content outside of those terms or seeking to acquire data through the use of software robots or *bots* (Wan, Jabin, Yazdani & Ahmadd, 2018). Specifically, Facebook says multiple accounts, or 'inauthentic behaviour' about the identity, purpose or origin of a profile is a breach of their community rules (Meta, 2024a). Worryingly however, there is no limit to the number of fake identities/personas a person can adopt online with little oversight. As noted by Donath (1995), since the time of the conception and early introduction of the internet, a person can have as many virtual identities as they have time and energy to create. In a sense nearly each and every online profile on Earth is deceptive, for whom amongst us can truly say we use them openly and unambiguously? Whether it be flattering pictures, holiday photos, use of a filter, the online profiles of life that social media platforms present to the world is sometimes barely recognisable as the mundanity of the real person's life and yet none of this behaviour may even be classed as fake. Even choosing not to post about something means that a form of deception has been carried out, for it can be said that a lie by omission is a lie, nonetheless.

Scale of problem of Fake Profiles

The scale of the proliferation of fake profiles is, by its nature, hard to quantify. Elon Musk's massive $44 billion Twitter takeover was nearly scuppered over legal disputes regarding the number of fake accounts on the site (Conger, 2022), and there are now reports that a large proportion of Musk's followers may themselves be non-active or fake accounts (The Economic Times of India, 2023). In 2018, Twitter reported the removal of 70 million fake accounts (Timberg & Dwoskin, 2018), and in 2019 Facebook removed 2.2 billion fake accounts (their highest number of fake accounts to date) in the first quarter of the year alone (Statista, 2024b). The staggering height of these numbers demonstrates there is an evident problem with fake profiles on social media.

The air of suspicion which lingers around the online space means people have an expectation of deception online (Caspi & Gorski, 2006) and this atmosphere of mistrust generates others to themselves engage in deceptive practices which they might deem common (Drouin, Miller, Wehle & Hernandez, 2016). Entertainment shows based solely on the fake profile phenomena now exist, like *Catfish: The TV Show* with over 230 episodes up to January 2024 (Jarecki et al., 2012 – present) , and a BBC production *For Love or Money* (Short, 2019 - 2022) examining the same theme. Similarly, the Netflix documentary *Untold: The Girlfriend Who Didn't Exist* (Vainuku & Duffy, 2022) examined a high-profile incident of catfishing sustained through a fake internet persona. However big the true problem, whilst any meaningful amount of fake profiles are present, it poses a risk to users, and the continued development of television series and documentaries on the topic again highlight the magnitude of the issue. How are people *still* getting fooled or duped by fake profiles?

Whilst platforms could potentially seek to manage human made fake profiles on a case-by-case basis, the sheer number of them means that they are unable to react accordingly in a timely manner. Studies have shown that 80% of users would accept Facebook friend requests from strangers if they met certain criteria, like mutual friends (Boshmaf, Muslukhov, Beznosov & Ripeanu, 2011). When users lack the judgement to protect themselves and thereby exacerbate the problem, the social media providers will never be able to keep up with the problem in order to tackle it.

Bots

Online deception is not only restricted to humans manipulating profiles, it also concerns non-human accounts commonly known as *bots* (Ferrara, Wang, Varol, Menczer & Flammini, 2016). Bots are typically created on a large scale to perform malicious activities such as rumour spreading, spamming, phishing, manipulating opinions, manipulating algorithms, and fabricating accounts (Aljabri et al., 2023; van der Walt, 2018). There are multiple types of bots; chatbots, shopbots, knowbots, spiders/crawlers, monitoring bots, transactional bots, spambots, socialbots etc. (Lutkevich & Gillis, 2022).

Socialbots, also referred to as *social media bots*, are defined as social media accounts that are controlled by a computer program rather than a human and can be either benign or malicious (Mbona & Eloff, 2023). Bots are now so prevalent that they are already their own subject of classification and categorisation – into types like broadcast, consumer and spam bots – and are the focus of research dedicated to this (Liu, Zhan, Jin, Wang & Zhang, 2023). As a whole, socialbots pose a large threat to social media credibility and currently only a very low percentage of people believe they could accurately identify them (Stocking & Sumida, 2018). Previous research has suggested that people struggle to detect profiles run by bots, particularly if these profiles align themselves with a viewpoint already held by the person making the judgement (Kenny, Fischhoff, Davis, Carley & Canfield, 2024). As social bots can create and run profiles so quickly, often without any human interaction, the ability of networking sites to respond to them becomes limited. Furthermore, as social bots more and more closely mimic human behaviour, the gained computer advantages which could detect them are therefore subsequently eroded. (Ferrara, Varol, Davis, Menczer & Flammini, 2016). Despite all our technological advances therefore, we are still often reliant upon human judgement to determine risk.

Artificial Intelligence (AI) has grown exponentially and continues to do so on a daily basis. In 1950, Alan Turing proposed the question of whether machines can think. The Turing test is a test of a machine's ability to demonstrate intelligence of that of a human brain. The probability of the computer being mistaken for a human is the measure for a computers' ability to think (Turing, 1950). Throughout the early to mid-1960's, computer scientist Joseph Weizenbaum created what is considered to be the first chatbot named 'ELIZA', a program which made conversations between human and computer possible (Weizenbaum, 1966), and thus one of the first documented attempts of creating a machine to act like a human. However, general consensus amongst computer scientists at the beginning of this study is that there is yet a system or machine that has passed the Turing test, despite multiple attempts. Most recently, the Google Duplex machine scheduled several appointments over the phone, interacting with humans at the respective businesses or companies, and each time the humans were unaware the caller was a computer (Leviathan & Matias, 2018). However, controversy surrounds this attempt, as the calls between the human and computer are argued to have been doctored in that no personal contact details were shared, and there was a significant absence of any background noise (Natale, 2021). These programs, along with virtual assistants on mobile phones or speakers (Amazon Alexa, Google Home, Siri, Cortana), are all examples of chatbots – programs designed to simulate a conversation with a human (Lutkevich et al., 2022). The recent online launch of the web game *Human or Not? (*Jannai, Meron, Lenz, Levine & Shoham, 2023) where users try to work out whether they are interacting with man or machine, demonstrates how accessible and popular Turing tests are, no longer the prerogative of science and computing alone. As we move into 2024 ChatGPT (OpenAI, 2022) has averaged over 1.5 billion visits per month for the last 6 months up to January 2024, as opposed to 152 million in November 2022 (Similarweb, n.d.). As with other online developments, AI brings with it new opportunities and new threats.

Computer Science Efforts to Tackle Problem

The majority of the current literature on fake profiles/identities, and the detection of such, is heavily based in computer science research - social media platforms have their own methods of detecting fake profiles, which often revolve around complex machine learning algorithms engineered for their respective platform. Algorithms, or more simply a set of mathematical rules that will help a computer calculate an answer (Merriam-Webster, n.d.), have been designed in an effort to enable a computer to automatically detect the prevalence of fake profiles online. Common machine learning methods, as evidenced in the computer science literature, include support vector machines (SVM), Bayesian classifiers, and neural networks.

SVMs are supervised learning algorithms used to solve classification problems and regression tasks and are particularly useful when analysing complex data that cannot be separated by a straight line (Fred Tabsharani, n.d.). Bayesian classifiers are a family of classification algorithms based on Bayes' Theorem (Bayes, 1763), a theory used to determine the conditional probability of an event based on prior knowledge of a previous related event, and aid in the rapid development of machine learning models (Rohith Gandhi, 2018). Bayesian classifiers are often used in filtering spam and have been successfully applied to detecting false information on social media with reasonable accuracy (Alowibdi et al., 2014). Neural networks are a complex series of algorithms designed to recognise underlying relationships in data by mimicking the operational functions of the neurons in the human brain (James Chen, 2024).

Understandably, with the exponential growth of social media, and the consequent increase in terms of soft power and wealth it brings a large amount of research that has been conducted on how to detect socialbot accounts, with researchers finding that machine learning models have consistently high accuracy in detecting bots.

Kudugunta & Ferrara (2018) developed a deep neural network model to ascertain if the possibility of accurately predicting if a tweet was posted by a real human account or a 'bot' account. Using a dataset of over 8,000 tweets, Kudugunta & Ferrara (2018) reported their model had achieved 96% accuracy in separating bot accounts from human accounts. Similarly, Chavoshi, Hamooni and Mueen (2016) reported 94% accuracy by using a real-time cross correlation of bot activities method; a self-reported 'new' approach to bot detection, and Adikari and Dutta (2014) achieved 84% accuracy by using neural networks and SVMs to identify fake profiles on LinkedIn. Yang, Harkreader and Gu (2011) analysed spam accounts on Twitter using a graph-based method and achieved 86% detection accuracy, and Lee, Caverlee and Webb (2010) achieved 88.98% detection accuracy of Twitter spam accounts using a meta-classifier. Researchers Chu, Gianvecchio, Wang and Jajodia (2010) achieved high success rate of detection on Twitter with a text classifier program, and Gupta and Kaushal (2017) achieved a 79% detection accuracy rate of fake accounts on Facebook using supervised machine learning classification techniques. Such techniques work by analysing different aspects of the profile (Elyusufi, Elyusufi & Kbir, 2020), the behaviour of the profile i.e., number of comments made by the profile per day and number of rejected friend requests (Ajith & Nirmala, 2022), and the links between multiple profiles – the social network (Kagan, Elovichi & Fire, 2018). However, the general consensus within the literature is that overall, computer science methods are not 100% effective at either detecting or removing fake accounts.

Popular platforms, such as Twitter and Facebook, have reported their tailored algorithms *cannot* detect and remove all fake profiles. Facebook, the most used platform on a global scale with 3.04 billion users (Statista, 2024c), recently reported that in the fourth quarter of 2023, they had removed 691 million fake profiles and accounts from their platform, mainly by using machine learning detection (Meta, 2024b; Statista, 2024b). They also reported use of employees who are specifically tasked with implementing the decision to remove the content that has been flagged as fake or malicious by Facebook users. However, the manual removal of such profiles/accounts by Meta employees is not only a much slower process than their tailored algorithms, but it only contributes to a very minor percentage of removals overall, due to the sheer magnitude of the fake profiles issued and relies upon users reporting accounts for such accounts to be investigated and removed. Additionally, Meta did not specify the individual checks employees use when reviewing the flagged accounts, so it is unknown whether all flagged accounts are automatically removed or whether employees do further investigative checks into the authenticity or behaviour of these accounts before removal, and if so, what employees look for to signify a fake account. Despite these large and albeit impressive numbers, within the same report Meta also stated that they expect that the number of Facebook accounts that they can detect will vary over time as the creation of fake accounts is unpredictable in nature, alluding to the fact that the creation of fake profiles is ever morphing as users continue to find new ways to avoid detection. This raises the question of whether perhaps some users will always fall through the metaphorical net?

Another major issue found universally across social media platforms is that, as the researchers Romanov, Semenov, Mazhelis and Veijalainen (2017) reported, each of the platforms do not meticulously check that the identity shown on the profile is an accurate representation of the person's identity in real life, i.e., that the user is who they say they are. This in itself is a massive problem in terms of identity authenticity on social media. Herein lies the problem – the social media platforms are actively trying to combat the ever-growing issue with fake profiles but are yet to develop a method that can wholly protect their users from said profiles.

Human Efforts to Tackle Problem

In comparison to human deception detection rates, computer aided deception detection has a much higher accuracy rate; many researchers have reported that human deception detection is not much higher than chance, with most finding a similar rate to the well-known Bond and DePaulo (2006) meta-analysis average rate of 54% human accuracy. Human deception detection online has had nowhere near the wealth of research as computer deception detection has - there has been research which has skirted around the edges of human detection methods (Kenney et al., 2022).

Studies that have already been conducted in the online sphere have measured some aspects of our personality traits and how judgement accuracy of others personality types can be measured (Darbyshire, Kirk, Wall & Kaye, 2016) and considered computer-based personality judgements based on social media and digital footprints (Hinds & Joinson, 2019). However, there have been very few studies that directly look at human ability at detecting deception online. Kenney et al (2022) highlighted this issue and in an effort to tackle said issue developed a study on human ability to detect social bots on Twitter, finding that humans had limited ability to detect social bots and were more likely to mistake a bot for a human and vice versa.

Further related research on human detection ability in the online sphere focuses on misinformation and is especially prevalent in political and health fields. Such research seeks to increase user judgement of fake news in an attempt to prevent the dissemination of it (Walter & Murphy, 2018; Suarez-Lledo & Alvarez-Galvez, 2021). However, there is still a continued focus towards this judgement being incorporated into the design of the social media platforms, meaning that it is still the provider, rather than the user, that is putting steps in place to detect deception. The end user is wholly reliant on the provider to continue that service, something which is ultimately not necessarily in their interest (Aiken, 2016).

Studies focusing solely on human detection of fake social media profiles are very few and far between - more attention now has to be paid towards the role of the individual in determining judgements of fake or deceptive information in the online sphere. A lie has always been able to travel farther than a truth and in the digital world, it now reaches further as well. Attempts to stop deceptive practices are currently very cumbersome and unresponsive (Shao, Ciampaglia, Flammini & Menczer, 2016). This thesis looks to strip back some of the work on deception in the online sphere and distil it back to a more original form, where emphasis is placed on the individual's ability to make an accurate judgement on deception and apply it themselves rather than rely on computer science technology or the online service providers to instigate checks and balances.

The chapters that follow seek to place the emphasis for individuals' safety from fake profiles wholly back into the hands of the user. If it can be established that individuals either possess, or can be trained to possess, the judgement skills required to make an accurate determination of whether a profile is fake, this not only removes their reliance upon the social media platforms to protect them, but it also means that any detection is instantaneous rather than responsive. Each study measures human deception detection accuracy in the form of a judgement task, specifically judgements of fake and real Facebook profiles. Study 1 (Chapter 2), Study 2 (Chapter 3), and Study 3 (Chapter 4) explore the different characteristics of a Facebook profile that users rely upon to make judgements as to the authenticity of the profiles and the accuracy of said judgements. Studies 4, 5, and 6 (Chapters 5-7) further measure human deception detection accuracy of Facebook profiles with added manipulations, namely judgement time pressures, cross-cultural profiles, and training interventions respectively.

In an early anthropological work, it was claimed that “Just as natural selection inevitably produces would be cheaters, it will inevitably give rise to individuals capable of detecting cheating” (Leakey & Lewin, 1978, p. 192). This thesis hopes to set in motion a body of work which can help that ideal become a reality.

Chapter 2: Study 1

Introduction

A growing body of literature seeks to understand the detection of fake profiles on social media platforms, specifically how accurate both computers and humans are at identifying fake profiles. Within this body of work is a distinct absence of knowledge on the motivation and intentions of fake profiles and how these may shape the form that fake profiles take, the cues that people use when deciding if a profile is fake, and the individual differences that can moderate the ability of a person to detect a fake profile. Furthermore, the current literature fails to investigate the accuracy of detecting fake human profiles rather than the fake accounts created by computers, accounts more commonly known under the umbrella term bots, and there is little to no explanation as to what characteristics found on a social media profile denote a fake profile, i.e., the number of friends, the number of photos etc. It is of great importance in the social media sphere to identify these characteristics and understand the ways in which they are used, or rather manipulated, to create fake profiles. Doing so means effective strategies can be developed to detect and combat such profiles, of which can be disseminated to inform social media users, in turn providing said users with an enhanced layer of security and personal protection in regard to their user experience.

Understanding human decision-making on whether social media profiles are real or fake is important for several reasons. Firstly, knowing how human users decipher such information can help to inform the platforms of what is typically manipulated on a profile to make it fake and therefore assist them in warning users that a profile may be fake. Secondly, this understanding could potentially help improve machine learning techniques – if the algorithms understand human decision-making strategies, they could incorporate these into their current techniques to make them more effective and in turn improve their efficacy. Finally, it can provide further insight into deception detection techniques, particularly those in the online domain.

Research on social perception and realistic accuracy judgements (Funder, 1999), can provide an insight into how humans detect or 'judge' a fake identity in the form of a fake social media profile. Funder (1995) introduced the Realistic Accuracy Model (RAM) which details the process of making accurate judgements of a person's personality using cues in the environment. The model posits a four-stage process to achieve an accurate judgement: *Relevance* (the quality of the information/cue from the target/observed person); *Availability* (the cue becomes available to the judge when it can be detected by said judge); *Detection* (judge is aware of the *relevant* and *available* cues from the target); *Utilisation* (judge correctly uses the cues available to make an accurate judgement). Funder (1995) outlines that to make an accurate judgement all four stages of the judgement process must be completed successfully in the correct order for accuracy to be achieved, and in later works reported that the detection of cues is reliant on there being a large number of information available (Funder, 1999). Although this model was created to explain the process of personality judgements, the underlying premise is applicable here, where the cues referred to are different areas of the social media profiles that might be used to inform judgements. However, there is a distinct lack of research into what these specific profile cues may be and how they are used to inform authenticity judgements. The only available literature is from websites, forums, and blogs of people's opinions on what makes a fake profile – there is yet to be any scientific research investigating the specific areas of a profile that denote a profile to be fake, hence the need for this research.

This research is designed to aid in bridging the gap in the online deception and fake identity literature to assist in a better overall understanding of the constitution of fake profiles online, how human users make authenticity judgements of said profiles, and the ramifications associated with those profiles that go undetected. Based on the limited literature surrounding this topic, a few hypotheses have been developed.

Of the research that has been conducted into human judgement of real and fake stimuli, researchers Köbis, Doležalová and Soraperra (2021) found that when participants were required to detect deepfakes, they could not reliably identify a deepfake, and they were biased towards believing the deepfake was real. As such, it is expected that this study will find a difference between the accuracy of profile judgments of real profiles and fake profiles (Hypothesis 1). Further, based on research on available information, it has been shown that people inherently believe that having more information helps to make an accurate judgement of a persons' personality (Beer, 2019), as such it is expected that accuracy will improve as the number of fake characteristics increases. Specifically, fake profiles that are constituted of four fake characteristics will be judged more accurately as fake as they contain a higher number of cues to the authenticity of the profile (Hypothesis 2).

In regard to the human behind the judgement, researchers have studied whether certain personality types have an effect on decision-making and judgement processes, finding that personality types do influence credibility judgements (Ahmad, Wang, Hercegfi & Komlodi, 2011), and that certain types of personality can make sub-par decisions when under pressure (Byrne, Silasi-Mansat & Worthy, 2014). As such it is expected that there will be a relationship between participants' personality and judgement accuracy (Hypothesis 3). As an extension of this, it is also expected that scores on the social sensitivity scale, a scale that measures interpersonal sensitivity - the ability to accurately assess the traits and states of another from non-verbal cues (Carney & Harrigan, 2003), will be related to judgement accuracy scores (Hypothesis 4). Social sensitivity was included as a measure of interest due to the key role it plays in social cognition, particularly the mental processes involved in perceiving and understanding others (Zaki & Ochsner, 2011). In specific relation to human judgement, social sensitivity can directly influence a person's ability to make accurate judgements of others (Hall & Andrzejewski, 2008), and including such a measure allows the researcher to account for any individual differences that may influence judgement accuracy.

With a focus on social media experience and usage, researchers have found that time spent on social media is positively correlated with the ability to detect deepfakes (Nas & de Kleijn, 2024). Although this is in reference to deepfakes, it is expected that a similar effect will be found with fake profiles – those who self-report that they frequently use social media are expected to perform better when judging profiles (Hypothesis 5). Along the same thread, whilst there is no available literature regarding previous experience of creating a fake profile and the relationship this might have with the ability to detect fake profiles, Kenny et al. (2022) have found that participants with greater experience on social media are less susceptible to being duped by a fake profile created by a bot. Based on these findings and the fact that social media usage correlates with detection ability, it is also expected that there will be a relationship between those who have previous experience of creating a fake profile and their judgement accuracy (Hypothesis 6). Having prior understanding of the areas that can be manipulated to create a fake profile may help in detection of fake profiles.

Some researchers have found that people display a level of over-confidence in their detection ability of fake or manipulated stimuli. Köbis et al., (2021) found participants overestimated their ability to detect deepfakes, and when identifying manipulated photographs people exhibit confidence that their judgements of the authenticity of the photographs, and therefore identifications of real and fake images, are correct (Nightingale et al., 2022). As such, it is expected that there will be a relationship between self-reported accuracy and actual accuracy of the profile judgements (Hypothesis 7).

Finally, the profile characteristics manipulated within this study are expected to have an effect. Previous research has found that particular stimuli – photos and imagery – are associated with influencing personality judgements of the photo subject (Turner & Hunt, 2014). However, there is no research regarding any other areas of profiles and how these may have an effect on authenticity judgements. As such, the hypotheses in regard to the manipulated characteristics will be non-directional. It is expected that the manipulated characteristics will have an effect on participants' judgements, i.e., whether that be a judgement of real or fake (Hypothesis 8), and the accuracy of their judgements i.e., whether their judgement is correct or incorrect (Hypothesis 9).

Due to the size, popularity, and global use of Facebook, this platform was chosen as the platform of choice for this research. This study sought to examine the characteristics of a Facebook profile (i.e., number of photos, number of friends etc.) that are used when making judgements as to whether the profile is fake (deceptive), i.e., what denotes a fake profile. The purpose of doing so is to create a succinct evidence-based list of fake profile characteristics that can be disseminated to society as a list of characteristics/cues, or 'red flags', to be aware of when looking at social media profiles. Further investigations will involve analysing the accuracy of the judgements in an effort to answer the age-old question of whether humans are better than chance at detecting deception, and expanding on this by tapping into a lesser researched aspect of deception detection; 'are humans better than chance when the deception is *online?*'

This study, along with subsequent studies in the following chapters, employed an experimental Turing test design. As discussed in the literature in Chapter 1, the Turing test serves as a benchmark for evaluating a machine's capacity to exhibit human-like thinking and behaviour, whereby the probability of a computer being mistaken for a human is the measure of a computers' ability to think (Turing, 1950). In the context of this research, the Turing test was adapted to evaluate the ability of digital entities, namely the Facebook profiles, to mimic human behaviour online. Essentially the 'machine' in the original Turing test is represented here by the Facebook profiles, and the method of a human judging the outputs (profiles) of a machine (social media), and the authenticity of the outputs, mirrors the method of the original Turing test.

Method

Participants

G*Power analysis (Faul, Erdfelder, Lang & Buchner, 2007), using an A-Priori power analysis of a repeated measures, within subjects ANOVA, was conducted prior to data collection to determine the appropriate sample size for this study. The analysis indicated that a sample size of 28 participants would be sufficient to detect a medium effect size of $f = 0.25$, with an alpha level of $\alpha = 0.05$ and a power of $1-\beta = 0.80$. However, 200 participants were recruited. The reasoning behind this decision is threefold. Firstly, increasing the sample size beyond the minimum needed enhances the reliability and generalisability of the findings by reducing the errors associated with the estimates. Secondly, the representativeness is improved as a sample of 200 participants is more likely to include a more diverse variety of the population studied, and thirdly increasing the suggested sample size from 28 to 200 reduces the risk of Type 1 and Type 2 errors thus reducing the likelihood of drawing erroneous inferences from the data.

Two hundred and four participants completed the study. Participants were recruited via Prolific.co through volunteer sampling. Data from three participants were removed because of an error in the Prolific ID, which meant that participants' entries would be unidentifiable if they wished to withdraw from the research. As this would violate the ethics of the study these participants were removed from final analyses.

Of the 201 participants included in the final analysis, 132 (65.7%) identified themselves as Male, 67 (33.3%) identified themselves as Female, 1 (0.5%) identified themselves as Non-binary, and 1 (0.5%) identified themselves as Transgender. Participants were aged between 18 – 63 years (*M* = 26.60; *SD* = 8.57), identified ethnicities as Asian or Asian British (*N* =11, 5.5%), Black, African, Black British or Caribbean (*N* = 8, 4%), Mixed or Multiple Ethnic Groups (*N* = 4, 2%), White, including any white backgrounds (*N* = 170, 84.6%), Another Ethnic Group (*N* = 4, 2%), and prefer not to say (*N* = 4, 2%). The location of participants spanned across six continents: Africa. Asia, Australasia, Europe, North America, and South America, with the three most popular locations being Poland (*N* = 49, 24.4%), United Kingdom (*N* = 36, 17.9%), and Portugal (*N* = 25, 12.4%). Of the 201 participants, five did not enter information regarding their location.

Design

The study used a 3 (Profile Type; Real profiles, 2 Fakes profiles, and 4 Fakes profiles) x 2 (Accuracy; Accurate judgement vs. Non-accurate judgement) experimental Turing test design. Both of the two factors are within subjects' measures, following a repeated measures design. The Independent Variable (IV) is the type of Facebook profiles presented to the participants, with three levels: real profiles, 2 fake profiles, and 4 fake profiles. The concept of accuracy is measured in two ways: *Subjective Accuracy* (treated as an IV) and *Objective Accuracy* (treated as a Dependent Variable [DV]).

Subjective Accuracy reflects the participants' perceptions of how accurately they judged the authenticity of the profiles (self-rated confidence of the accuracy of their profile judgements), and *Objective Accuracy* is a continuous variable that reflects actual accuracy/correctness of the participants' judgements – the judgement of Fake or Real.

Measures and Materials

Measures

Social Media Use. Participants' use of social media was measured with five questions asking if they used social media, how much time they spend on social media, and the reasons that they use social media. The two remaining questions asked participants to select and rank the social media platforms they use from used most often to used least often (Appendix A).

Personality. Participants' personality was measured on the 10-item TIPI (Gosling, Rentfrow & Swann, 2003), which captures the Big-5 personality traits – Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism. These traits are measured on a 7-point Likert scale ranging from 'Disagree Strongly' (1) to 'Agree Strongly' (7) (Appendix B). To score this scale, items 2, 4, 6, 8, and 10 were first reverse scored, based on the inventory manual (Gosling et al., 2003), so a 'Disagree Strongly' with an original score of 1 would become a 7. A score for each of the five traits was calculated by summing the two specific items for each trait (e.g. Extraversion = Item 1 + Item 6), using the newly calculated reverse scored items. Each score for the five traits was then averaged (divided by 2), and these final scores were summed together to achieve an overall score across all 10 items on the scale.

Social Sensitivity. Social sensitivity was measured using the shortened 15-item social sensitivity (SS) scale (Riggio & Carney, 2003). The SS measures Interpersonal Sensitivity by assessing respondents' ability to interpret and decode social situations and the verbal communication of others. The scale is measured on a 5-point scale; 1: 'Not at all Like me', 2: 'A little like me', 3: 'Like me', 4: 'Very much like me', 5: 'Exactly like me'. Each of the 15 items are scored by summing the number value of the responses selected, i.e., 'Like me' has a score of 3, and 'Very like me' has a score of 4 etc. Questions 1, 3, and 7 on the scale were reverse scored in line with the Social Skills Inventory Manual (Riggio & Carney, 2003). (See Appendix C for the scale).

Accuracy judgements. *Objective Accuracy:* After each profile, participants were asked to judge whether they believed the profile to be real or fake. *Subjective Accuracy:* Once all stimuli had been presented participants were then asked at the end of the study to rate on a 7-point scale how confident they were of their judgements overall , from 'Unconfident' (1) to 'Confident' (7).

Fake profiles. Participants were asked to indicate if they have ever created a fake social media profile using a binary yes/ no response (Appendix D). If participants answered 'yes', they were asked to detail the circumstances.

Facebook Profiles

A total of 36 Facebook profiles were used; 30 fake profiles and six real profiles. The 30 fake profiles were created for this research, based upon research and statistics gathered from the internet (Salman Aslam, 2020), and related personality judgement research (Krombholz, Merkl & Weippl, 2012; Darbyshire, Kirk, Wall, & Kaye, 2016) as to what characteristics are said to be most likely to represent a fake profile. The characteristics manipulated within this study and the breakdown of them (as most typically identified as being used in fake profiles) were; 'Age' (under 17 or over 80), 'Photo – Number' (very few or thousands of photos), 'Photo – Type' (non-descript photos e.g., landscapes, artwork, cartoons, celebrities etc.), 'Friends – Number' (very few friends e.g., under 100, or a lot of friends e.g., over 200), 'Posts – Number and

Regularity' (very few and non-regular posts on the timeline), and 'Groups' (lack of subscriptions to groups or subscription to hundreds of groups) (See Appendix E for the lists of manipulated characteristics). Fifteen of the fake profiles comprised two of the aforementioned fake characteristics and the remaining 15 fake profiles comprised four of these fake characteristics, the purpose of such being to analyse whether the fake profiles with more fake characteristics are easier to identify and therefore more accurately judged by participants as fake than the profiles with fewer characteristics.

The six participants who provided their real Facebook profile were recruited via word of mouth and were all known to the researcher for the purposes of validating that their profiles were 'real' and a true and accurate representation of their identity. Demographic information was not recorded via means of a questionnaire, however as they are known to the researcher, it can be reported that their ages ranged from 27-59 years (*M* = 40.67, *SD* = 14.51), three of these participants were Male (50%) and three were Female (50%), with all six identifying as White and being located within the United Kingdom.

Both the real and fake profiles were formed by taking a series of screenshots of the main page of the Facebook profile (previously known as 'the wall'), then cropping/editing these and composing together using Adobe Photoshop software, to form one static screenshot of the profile that contains all relevant profile information in one accessible place (Appendix F).

Participants who provided their real Facebook profile for the purposes of this research were invited to participate via e-mail. Each of the six participants were sent an initial consent form outlining the study and details regarding the need for the use of their profile and how their profile will be used (Appendix G). Participants were also provided with information regarding how their data will be used and stored and of their right to withdraw from the study. After providing initial consent, participants were e-mailed a set of instructions detailing how to screenshot their Facebook profile and were asked to e-mail the screenshots back to the researcher to allow for the researcher to Photoshop the screenshots together. Using Photoshop, the researcher 'stitched' all the screenshots together to emulate a 'real' Facebook profile and omitted all the identifiable data on the profile to ensure the privacy of the participant is protected. These omissions included the name of the participant as it appeared in any form (i.e., on posts, comments, tags etc.), the names of others in any form (i.e., friends list, comments, tags etc.), the names of any other social media profiles or email addresses that may be shown on the profile, and the profile pictures of anyone who had commented on or tagged anything on the participants' profile timeline. Participants were sent the final screenshot image of their Facebook profile for their approval, along with a further consent form (Appendix H). Participants were given the option to provide final consent or withdraw from the study at that point. If participants chose to provide their final consent, they were informed they had one week after providing their consent to withdraw their profile from use within the study. After this duration they would be unable to withdraw due to their profile becoming inextricable once the study was live on Prolific and Qualtrics.

In preparation for analysis, each participant was given an accuracy score for their profile judgements based on how many profiles they accurately judged as real or fake. The accuracy score was split across the three different types of profiles: *2 Fakes* accuracy (maximum score of 3), *4 Fakes* accuracy (maximum score of 3), and *Real* accuracy (maximum score of 6). Additionally, an *overall accuracy* score for each participant was calculated across all types of profiles, giving a maximum score of 12.

Procedure

A link to the study was published on Prolific where the study was described as 'an investigation into the detection of fake profiles on social media'. If participants wished to take part in the study, they clicked on the link provided which automatically redirected them to the survey on Qualtrics. Once on Qualtrics, participants were presented with information about the study and asked to indicate their consent to take part. Those who clicked 'no' were thanked for their time and instructed to close the window. Those who clicked 'yes' were presented with the first page of the study whereby they were required to input their unique Prolific ID number, as provided to them by Prolific when they signed-up to participate, to allow for their data to be identified and withdrawn if later requested.

The first stage of the study required participants to complete the Social Media questions, the TIPI, and SS scales. They were then presented with 12 profiles; six real profiles and six fake profiles; all of which were randomised. Participants were instructed to view each profile carefully and make a judgement as to whether they thought each profile was real or fake. Participants were asked to indicate the specific areas of the profile they used to reach their decision by clicking on areas of the profile, which was later converted into heatmaps. The heatmap function allows for graphical analysis of the most clicked on areas using a colour scale - the clicks on the profiles by participants present as coloured circular areas where blue areas are at the lower, or 'cooler', end of the scale and represent few clicks, and red areas are at the higher, or 'warmer', end of the scale and represent a larger number of clicks. To capture the clicks, regions were outlined on each profile surrounding each of the manipulated characteristics and were made as wide as was possible without interfering with other regions, to ensure slightly inaccurate clicks were encapsulated within the regions. For example, if a participant was relying on the Number of Comments and didn't click directly on the number, but just to the side of it, this would still be included within the Number of Comments region (within reason). These regions were visible only to the researcher (Appendix K). All characteristics – including those manipulated in the profiles – were clickable. Once participants clicked the image as instructed, a red dot appeared in the place that they clicked and remained in place (Appendix L). Participants were informed that they were allowed up to 10 clicks on the image and to be as accurate as possible when selecting each area of the profile to allow for accurate data analysis of the regions of the 'heatmap'.

Following each profile were two short multiple-choice questions, asking if participants needed more information to make their judgement and if they believed the profile was created with malicious intent and deceptive motivations. If participants selected 'yes', they were asked to detail why in a box situated below their answer (Appendix M).

After completing the 12 profile judgements, participants were asked two final questions – how confident they were in their judgements, and whether they had experience making a fake profile. They were then asked to provide demographic details on age, gender, ethnicity, and country where they live (Appendix N).

Once participants had completed the study they were thanked for their time and participation and fully debriefed, whereby they were provided with further information on the purpose of the study and given contact details of the researcher and supervisors should they have any issues regarding their participation, or further questions regarding the research or the results (Appendix O).

Ethics

This research was fully approved by the ethics committee at Lancaster University on 11th May 2020. All participants were provided with the appropriate information to give their informed consent, including information on voluntary participation, their right to withdraw and the use and storage of their data. Additionally, anonymity of the participants was adhered to through the omission of all identifiable details on the real Facebook profiles and by use of the individual Prolific ID numbers for those participating in the online study. All data was only accessible to the researchers and stored on a secure hard drive and university approved applications, in line with GDPR guidelines.

Results

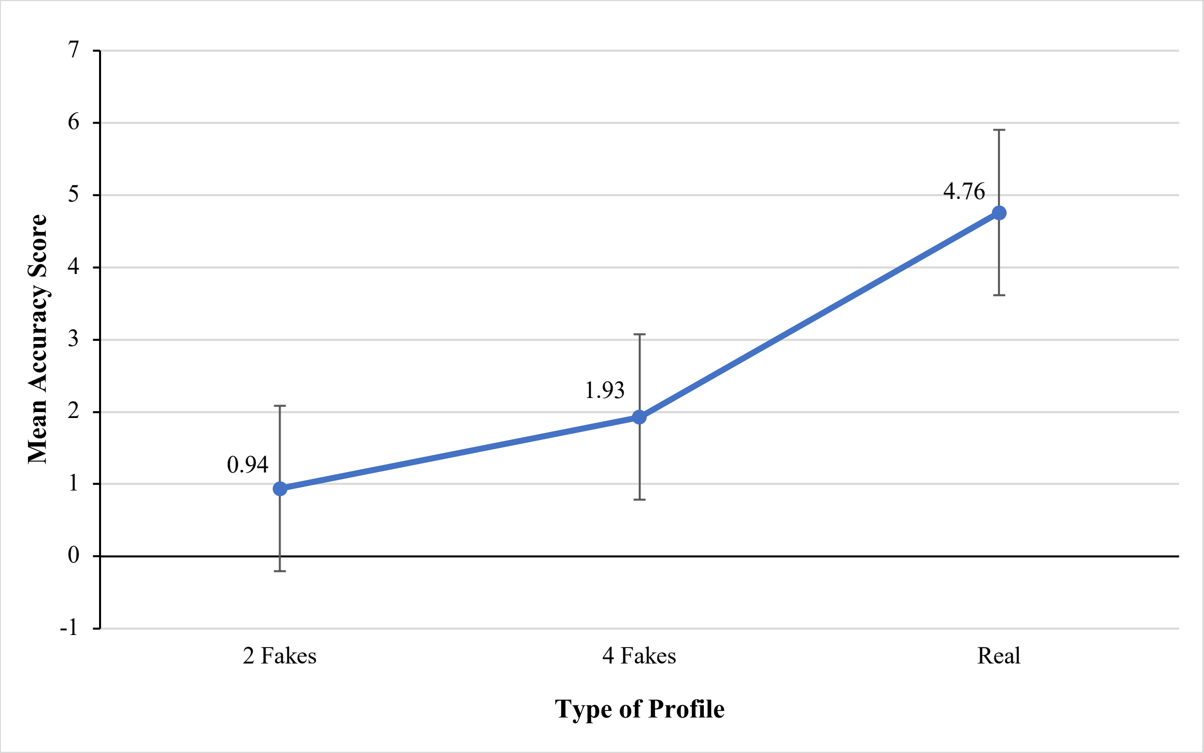
Raw data were exported from Qualtrics and initially sorted in Excel prior to importing into IBM SPSS Version 26.0 for Mac and R Version 1.3.1073 (R Core Team, 2020) for analysis.

Prior to conducting the statistical analyses, multiple normality tests were conducted to assess the appropriateness of the data for testing. The normality tests showed the data had four outliers, of which none were extreme, histograms showed mainly normal distributions (a small number had minimal positive or negative skew, but none were extreme), and all Q-Q plots showed data of a linear pattern.

Profile Accuracy

A linear pattern of mean accuracy across the three different types of profile, *2 fakes, 4 fakes*, and *real* profiles was found. Mean judgement accuracy scores, displayed in Figure 1, show that participants were more accurate at judging real Facebook profiles (*M* = 4.76; *SD* = 1.13) than fake Facebook profiles (*M* = 2.87, *SD* = 1.14); and were more accurate at correctly judging fake profiles with four fake characteristics (*M* = 1.93, *SD* = 0.82) than with two fake characteristics (*M* = 0.94, *SD* = 0.75). The results showed these differences to be significant, *F*(1.79, 357.81) = 906.92, *p* <.001[1]. These findings support H1 and H2 that respectively predict that participants' judgement accuracy will differ between fake and real profiles, and that profiles with more fake characteristics will be more accurately judged than those with fewer fake characteristics.

Figure 1
Mean judgement accuracy scores for each type of profile (N = 201).



To further investigate participants' mean accuracy scores, specifically whether their scores were better than the level of chance, a t-test was conducted using participant's overall accuracy scores. Results show that participants' overall accuracy levels were significantly different to the chance level of 6, *t*(200) = 15.31, *p*<.001,

[1] Mauchly's Test of Sphericity was violated, $x^2(2) = 27.08$*, p<.001.* As the Greenhouse-Geisser correction statistic was more than .75 (ε = .89), the Huynh-Feldt correction statistic (ε = .90) was used to correct the degrees of freedom.

suggesting that participants were able to judge profiles with a level of accuracy that was statistically significantly higher than what would have been expected by chance alone.

Maximum judgement accuracy, i.e., accurate judgements of all profiles seen, was also measured. A similar linear trend was found for maximum judgement accuracy as was found for mean judgement accuracy; 1% (*N* = 2) of participants accurately judged all *2 fakes* profiles as fake, 25.9% (*N* = 52) judged all *4 fakes* profiles as fake, and 33.3% (*N* = 67) of participants accurately judged all the *real* profiles as real. However, zero participants achieved a maximum judgement accuracy score of 12 across all profiles viewed. Only 0.9% participants were close to achieving a maximum accuracy score as five participants scored 11 out of 12.

To further analyse participants' judgement accuracy across both fake and real profiles, Signal Detection Theory (SDT) was used. SDT seeks to measure the accuracy of decision-making in forced-choice tasks such as 'yes/no' tasks (Green & Swets, 1966). SDT proposes that decisions are made under conditions of uncertainty, and the decision-maker must decipher between the signal (stimulus), and the background noise (random variables) when making their decision. To test this, the probability that the participant says 'yes' when the stimulus is present and the probability that the participant says 'yes' when the stimulus is *not* present are measured. These probabilities are known as the 'Hit Rate' and 'False Alarm' rate respectively.

In regard to the decision-making process of judging the authenticity of the Facebook profiles, fake profile accuracy and real profile accuracy scores were transformed into hit rate (signal) and false alarm rate (noise) scores. A hit rate is where the signal was present (fake profile), and the participants accurately detected the signal (judged the profile as fake). A false alarm rate is where the signal was absent (real profile), and the participants accurately rejected the signal (judged the profile as real).

From this, d-prime (*d'*) values and criterion (*c*) scores were computed; *d*' is a measure of sensitivity used to indicate participants' abilities at discriminating between the signals (fake profiles) and the noise (real profiles) within the study, and was calculated by subtracting the *z* score of the Hit Rate (HR) from the *z* score of the False Alarm Rate (FA): *(d*' = z(FA) – z(HR)). The *c* score is a measure of participants' response bias (i.e., were participants biased more towards answering yes or no, or in this case, fake or real, when judging the profiles?) and was calculated by summing the *z* scores of the Hit Rate and False Alarm Rate together and multiplying this by -0.5 (*c* = -0.5 x (*z*(HR) + *z*(FA)).

Results indicate that participants were unable to distinguish between the signals (fake profiles) and the noise (real profiles) within the study and performed below chance (*d'* = -0.90). A negative *d'* is indicative that participants' performance may have been due to a misunderstanding or confusion with the task and what was required of them, sampling error or some other methodological effect. As participants were unable to accurately distinguish between the fake and real profiles, any inferences made from the mean accuracy scores for each profile type and corresponding ANOVA test reported above are to be taken with caution.

Participants showed a bias towards responding 'yes' with a liberal criterion of *c* = -1.83, meaning that they were biased towards judging the profiles as fake. However, as reported previously, participants were more accurate at judging real profiles than fake profiles, hence the criterion score suggests that the fake profiles were deceptive enough to fool participants into judging the profiles as real, thus going against their response bias.

Self-Reported Accuracy

Participants were asked at the end of the study, after making all judgements of each profile, how accurate they think their judgements were. The overwhelming majority (*N* = 64, 31.8%) reported feeling ‘Slightly Confident’. Very few participants (*N* = 6, 3%) reported feeling ‘Unconfident’ and only (*N* = 8, 4%) reported they were ‘Confident’ in the accuracy of their judgements. To investigate whether participant confidence levels in the accuracy of their judgements had a relationship with their actual accuracy, three Pearson’s correlations for each accuracy score were ran. As self-reported accuracy was scored based on a 7-poimt Likert scale it was treated as a continuous variable, hence the use of Pearson’s correlation test. Table 1 displays the results.

Table 1.
Pearson’s Correlations for Self-reported accuracy against each accuracy score

Variables	*M*	*SD*	*1*	2	*3*	*4*
1. Fake Profile Accuracy	*2.87*	*1.14*	*1*			
2. Real Profile Accuracy	*4.76*	*1.14*	*-.09*	*1*		
3. Total Accuracy	*7.64*	*1.52*	*.67***	*.66***	*1*	
4. Self-Reported Accuracy	*4.32*	*1.51*	*-.05*	*.08*	*.06*	*1*

Note. ***p* <.01

As is evident from Table 1, there are no significant relationships between participants’ self-reported accuracy and actual accuracy scores, whether that be for fake, real, or total accuracy. As such, H7 that stated that there will be a relationship between self-reported accuracy of judgements and actual accuracy of judgements, cannot be accepted.

Table 1 does show two significant relationships present between fake accuracy and total accuracy, and real accuracy and total accuracy. As total accuracy is the sum of real and fake accuracy such result can be expected.

Personality

To test for the effect of personality and Social Sensitivity (SS) on participants' accuracy score three multiple regression models were conducted: one for each accuracy score (fake accuracy, real accuracy, total accuracy). Prior to analysis, assumption testing was conducted to ensure the appropriateness of the data for the statistical test. Linearity was assessed via visual inspection of the personality predictor scatterplots. The data points were overplotted on one another due to the data being discrete. As such, noise was introduced around each data point by 'jittering' to allow the points to split apart from one another. Adding this noise showed a linear relationship, and homoscedasticity of the data, between each personality trait or score and accuracy scores. The data also met all other assumptions, including independence of residuals and Cook's & Leverage values all being within the appropriate range. As such the multiple regression was deemed a suitable test for the data, so all five personality traits and SS scores were entered as predictors for each accuracy model. It was expected that a relationship will be found between personality type (H3) & social sensitivity (H4) and judgement accuracy. Results reported in Table 2.

Table 2.
Multiple regression for personality predictors of fake accuracy, real accuracy, and overall judgement accuracy.

Predictors	Fake Profile Accuracy			Real Profile Accuracy			Total Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
TIPI	.041			.006			.036		
Extraversion[a]		0.04	0.06		0.05	0.06		0.09	0.08
Agreeableness[a]		-0.05	0.08		-0.04	0.08		-0.10	0.11
Conscientiousness[a]		-0.02	0.07		-0.02	0.07		-0.04	0.09
Emotional Stability[a]		0.04	0.07		<.001	0.07		-0.04	0.10
Openness to New Experiences[a]		0.17*	0.07		<.001	0.07		0.18	0.10
SS		0.01	<.001		<.001	<.001		0.01	0.01

Note. [a]*df* = 6, 200. *$p < .05$.

The results, as in Table 2, show that those high in openness to experience are most likely to accurately predict fake profiles (p = .023), however the regression model was non-significant, F (6, 200) = 1.38, p = .224. There were no significant predictors of real profiles, or overall accuracy, with both models being non-significant; Real profiles (F(6, 200) = 0.19, p = .978), Total accuracy (F(6, 200) = 1.19, p = .312). Overall, personality traits and levels of social sensitivity are not good predictors of judgement accuracy, thus H3 and H4 cannot be accepted.

Social Media

Platforms

Participants were asked, in relation to their social media use, to select which social media platforms they use and then rank them in order of platforms they use most often. All 201 participants selected and ranked at least one social media platform they use, with only one participant selecting and ranking all seven platform options. Even the two participants who stated they do not regularly use social media ranked at least one platform that they do use when they use social media. Table 3 outlines the social media platforms used by participants and the frequency of their individual rankings.

Table 3.
Participant rankings of social media platforms based on the platforms they use most often (1^{st} ranking) to the platforms they use the least (7^{th} ranking).

Social Media Platform	Ranking Order							Total number of participants who use each platform
	1st	2nd	3rd	4th	5th	6th	7th	
Facebook	171	0	0	0	0	0	0	171
Twitter	13	68	0	0	0	0	0	81
Instagram	10	94	57	0	0	0	0	161
Snapchat	3	7	31	31	0	0	0	72
TikTok	0	1	17	14	12	0	0	44
YouTube	3	23	55	54	32	12	0	179
Other	1	2	12	9	5	6	1	36
Total number of rankings made	201	195	172	108	49	18	1	

Facebook was ranked as the most used social media platform by all participants that selected Facebook as one of the social media platforms they use (*N* = 171), thus being the platform with the highest overall use amongst this sample. However, YouTube is the most popular platform amongst the participants (*N* = 179) but is not the platform used the most by the overwhelming majority of these participants, with only three ranking it in first place. Thirty-six participants listed Other platforms they use as WhatsApp, LinkedIn, Pinterest, Discord, Netflix, Quora, Spotify, Reddit, and Tumblr.

Purposes

Participants reported using social media most frequently for 'Socialising with friends/keeping in touch with friends' (*N* = 179, 89%) and 'Watching videos (TV series/Films/YouTube etc.)' (*N* = 179, 89%); followed by 'News (keeping up with current events)' (*N* = 132, 65%), 'Listening to music' (*N* = 125, 62%), 'Business purposes (Advertising/promoting products or brands)' (*N* = 29, 14%), and Other - "Learning to school", "Look at cute rabbit pictures", "entertainment", "Participation in groups to do with my interests", "Pass the time when bored", "Finding out things going on in the world", "Staying up to date on areas of interest", and "following organisations and activists" (*N* = 8, 4%).The purpose of 'Making friends/meeting new people' was only selected by 52 participants (26%), which suggests that the original purpose of social media being a platform to connect people is no longer the main reason people are using the platforms.

Daily Usage

To investigate whether participants' use of social media has an effect on their judgement accuracy, participants were asked if they were regular users of social media, and how many hours per day they use social media. Two of the 201 participants (0.99%) reported that they were not regular users of social media. Of the 199 participants who reported that they were regular users, the vast majority (91%) reported that they use social media for more than one hour per day. The most frequent was '2-3 hours' per day (N = 57, 28.4%), closely followed by four or more hours per day (N = 51, 25.4%).

To analyse the impact of social media use on accuracy judgements, three multiple regression models were run (one for each accuracy outcome variable) with hours spent on social media and as the predictor variable. Hours spent on social media was dummy coded with 'Less than one hour' as the constant in this regression model, against which the categories '1-2 hours', '2-3 hours', '3-4 hours', and 'More than four hours' were compared. Table 4 presents the results of the regressions.

Table 4.
Multiple regression of time spent on social media per day and fake profile accuracy, real profile accuracy, and overall accuracy scores.

Predictors	Fake Profile Accuracy			Real Profile Accuracy			Total Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Hours spent on social media per day	.046			.004			.045		
Constant		2.28	0.27		4.44	0.27		6.83	0.35
1-2 Hours		0.64	0.31*		0.24	0.32		0.76	0.42
2-3 Hours		0.60	0.30*		0.22	0.31		0.71	0.41
3-4 Hours		0.37	0.34		0.63	0.34		0.88	0.45
4+ Hours		0.88	0.31**		0.44	0.31		1.21	0.41**

Note. df = 4,196· *$p < .05$, **p <.01.

The results, as in Table 4, show that time spent on social media predicts accuracy for fake profiles, but not real profiles. Fake profiles are most accurately judged by those who spend up to 3 hours and over 4 hours on social media when compared to those who spend up to 1 hour on social media. Those who spend the longest on social media also show the best overall accuracy. However, overall, these models were non-significant.

Previous Experience in Creating a Fake Profile

Participants were asked if they had any experience in creating fake social media profiles. A total of 61 participants (30.3%) reported that they had: 44 Males (72.1%), 16 Females (26.2%), and 1 Non-Binary (1.6%). The reasons given for creating fake profile included surveillance (e.g., "For a friend to look at her ex-boyfriend", "To enter in some groups and try to investigate people", "To spy on my boyfriend at the time.") and anonymity/privacy ("Created accounts under another name to create some privacy for browsing through pages", "When websites ask to link a social media page and I don't want them to have access to my real data", "I didn't want some of my friends to know too much about me"). Only a few participants reported that they created the fake profile for malicious purposes ("To trick and have fun with some classmates that I didn't like", "For trolling", "…to scare a friend").

To investigate whether having previous experience of creating a fake profile can predict accuracy scores, a multiple regression was conducted using real profile, fake profile, and overall (real and fake) accuracy scores. As 'previous experience in creating a fake profile' is a categorical variable ('Yes'/'No') it was dummy coding to allow for input into the model. 'No' was used as the constant against which 'Yes' responses were compared. Results of this regression are presented in Table 5.

Table 5.
Multiple regression of previous experience creating a fake profile and fake profile accuracy, real profile accuracy, and overall accuracy scores.

Predictors	Fake Profile Accuracy			Real Profile Accuracy			Total Accuracy		
	R^2	*B*	*SE*	*R*^{2}	*B*	*SE*	*R*^{2}	*B*	*SE*
Experience in creating a fake social media profile	.002			<.001			.001		
Constant		2.83	0.10		4.76	0.10		7.61	0.13
Yes		0.12	0.18		-0.01	0.18		0.09	0.23

Note. df = 4,199.

Table 5 shows that participants' previous experience in creating a fake social media profile is not a significant predictor of either fake profile accuracy, real profile accuracy, or overall accuracy.

Through both descriptive and inferential testing, it has been shown that there are partial relationships between time spent on social media per day and judgement accuracy, and no relationship between previous experience in creating a fake profile and judgement accuracy. As such, H5 can be partially accepted and H6 cannot be accepted.

Manipulated Characteristics of Profiles

The impact on accuracy of the manipulated characteristics of the fake profiles (*Age, Number of Photos, Photo Type, Number of Friends, Number and Regularity of Posts,* and *Groups*) was analysed using general linear models using the 'lme4' package (Bates, Machler, Bolker, & Walker, 2015) in R. Two different models were run, the first using participants' judgement of the profiles (real or fake) and the second using participants' accuracy scores (total of correct and incorrect judgements). Both of the judgement and judgement accuracy models reported were ran with 'Prolific ID' as a random effect and 'Profile Number' (the random profiles participants viewed) as a nested random effect with 'Prolific ID'. The addition of 'Profile Number' as a random effect made little statistical difference to the results, with variation only observed on the fifth decimal place. As the randomisation was accounted for in the method (i.e., participants were randomly assigned six of the fake profiles), the *glmer* models including 'Profile Number' *have not* been reported here to avoid duplication of results.

Table 6 reports the results for both Models; Model 1 reports participant's judgements and Model 2 reports participant's accuracy scores. Both used the six manipulated characteristics as predictor variables and the participant Prolific ID number as the random effect.

Table 6.

Results from 'glmer' Model's 1 & 2 where judgement and accuracy are regressed on the manipulated profile characteristics.

Predictors	Model 1 – Judgement [b]					Model 2 – Accuracy [c]				
	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*	
Fixed effects										
Intercept	-1.49	0.08	0.19	0.26	<.001***	0.95	0.06	2.28	2.90	<.001***
Age [a]	0.30	0.11	1.08	1.69	.008**	-0.41	0.11	0.54	0.82	<.001***
Number of Photos [a]	0.51	0.11	1.34	2.08	<.001***	-0.20	0.11	0.66	1.01	.067
Photo Type [a]	1.21	0.11	2.69	4.19	<.001***	0.47	0.11	1.28	1.98	<.001***
Number of Friends [a]	0.29	0.11	1.07	1.67	.010**	-0.44	0.11	0.52	0.80	<.001***
Number and Regularity of Posts [a]	0.48	0.11	1.29	2.02	<.001***	-0.27	0.11	0.62	0.94	.011*
Groups [a]	0.17	0.11	0.95	1.47	.142	-0.59	0.11	0.45	0.68	<.001***
Random effects										
Residual Variance (σ^2)	3.29					3.29				
τ_{00} PROLIFICID	0.11					0.04				
Intraclass Correlation Coefficient	.03					.31				

Note. Number of Participants = 201, Number of Profiles = 74, Number of Observations = 2412. *p =.05, ** p =.01, *** p<.001.

[a] Model 1: 0 = Judgement of Real, 1 = Judgement of Fake; Model 2: 0 = Non-Accurate Judgement, 1 = Accurate Judgement. [b] Conditional R^2 = .209. [c] Conditional R^2 = .326

In regard to Model 1, Table 6 shows that all of the characteristics, with the exception of *Groups*, are significant predictors of participants' judgements. *Photo Type* is the strongest predictor, meaning that if *Photo Type* is manipulated on the profile (i.e. profiles display photos showing Celebrities, cartoon characters, landscapes, or artwork), participants are more likely to judge that profile as fake (B = 1.21). The same results are found for all other characteristics, except for groups. Interestingly, none of the manipulated characteristics predicted a likelihood of judging the profile as real as none were below zero. These results support H8, which predicted that the different types of manipulated characteristics on the profiles will have an influence on the judgement participants give (fake or real).

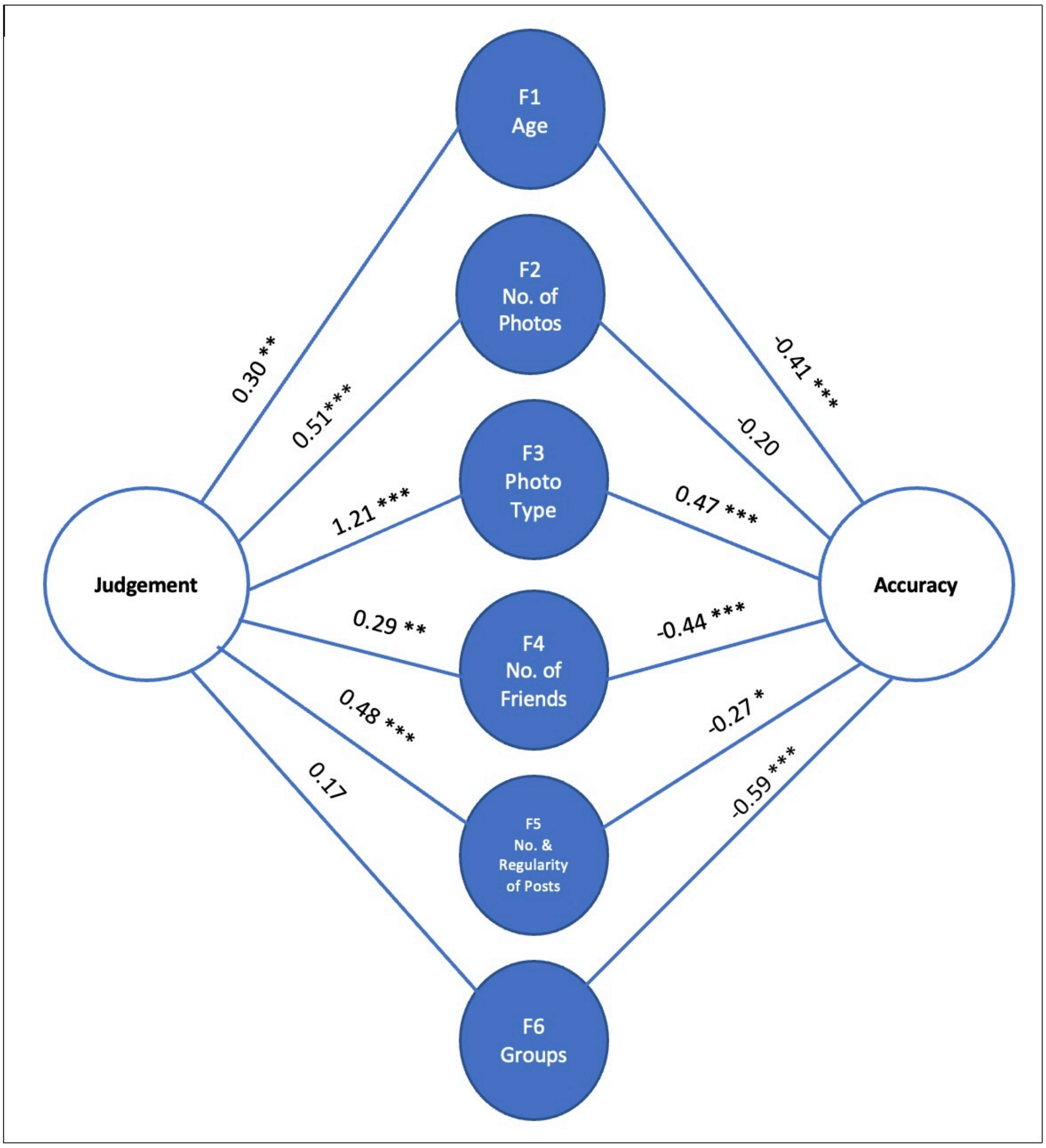
Model 2, reported in Table 6, shows that *Number of Photos* was the only characteristic that failed to predict judgement accuracy. With the exception of *Photo Type* (*B* = 0.47), the manipulation of all other characteristics significantly increases the probability of a non-accurate judgement, i.e., participants are more likely to inaccurately judge a profile as real or fake (respective to the type of profile being judged), when the factors of *Age, Number of Friends, Number and Regularity of Posts,* and *Groups* were manipulated on the profiles. In contrast, *Photo Type* increases the probability that an accurate judgement will be made. As the different characteristics had an effect on the accuracy of participants' judgements, H9 can be accepted.

Considering both models collectively, when *Photo Type* is manipulated on the profile participants are more likely to judge that profile as fake and this judgement is more likely to be an accurate judgement; photos showing celebrities, cartoon characters, landscapes, or artwork, aid participants in judging the profile accurately as fake. This suggests that participants may be relating authenticity of the profile to identity.

Both Model 1 and Model 2 were compared against each other using a Brunswik Lens Model approach. The Brunswik Lens Model, developed by Egon Brunswik (1956), is a probabilistic theory of perception and decision making that seeks to assess how characteristics influence judgement. The model posits that the decision-making/judgement process is composed of three elements; cues available, observed decision, and correct decision, and recognises a persons' decision or judgement, and the criterion being predicted, as two separate functions of the cues available within the environment. The accuracy of the decision is dependent on the predictability of the criterion on the basis of the cues and how these cues match the environment (Karelaia & Hogarth, 2008). Regarding this study, the cues available are the manipulated characteristics in the profiles, with the observed decision (criterion) being participants' judgement (whether the profile is real or fake), and the correct decision is whether that judgement was accurate. Figure 2 presents the lens model approach comparison between Models 1 and 2 with the estimates and statistical significance for each of the factors shown.

Figure 2.

Lens model diagram showing the estimates and statistical significance for each predictor variable (factors) and both outcome variables (judgement and accuracy).



Note. *p = .05, ** p = .01, *** p<.001

Figure 2 shows that across both models, Factor 3 (*Photo Type*) is the strongest predictor of judgements: the manipulation of *Photo Type* in a profile significantly increases the probability that participants will judge the profile as fake (Model 1) and said judgement will be accurate (Model 2). This suggests that participants may be over-relying on *Photo Type* in their judgement.

Post-Hoc Analysis

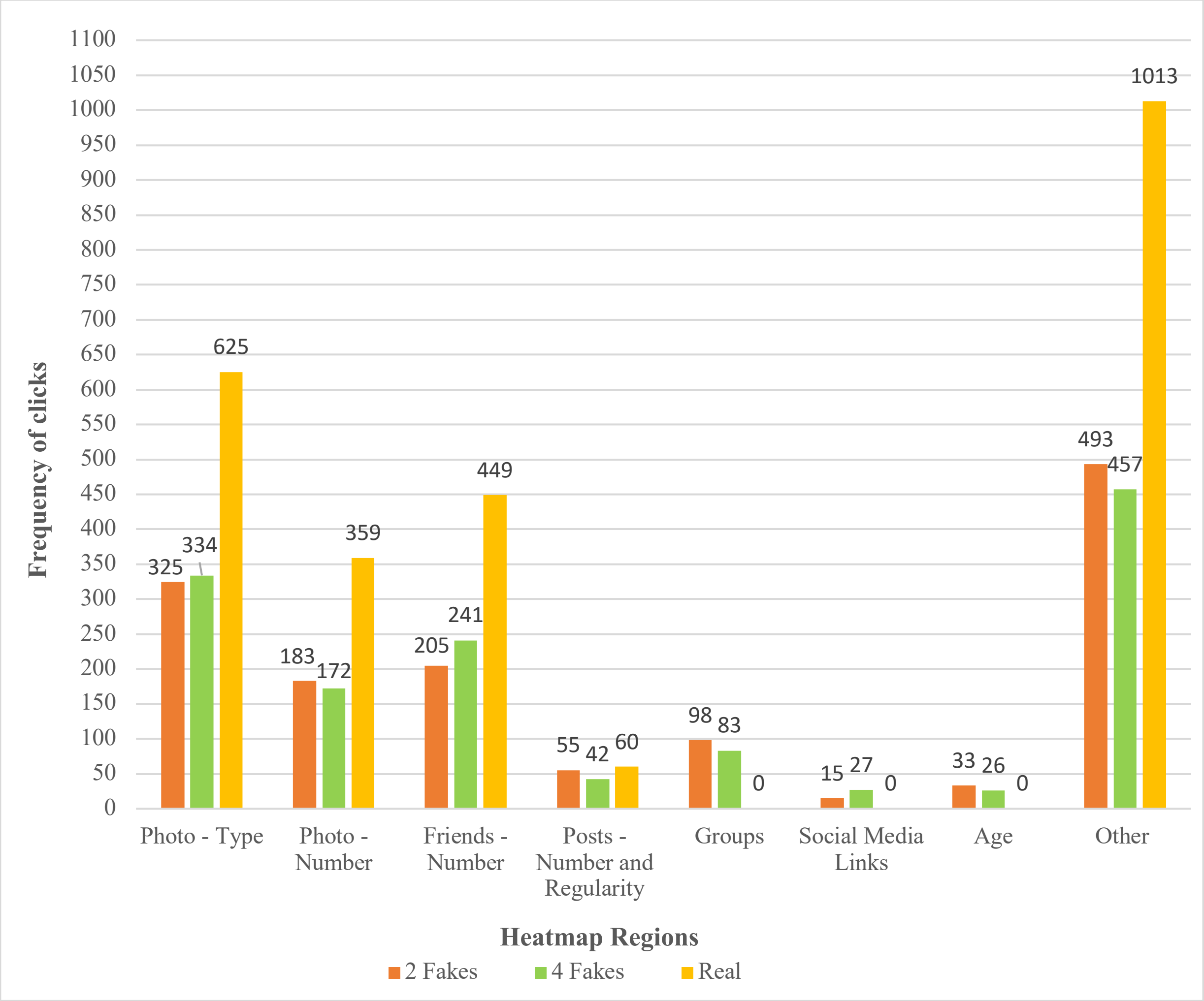
Heatmaps

Heatmap images were used to analyse the specific areas of each profile participants used when making their judgements. Each profile shown within the study had a heatmap layer, meaning any clicks on the profile were recorded. To capture the areas clicked on, heatmap regions were added to each profile by the researcher. These regions related to the characteristics on each profile that were manipulated; *Photo – Type, Photo – Number, Friends – Number, Posts – Regularity, Posts – Number, Groups, Social Media Links,* and *Age*. Any clicks within these regions would be recorded as a click for the respective manipulated factor listed previously. The regions were made as wide as possible, without overlapping with other regions, to ensure that slightly inaccurate clicks were encapsulated within the regions. The heatmap regions were only visible to the researcher during the creation of the study and the analysis stage (Appendix K).

Each participant was allowed a maximum of ten clicks per profile viewed, with each click showing on the profile as a red dot (Appendix L). The results of the recorded clicks on each profile are denoted by colours on a heatmap scale. The blue areas are at the lower, or ‘cooler’, end of the scale and represent few clicks, whereas red areas are at the higher, or ‘hottest’, end of the scale and represent a large number of clicks. The frequency of the clicks in each area for all participants across the fake profiles with both two and four fake characteristics, and real profiles are shown in Figure 3.

Figure 3.

Graph detailing the number of clicks in each heatmap region for both profiles with 2 fake characteristics and 4 fake characteristics, and real profiles.



As shown in Figure 3, the characteristic participants relied upon most, across all profile types when making judgments is Other. The characteristics of *Photo-Type, Photo – Number, and Friends – Number*, were also relied upon heavily. As reported above, the factor of *Photo-Type* was the strongest predictor of *both* participants' judgement and accuracy, which is reflected here in that it is the second most clicked area of both types of fake profiles and real profiles and is the most clicked area of the manipulated factors.

To assess the areas of the profile encapsulated under the umbrella of *Other*, each of these clicks were manually checked. These areas included relationship status, content of posts and comments, number of likes on posts and comments, Intro section, life events, location of posts, employment, university, memories, and recommendations.

Further analyses were carried out on the heatmap frequencies to examine if participants were clicking on different areas to inform real/fake judgements across the different profile types. Figure 4 shows the frequencies of clicks.

Figure 4.
Number of clicks in each heatmap region split by real/fake judgement for each type of profile.

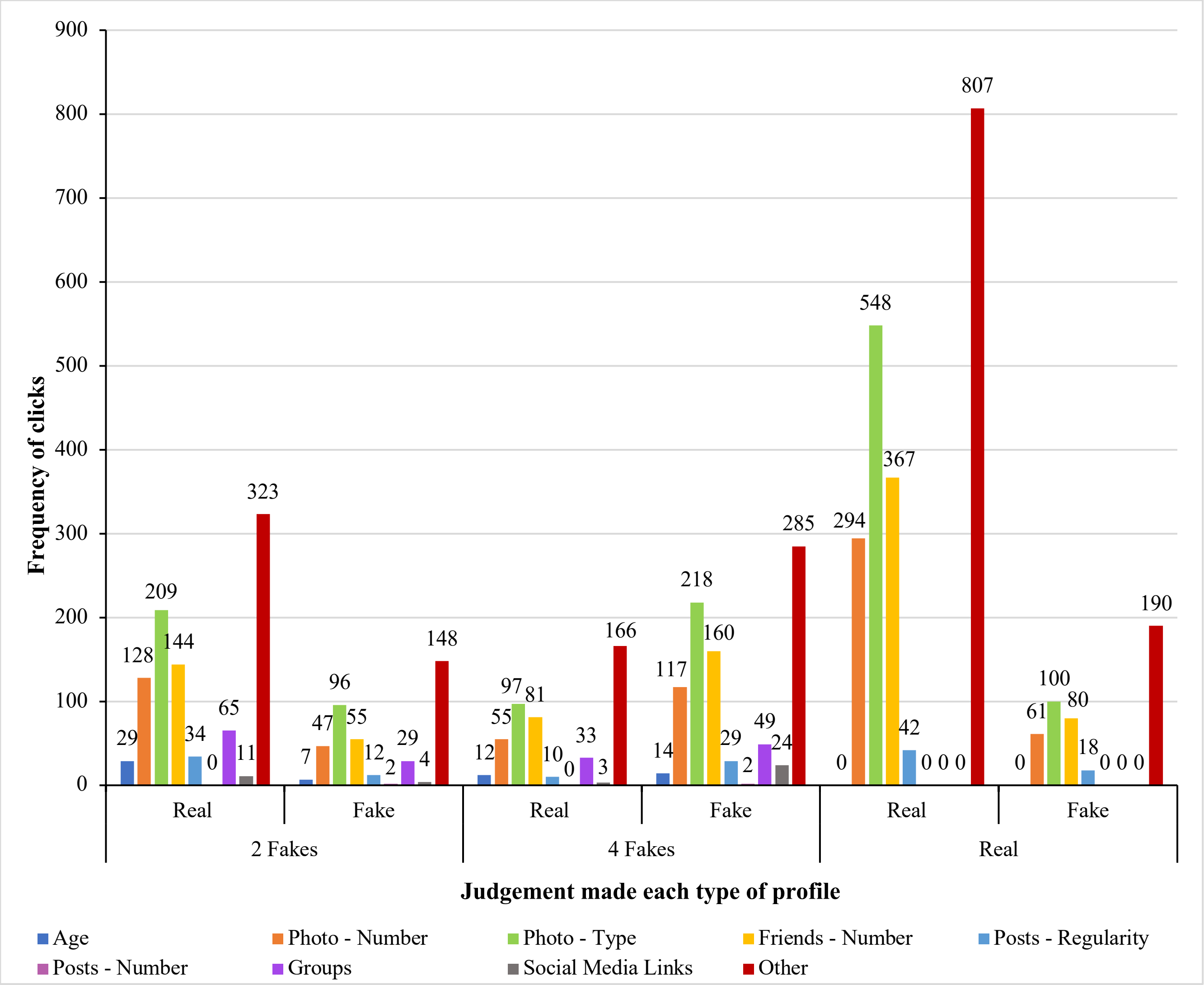


Figure 4 shows that when presented with two fake characteristics, participants clicked most on Other followed by 'Photo-Type' to inform their judgement, irrespective of the whether this judgement was real or fake. More absolute clicks were related to real judgements (*N* = 943 total clicks versus *N* = 400 total clicks when making a fake judgement). Interestingly, the profiles with 4 fake characteristics revealed that a higher number of absolute clicks were related to fake judgements rather than real judgements (*N* = 898 and *N* = 457, respectively). However, the pattern of clicks was the same as that found for profiles with 2 fake characteristics: *Other* and *Photo – Type* were the most clicked on areas.

The same pattern of results was found for real profiles, with the vast majority of clicks being in the *Other* category, followed by *Photo-Type*, again irrespective of whether the judgement is real or fake. The characteristics of *Age, Posts – Number, Groups*, and *Social Media Links* were not clicked at all, which may be attributed – in part – to the fact that these characteristics were not present on the profiles (as they were real profiles and not manipulated by the researcher) and thus were not outlined as heatmap regions.

Deceptive Purposes of Profiles

Participants were asked after judging each profile whether they believed the profile had been created with deceptive purposes or malicious intent. Figure 5 displays the frequencies of yes/no answers for each type of profile.

Figure 5.
Graph showing the frequency of yes/no answers to the question 'Do you think this profile was created for deceptive purposes?' for all profile types.

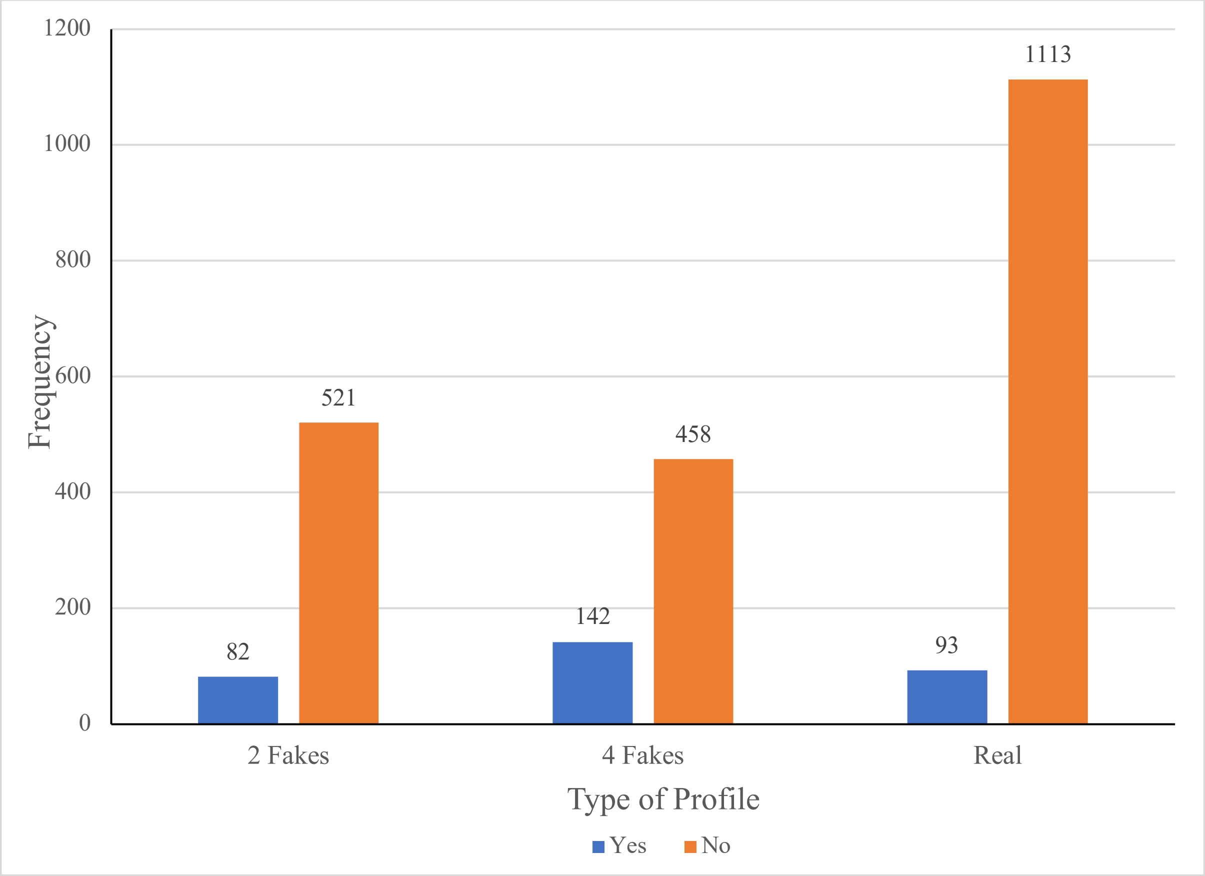


Figure 5 shows that across all profile types, participants were most likely to perceive no deceptive intent. This difference between those that agreed with this view and those that disagreed is most noticeable for real profiles and least noticeable for profiles with 4 fake characteristics. Interesting to note here is that participants thought the real profiles were the most deceptive.

Discussion

The purpose of this study was to investigate online deception in the form of fake Facebook profiles; specifically identifying the aspects of profiles that people rely upon when judging their authenticity. Overall, the results show that participants' levels of accuracy differ when judging different types of profiles, with highest accuracy being found for judgements of real profiles, and lowest accuracy for judgements of fake profiles with *2 Fakes.* Additionally, different manipulated characteristics were found to have different influences over both participants' judgements and judgement accuracy. The manipulated factor that was consistently highly significant for both participants' judgements and the accuracy of such judgements was *Photo Type*, suggesting that participants may rely more so on the visual elements of the profile when making their judgement. However, as the heatmap data has shown that *Other* was the most clicked on area of the profile, this suggests that perhaps participants rely more so on a completely different area of the profile than the type of photo presented. Further, judgement accuracy was not influenced by a person's personality, social media use, self-reported accuracy or personal experience in creating a fake profile.

One possible explanation of participants' improved ability to judge real profiles more accurately may relate to the content of profile. The real profiles showed several posts from others wishing the profile owner a happy birthday or 'check-ins' where profile owners and friends share their location, for example 'check in' as being at their holiday hotel, or a friend's house. These types of posts were absent from the fake profiles and with this so was the social proof such messages can signal of appropriate behaviours or opinions. Social proof is defined as a tendency to view behaviours as correct, or more appropriate, when others engage in the same behaviours (Gass & Seiter, 2014, p.132). Social proof has been found to effect; trust of an organisation when buying an item from said organisation (Talib & Saat, 2017), moral decision-making in ethical dilemma tasks (Pitsea & Thau, 2013), and behaviour in virtual environments (Shepherd, Lane, Tapscott, & Gentile, 2011). Regarding social media and the link with social proof, it has been found that individuals are more likely to 'like' or 'follow' a page if they see that several others are doing the same (Muscanell, Guadgno, & Murphy, 2014). Therefore, showing that the user has an active social circle through the happy birthday or ‘check-in’ posts on the real profiles could demonstrate authenticity of the profile - as others are contributing to the person’s profile they must be under the impression, or know factually, that this profile is that of a ‘real’ person. Thus, others are influenced to also believe the person to be real, and in turn the profile, meaning they judge the profile to be real. The absence of these posts on the fake profiles could explain why they are judged less accurately than the real profiles.

The linear trend regarding the profiles with four fake characteristics being judged more accurately than those with two fake characteristics showed that the presence of more manipulated factors aided participants in making an accurate judgement of the authenticity of the profiles. Research has found that when explicit expressions of social proof on social media are unavailable, such as comments on the profiles from others, people will turn to finding more subtle cues of social proof such as number of likes (Amblee & Bui, 2011; Lee, Lee & Oh, 2015). The profiles with four fake characteristics had more cue information that suggested they are fake than those with two characteristics, for example one of the *4 Fakes* profiles had only two posts (including minimal likes and comments), one photo, eight friends, and zero groups. Whereas one of the *2 Fakes* profiles had only two posts and zero groups but did show lots of photos (selfies and group photos) and lots of friends. Based on the theory of social proof, participants would be able to shift their attention to different cues regarding authenticity more so with four fakes profiles than two fakes, as more of the information denoting it as fake is readily available, i.e. not much engagement from either the profile user or friends on this 4 Fakes profile demonstrates low social proof, suggesting participants would have little to no social cues from others to suggest the profile is real, thus leading them to make a judgement of fake. Whereas, on the *2 Fakes* profile described above, participants would be able to shift their attention from the posts (including likes and comments, or lack of) to other areas that may suggest social proof such as high number of friends and lots of photos of the individual both alone and in groups. Thus, higher levels of judgement accuracy for *4 Fakes* profiles when compared to *2 Fakes* profiles could be explained by social proof, in that the presence of more manipulated characteristics on the *4 Fakes* profiles resulted in less available information from others to base authenticity judgements on. However, as the likes and comments were not specifically manipulated within this study, further research is warranted where these areas are manipulated. Doing so may provide further evidence to support the use of social proof as an explanation of the linear trend found with judgement accuracy.

The heat map data offered a way to explore areas of each profile that were examined by the participant, which we assume was used to inform their judgements. These data showed a similar pattern of click rate across the three types of profile. *Photo Type* emerged as the most clicked area of manipulated characteristics, while *Other* was the highest overall. Focusing on the specific area's participants used when making their judgements, the heatmap click data showed that it is evident that participants used/relied upon the factors of *Photo Type*, and *Photo Number* the most when making their judgements, particularly the characteristic of *Photo Type* as this has the second highest number of clicks for both real and fake profiles. Interestingly, the region with the highest frequency of clicks across all profiles is *Other*, meaning participants did not click on any of the manipulated factors but other areas entirely, such as the bio/intro section (relationship status, location, job etc), and number of likes or comments on the posts. Due to the volume of clicks under the umbrella of *Other*, it would be reasonable to assume these are areas that participants rely upon when making judgements. Further studies could focus on manipulating some or all of these areas to analyse whether *Other* clicks are still made elsewhere, or whether *Photo Type* is still the most clicked on area of the profiles.

The high absolute and relative click rate for *Photo Type* suggests that visual information is used by people as diagnostic information of authenticity. This has been shown in other areas. For example, Facebook profile photos have been shown to influence personality judgement of the person depicted, with smiles also being used to inform the person's trustworthiness (Toma, 2014; Willis & Todorov, 2006). The link between smiles and trust is not specific to Facebook profiles and has been shown to shape trustworthiness ratings in other domains too such as political voting (Ballew & Todorov, 2007) and criminal sentencing decisions (Blair, Judd, & Chapleau, 2004), with some suggestion it may operate at an automatic level (Todorov, Pakrashi, & Oosterhof, 2009; Getov, Kanai, Bahrami, & Rees, 2015). However, as an explanation, this only partly explains the results as not all images were of faces. Although, Ivcevic & Ambady (2012) found that when comparing raters' personality judgements of individuals' Facebook pages, the consensus between raters was greatest when their judgements were based on profile pictures alone compared to when using all available information available. This suggests that profile pictures are replied upon most when making judgements, or forming impressions, of others online, and when done so can be done so accurately. However, in specific relation to credibility judgements Sandi, Rusconi and Li (2017) found that profile photos on social media accounts *did not* influence participants' credibility judgements of the accounts, which suggests that there are other areas of the profile that influence participants' judgements. Thus, further research is needed to identify the types of images that inform judgements and specific areas relied upon most when making judgements as to the credibility or authenticity of the profile being judged.

The findings of this research highlighted important limitations. Firstly, not all fake profiles made visible *Age, Post-Number, and Social Media Profile Links*. Addressing the factor of *Age* first, age was one of the six directly manipulated factors, however only 16 out of the total 30 fake profiles created displayed the age. Due to Facebook not having a specific section for age to be displayed on the profile, the researcher displayed the age in either the *Bio* section, or in specific posts on the profile timeline. In an effort to create a wide range of realistic profiles, age was not categorically displayed in either of these ways in 14 of the profiles, but rather through the photos displayed, i.e., photos of the young teenagers, or photos of the over 70 generation. However, in doing this, an overlap between the two factors of *Age* and *Photo Type* has been created, thus meaning it is difficult to say whether a participant who clicked on the profile photo of the profile was clicking on it because of the type of photo shown or because of the age of the person within the photo.

Regarding the factor of *Post Number*, Facebook does not have a specific feature for displaying the number of posts on each profile timeline, but rather a large dot underneath the first post on the timeline to signify that there are no more posts on the timeline before that point. Each profile screenshot displayed between 6-8 posts, unless *Post Number* was being directly manipulated, in which case the screenshot would display 3-4 posts and the large dot at the bottom of all posts. Regular users of Facebook would most likely be aware of this large dot and the significance of it, and as all participants stated they used Facebook (with 86% ranking it as their most used platform), it is warranted for the researcher to have assumed participants would have knowledge of the dot on the profile in question. However, it is recognised by the researcher that this particular factor is relatively obscure and may question the reliability of the results found. Additionally, it is rather presumptuous to expect

Facebook users to firstly notice the dot, and secondly to understand what this dot signifies. However, it must be remembered that this is an original study with little to no previous literature base to expand upon and can therefore be seen as an exploratory piece of work. In future, it would be advisable to emit this factor as a directly manipulated factor in further studies due to the grey area associated with it. It can however provide some input to the literature and perhaps can be expanded upon further in the future.

Similarly, *Social Media Profile Links* was not a directly manipulated factor due to the lack of literature to inform the researcher and provide a basis for inclusion within the research. It was included as an exploratory factor to investigate whether the presence of links to other social media profiles on different social media platforms, such as Twitter, Instagram, or Snapchat, had an effect on profile judgements, i.e., did participants rely upon the presence of these links to make their decision as to whether the profile was real or fake, due to the assumption that the profile was legitimate as it links to other social media platforms? As this was an exploratory factor, and due to the need to create a wide range of diverse profiles, this factor was only included in 19 profiles. Overall, *Social Media Profile Links* was one of the areas with lowest number of clicks on the heatmaps, meaning that it was not heavily relied upon when making judgements, and is not an element that immediately highlights the need for further manipulation. However, this is an interesting finding which suggests that perhaps the presence of other profiles does not suggest that the profile itself is authentic, but rather other, more visual factors, are relied upon more when making judgements of authenticity.

A further issue with the Facebook profiles is that the *Real* profiles obtained did not have the following factors present on the screenshot; *Posts-Number, Groups, Social*

Media Profile Links, and *Age.* This is because these profiles were not directly manipulated in any way. The only changes made to the profiles was the omission of all identifying information from the screenshot for the purposes of anonymity. As such, this means the real profiles and fake profiles are not directly comparable in that they do not have the same factors present on each. However, each of the fake profiles created did not have the same manipulations – each profile had a different combination of either two or four fake factors, thus there should be no significant impact of the real profiles lacking in some of the factors, and thus no issues with comparing the real profiles to the fake profiles, particularly as the fake profiles are directly compared to one another throughout the research.

As mentioned previously, the real profiles also had different types of posts on them than were on the fake profiles, namely the happy birthday posts, and so could be said to have stood out more than the fake profiles. Posts such as these were not included on the fake profiles as it is very difficult to replicate such a post using Photoshop software, without creating a fake Facebook profile on Facebook itself along with a series of fake profiles to act as 'friends' and add comments on the original fake profile. This element is one which is difficult to circumnavigate, mainly for ethical reasons, but is one which should be explored within further research to ascertain whether participants are clicking on the content of the posts and relying on such to make their judgements.

Additional to the limitations of the factors manipulated on each profile, the profiles as a whole are limited in that the fake profiles were physically created and manipulated by the researcher, so effectively it could be argued that all aspects of the profile were manipulated and therefore fake, not just the specific factors investigated within the study. However, in response to this, it is near impossible for the researcher to gather fake profiles directly from Facebook for many reasons. Firstly, the researcher is unable to collect such information without contacting each profile user, informing them of the study, and obtaining their consent, which is something that would not only be time consuming and unfeasible, but would also violate Facebook's community guidelines and ethics. Second, and perhaps most problematic, is the issue of identifying which profiles are fake. Without a list of the specific aspects of the profile to look out for that would denote the profile as fake, the researcher would be unable to categorically say the profile is fake, hence the need for this research and the development of such list or framework. As such, the method of creation of these profiles is the closest replication of an actual Facebook profile, and therefore provides an as accurate as possible study on Facebook profiles and their content. Additionally, it has been found that judgements of online profiles are equally as accurate when judging a condensed profile showing limited information as they are when using the full profile (Stecher & Counts, 2008), suggesting that a screenshot of the profile should provide sufficient information for accurate judgements to be made.

Another important point to note is that not all participants judged six profiles as real, and six as fake, some judged more than six as either real or fake, meaning they did not adhere to the six and six rule. This may be due to several different factors, firstly participants may have misunderstood the instructions shown prior to the profile judgement section of the study and not noted the phrase 'you will be shown 12 profiles, six of which are real and six of which are fake'. Additionally, participants could have also forgotten this instruction after the first few profiles seen and/or also lost track of the judgements they had already made, thus making more or less of the required amount. Finally, and perhaps of most interest, participants may have genuinely thought that more than six of the profiles they viewed and judged were either fake or real, hence why they made more than six judgements. In future research, the instructions should be made clearer to try and reduce, and potentially irradicate such an issue, however if participants are of the view that more than the described number of profiles are either fake or real, then there will always be some who make more judgements of either fake or real than is specified. It is important to remember that such judgements of more than six profiles does not increase the accuracy, as there were only six fake profiles and six real profiles. However, it does highlight that perhaps participants are fixating on one element of the profile to make their judgements, and if this element or characteristic is present in the majority of the profiles, they see then their total number of judgements will likely be more than the six total.

The results of this study show that human judgement cannot be solely relied upon to accurately identify an inauthentic social media profile, due to the accuracy levels of participants' judgements being only slightly better than that of chance. However, what can be taken from this study is that humans *can*, to some extent, accurately identify a fake social media profile and distinguish between a fake profile and a real profile under these conditions. The implications of this are that human judgement *can* provide a layer of protection for users against fake profiles and their behaviours, specifically scamming behaviour, or behaviour with malicious intent. With further honing and testing, the levels of judgement accuracy in the future may increase extensively, which could mean that human judgement would become a vital method of online deception detection, and thus could prove useful to security services, those threatened by fake social media profiles, or the social media companies themselves.

Future research will involve further exploration into the individual factors or outside influences that may affect accuracy of judgements of the authenticity of social media profiles, as a response to the lack of significant results found in support of the co-variates and their related hypotheses. Additionally, as the most clicked on factor of the profiles is *Other*, a further research study will be undertaken with a focus on manipulating different factors within the profiles, such as the content of the posts, and the information included in the *Intro* section (i.e., relationship status, job, school, etc.), whilst removing the factors that were hardly relied upon *(Age, Social Media Links, Posts Regularity*). Doing so would help further in determining the specific areas of a Facebook profile that signify the profile as fake and contribute to the end goal of creating a framework of fake profile characteristics. Such a framework could provide a useful tool for those exposed to social media profiles, particularly those in enforcement who are trying to identify such profiles for the purposes of anti-terrorism or cybercrime.

Overall, the results of this study demonstrate that humans do have the ability to accurately identify a real social media profile and a fake social media profile and distinguish between the two at a level better than chance. However, as the accuracy levels were not at the top end of the scale, these results also highlight that there is a need for a comprehensive framework of factors within a social media profile that denote such profiles as fake, to enhance a humans' ability of accurately identifying a fake social media profile.

Chapter 3: Study 2

Introduction

The previous study showed that participants were more accurate at judging real Facebook profiles as opposed to fake profiles and were more accurate as the number of manipulated fake characteristics increased. Of these characteristics, *Photo Type* was a strong predictor of participants' judgements of the profiles and the accuracy of such judgements, suggesting that perhaps participants rely heavily on the visual aspects of the profile when making their decision as to whether the profile is authentic. Additionally, the heatmap data showed that participants clicked on the manipulated characteristic of *Photo Type* the most across nearly all of the profiles. They also showed a significant number of clicks on the *Other* category, a category found to encompass areas of the profiles such as relationship status, job, comments, likes on posts, likes on comments, content of the posts, and location.

This study systematically examines this *Other* category in more detail to identify specific other characteristics that may influence authenticity judgements. From looking through each heatmap individually and existing literature, *Bio, Intro* (including job, relationship status, school, university, location), *Posts Content, Number of Comments,* and *Number of Likes* were suggested as important. Supporting research has found that including name, photos, status, school, and gender in condensed profiles helped participants to form impressions (Stecher & Counts, 2008). Such research can provide support for the focus on the *Bio* aspects of the profile, as this can contain status, school, and gender, and a plethora of other personal information. Further, Young (2013) found that people primarily use status updates on profiles to make judgements about others and such updates that reflect the audiences' perception of the authors' real self are more highly valued in comparison to mundane status updates that are not as highly appreciated. Young (2013) also found that interactions on online profiles between friends through the comments section on posts is a source of analysis for others who are not part of the conversation, and that interactions are essential in facilitating the online social networking process.

Additionally, extensive research into impression formation online has resulted in the development of warranting theory (Walther & Parks, 2002), which suggests that when people are evaluating information in the online space, they are judging the warranting value of the information, or rather the extent to which the available information is immune to manipulation. Information with higher warranting value is perceived to be more immune to manipulation and thus more reliable than information with a lower warranting value. In relation to social networks, impression cues generated by the system (social media platform) and by others (friends) tend to be relied on most by people when forming an impression, rather than cues generated by the user themselves. The most salient cue being posts on the profile, specifically the messages, or comments, left on the posts by the friends of the user (Antheunis, Valkenburg & Peter, 2010).

These pieces of research suggest that content of the posts and the number of comments are important characteristics to consider and provide further support for the researcher's informed decision, from the analysis of the *Other* category, to manipulate these characteristics within this study.

Expanding on from Study 1, the profile characteristics of *Photo Type* and *Number of Photos* are again included in this study to judge the influence of these on the other characteristics introduced within this study, and also to measure whether participants still rely heavily on these characteristics, specifically *Photo Type*, above all others. This study also changed design from repeated measures to independent measures through the introduction of participant conditions. This revision was made to assess whether accuracy improves when participants are exposed to only one type of fake profile, or whether as in Study 1, participants continue to have higher accuracy when judging the real profiles. A further point of interest is whether participants' accuracy in Study 1 was influenced by the fact they were judging two different types of fake profiles, and using a judgement of one type to inform a judgement of the other, i.e., a judgement of fake for a sparse looking four fakes profile with only two friends may have informed a judgement of fake for a two fakes profile with ample information but only two friends. Thus, the participant condition was introduced to analyse whether this is the case by looking at whether participants have higher judgement accuracy scores when looking at only one type of fake profile.

Given this change in design the current study tested the same hypotheses from Study 1, with minor tweaks, to check for differences that may occur based on these changes. It is predicted that the real profiles will be judged more accurately than fake profiles (Hypothesis 1), and that participant condition will have an effect on accuracy scores, namely profiles with the highest number of fake characteristics will be more accurately judged as fake, than profiles with fewer, or no, fake characteristics (Hypothesis 2). Additionally, it is expected that participants' personality type will have an effect on their judgement accuracy (Hypothesis 3). Despite not accepting this hypothesis in Study 1, the different set of participants and slightly different study design warrant the inclusion of this hypothesis in this study, as an effect may be found. It is also expected that scores on the Social Sensitivity scale will be positively related to judgement accuracy (Hypothesis 4).

Regarding social media, it is expected that there will be a relationship between social media use and judgement accuracy, specifically the number of hours spent using social media per day and their familiarity with Facebook (Hypothesis 5). It is also expected that there will be a positive relationship between previous experience of creating a fake social media profile and judgement accuracy (Hypothesis 6). Finally, a relationship between self-reported accuracy and judgment accuracy is expected (Hypothesis 7). As human deception detection is only slightly higher than chance in a multitude of situations (Bond & DePaulo, 2006), it is difficult to predict the direction of this relationship.

Regarding the manipulated characteristics on the profiles, it is expected that, as in Study 1, there will be a relationship between the manipulated characteristics of the Facebook profiles and participants' judgements of the profiles (Hypothesis 8). Further, it is expected that the manipulated characteristics will also have an effect on the accuracy of participants' judgements of the profiles (Hypothesis 9). As this study introduced participant conditions in the design it is difficult to predict the direction of these relationships, hence the non-directional nature of these final hypotheses.

Method

Participants

An A-Priori power analysis was conducted using G*Power (Faul et al., 2007) to determine the appropriate sample size for the study. The analysis was based on a repeated measures between factors ANOVA design, with an expected medium effect size of f = 0.25, an alpha level of $\alpha = 0.05$, and a statistical power of $1-\beta = 0.80$. The results indicated that a total of 120 participants would be required to adequately detect expected effects whilst allowing for meaningful and robust comparisons across the experimental conditions.

A total of 120 participants completed the study online via Prolific through means of volunteer sampling. Participants were aged between 18 and 64 years, with a mean age of 26.58 years (*SD* = 8.52). Of these, 68 (56.7%) identified their gender as Male, and 52 (43.3%) identified as Female. Participant identified their ethnicities as; Asian or Asian British (*N* = 6, 5%), Black, African, Black British, or Caribbean (*N* =2 , 1.7%), Mixed or Multiple Ethnic Groups (*N* = 5, 4.2%), White, including any White backgrounds (*N* = 104, 86.7%), and Another Ethnic Group (*N* = 3, 2.5%).

Participants were asked to enter cultural details regarding the country they were born in, the country they currently live in, and how long they have lived in their current country. This was done to gain an understanding of participants' cultural backgrounds and beliefs. The locations of participants spanned six continents: Africa, Asia, Australasia, Europe, North America, and South America. The most popular locations of birth were Poland (*N* = 37, 30.8%), United Kingdom (*N* = 18, 15.0%), and Italy (*N* = 13, 10.8%). Countries where participants currently reside were Poland (*N* = 36, 30.0%), United Kingdom (*N* = 24, 20.0%), and Italy (*N* = 12, 10.0%). The majority of participants were residing in the same country they were born in, with only 14 participants (11.67%) moving location. Of these 14, nine (64.29%) have stayed within the same continent, and thus would be presumed from a cultural perspective as remaining within the same culture and sharing the same cultural beliefs as they would have done in their previous location.

Design

A 4 (*Real* profiles/ *0 Fakes* profiles/ *2 Fakes* profiles/ and *4 Fakes* profiles) x 2 (Accurate judgement vs. Non-accurate judgement) experimental Turing test design will be used within this study to investigate the hypotheses. The Dependent Variable (DV) is judgement accuracy with two levels - accurate and non-accurate - and is a within-subjects factor. The Independent variable (IV) is the participant condition - a between-subjects factor with 3 Conditions: Condition 1 – 6 *real* profiles & 6 *0 Fakes* profiles, Condition 2 – 6 *real* profiles & 6 *2 Fakes*, or Condition 3 – 6 *real* profiles & 6 *4 Fakes*.

Measures & Materials

Measures

The same self-report questionnaire scales, as used in Study 1, were used to measure social media activity, personality (Gosling, Rentfrow & Swann, 2003), Social Sensitivity (Riggio, 1986), and a follow-up questionnaire measuring participants' self-reported accuracy and previous experience in creating a fake profile.

Materials

A total of 74 profiles were used in the study. These comprised 68 fake profiles and six real profiles. The 68 fake profiles were created by manipulating up to seven characteristics that relate to: Type of photography (*Photo-Type*); Number of photographs (*Photo Number); Bio; Intro*; Content of posts (*Posts Content*); Number of comments (*Comments Number*); and Number of likes (*Likes Number*). These characteristics were identified by the heat maps as the most clicked areas across all profiles in Study 1.

Of the 68 fake profiles, 12 reflected '*0 fake* characteristics', 21 reflected '*2 fake* characteristics', and 35 reflected '*4 fake* characteristics'. The total number of each type of profile reflects the number of possible combinations of the seven fake characteristics manipulated within each profile. For example, there were 21 possible combinations of 2 fake characteristics, and 35 combinations of 4 fake characteristics (see Appendix P for breakdown of each different combination). The *0 Fakes* profiles were created with the main goal of creating a profile that no participant can identify as fake (i.e., a fake profile that is as convincing as possible as a real profile). Therefore, while the profile information is 'fake', it was designed to mirror the characteristics of a real profile and so no one characteristic was targeted for change. The design required six 0 fake profiles, which were randomly taken from a selection of 12 that were created for the study to allow for variability and randomisation. All profiles were created using the new Facebook layout that came into use in the time between the end of Study 1 and the commencement of this study (see Appendix Q for example of new layout).

For the purposes of validating the authenticity of the real profiles, the six participants who provided their Facebook profile were known to the researcher and were all recruited via E-Mail. Their ages ranged from 27-60 years (*M* = 41 years; *SD* = 14.71), all six participants identify as White British, and genders were split equally with three Females (50%), and three Males (50%).

Each participant had three different accuracy scores: one score for fake profiles (maximum score of 6), one score for real profiles (maximum score of 6), and a total accuracy score across all profiles (maximum score of 12).

Procedure

The procedure was the same as in Study 1. In brief, participants began by self-reporting their social media use, and completed the TIPI and Social Sensitivity personality scales online via Qualtrics. They were then provided with a set of instructions that they would see 12 profiles in a random order and be asked to make a judgement as to the authenticity of each profile. Additionally, participants were asked to indicate the areas of the profile they used when making their judgement by clicking on the areas of the profile screenshot. Participants are informed that six of these profiles are real, and six are fake. Different to the procedure of Study 1, prior to beginning the profile phase of the study, participants were randomly assigned into one of three conditions: Condition 1 (6 *real* profiles & 6 *0 Fakes*); Condition 2 (6 *real* profiles & 6 *2 Fakes*), or Condition 3 (6 r*eal* profiles & 6 *4 Fakes*). The random assignment of participants via Qualtrics resulted in the number of participants within each condition being Condition 1 - *N* = 40, Condition 2 - *N* = 41, and Condition 3 - *N* = 39.

After completion of the 12 profile judgements, participants were asked to; report how accurate they felt their judgements were using a 7-point Likert scale, declare whether they have previously created a fake profile, and if yes then why, and provide demographic details (e.g., age, gender, ethnicity, location/culture). Following this, participants were thanked for their time and fully debriefed. All participants were provided with further information regarding the purposes of the study and contact details of the researchers if they had any issues regarding their participation, or any further questions.

Ethics

This research was fully approved by Lancaster University's Faculty of Science and Technology ethics committee. All participant data were stored on a secure hard drive, in line with GDPR guidelines, and only accessible to the researcher.

Results

The raw data for this study were exported from Qualtrics and initially sorted in Microsoft Excel prior to conducting analyses in both IBM SPSS for Mac Version 27.0, and R Version 1.3.1073 (R Core Team, 2020). There were no incomplete entries, thus data from all 120 participants were used in analysis.

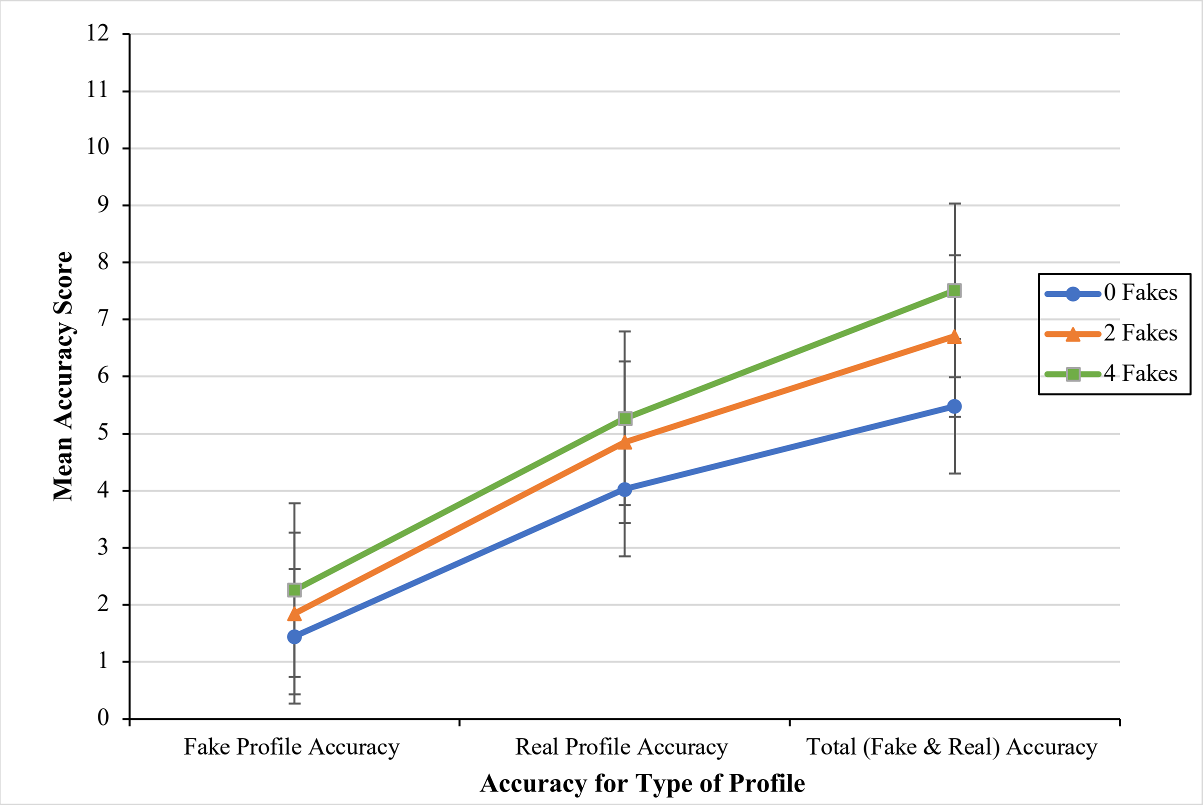
Prior to conducting the main analysis, the data was screened for missing data, non-normality, and outliers. The results showed that the data had only five outliers, none of which were extreme and therefore of any concern, and no significant negative or positive skew in the data.

Profile Accuracy

The results showed that overall, participants were more accurate at judging real Facebook profiles (*M* = 4.71, *SD* = 1.17) than fake Facebook profiles (*M* = 1.85, *SD* = 1.24), and that the overall total accuracy score (*M* = 6.56, *SD* = 1.65), was significantly greater than chance level of 6 (*t*(119) = 43.57, *p* < .001). These results provide support for H1 as real profiles were judged more accurately than fake profiles. Maximum judgement accuracy scores were only achieved by participants when judging real profiles; 15% (*N* = 6) correctly judged all 6 real profiles correctly in the *0 Fakes* condition, 26.8% (*N* = 11) did so in the *2 Fakes* condition, and over half (51.3%, *N* = 20) did so in the *4 Fakes* condition.

Mean accuracy scores for *fake* Facebook profiles, as shown in Figure 1, reveal a positive linear trend from *0 Fakes* (condition 1) (*M* = 1.45, *SD* = 1.43), *2 Fakes* (condition 2) (*M* = 1.85, *SD* = 1.01), and *4 Fakes* (condition 3) (*M* = 2.26, *SD* = 1.16). The same pattern was found for real profiles; mean accuracy scores in the *0 Fakes* condition (*M* = 4.03, *SD* = 1.27), *2 Fakes* condition (*M* = 4.85, *SD* = 0.94), and *4 Fakes* condition (*M* = 5.26, *SD* = 0.95). This trend provides support for H2 in that profiles with the highest number of fake characteristics were more accurately judged as fake, than profiles with fewer, or no, fake characteristics.

Figure 1
Mean judgement accuracy scores of 120 participants for each type of profile within each condition.



To further analyse the mean differences in judgement accuracy a 3x2 mixed design ANOVA (Analysis of Variance) was conducted where condition (*0 fakes, 2 fakes*, and *4 fakes*) was the between subject factor and fake vs. real profile judgements was the within subject factor. The data was assessed prior to conducting the test and all assumptions of the ANOVA (normal distribution, sphericity, no outliers of concern) were met. Results show that a significant main effect of accuracy was found, $F(1,117) = 311.77$, $p < .001$, $n^2 = .727$. Participants were significantly more accurate when judging real profiles ($M = 4.71$, $SD = 1.17$) compared to fake profiles ($M = 1.85$, $SD = 1.24$), There was also a significant interaction between participant condition and accuracy scores, $F(2,117) = 20.31$, $p < .001$, $n^2 = .26$. To understand this interaction further, Tukey's post hoc tests were conducted (Table 1).

Table 1.
Tukey's post-hoc comparisons of mean accuracy scores between each participant condition

Measures	*M*	*SE*	*95% CI*
0 Fakes & Real vs. 2 Fakes & Real	-1.23*	0.32	[-1.98, -0.48]
2 Fakes & Real vs. 4 Fakes & Real	-0.81*	0.32	[-1.57, -0.04]
4 Fakes & Real vs. 0 Fakes & Real	2.04*	0.33	[1.28, 2.80]

Note. ^{a}N = 40. ^{b}N = 41. ^{c}N = 39. * *p* = .05

Results displayed in Table 1 show that a significant mean *increase* in judgement accuracy scores was found between; participants in the *0 Fakes* and *2 Fakes* conditions (*M* = 1.23, *SD* = 0.32), participants in the *2 Fakes* and *4 Fakes* conditions (*M* = 0.81, *SD* = 0.32), and participants in the *4 Fakes* and *0 Fakes* conditions (*M* = 2.04, *SD* = 0.33), all of which are statistically significant mean differences at *p* = .05 level. These results further support the prediction that real profiles will be judged more accurately than fake profiles (H1), and that participant condition will have an effect on accuracy scores, namely profiles with the highest number of fake characteristics will be more accurately judged as fake, than profiles with fewer, or no, fake characteristics (H2). Thus, H1 and H2 can be accepted.

To further analyse participants' judgement accuracy across both fake and real profiles, and how participants make decisions under uncertainty, Signal Detection Theory (SDT) was used (Green & Swets, 1966). SDT proposes that decisions are made under conditions of uncertainty, and the decision-maker must decipher between the signal (stimulus), and the background noise (random variables) when making their decision. To test this, the probability that the participant says 'yes' when the stimulus is present and the probability that the participant says 'yes' when the stimulus is *not* present are measured. These probabilities are known as the 'Hit Rate' and 'False Alarm' rate, respectively.

In regard to the decision-making process of judging the authenticity of the Facebook profiles, fake profile accuracy and real profile accuracy scores were first transformed into hit rate and false alarm rate scores, or in other words the levels of signal and noise within the study. From this, several calculations were completed to obtain d-prime (*d'*) values and criterion (*c*) scores. *d*' is a measure of sensitivity used to indicate participants' ability to discriminate between the signals (fake profiles) and the noise (real profiles) within the study, and *c* is a measure of participants' response bias (i.e., where participants biased more towards answering yes or no, or in this case, fake or real, when judging the profiles?)

Results indicate that participants' ability to distinguish signals (fake profiles) from noise (real profiles) was greater than zero (*d*' = 0.17, 95% CI [1.62, 2.08]), meaning they were able to identify fake profiles as fake. Additionally, participants showed a bias to responding 'yes' (*c* = -0.48), meaning they were biased to judging the profiles as fake. However, judgement accuracy was higher for real profiles when compared to fake profiles, which, alongside the *d'* value and the *c* score, suggests that the fake profiles were deceptive enough to fool participants in to judging them as real: participants were statistically able to distinguish between the fake and real profiles *and* had a response bias towards judging the profiles to be fake, yet they still judged the majority of fake profiles as real.

Self-reported Accuracy

After all profile judgements were made, participants were asked to rate how confident they feel in their judgements on a Likert scale from 1 (Unconfident) – 7 (Confident), with 'Neutral' in the middle. Of the 40 participants in the 0 Fakes condition, the most frequently selected choice was 'Slightly Confident' (*N* = 14, 35.0%), with only 1 participant (2.5%) selecting 'Confident'. For those in the 2 Fakes condition (*N* = *41*), the most frequently selected choice was also ‘Slightly Confident’ (*N* = 13, 31.7%), with 0 participants selecting ‘Confident’. Finally, in the 4 Fakes condition (*N* = *39),* the most frequently selected choice was ‘Slightly Unconfident’ *(N* = 11, 28.2%), with only 7 (17.9%) reporting feeling ‘Slightly Confident’, and 2 participants (5.1%) selecting ‘Confident’. In summary, participants in the 0 Fakes were the most confident in their judgements, followed by participants in the 2 Fakes condition with those in 4 Fakes condition feeling the least confident.

To understand whether certain levels of self -reported accuracy have a relationship with accuracy scores a multiple regression was conducted for each participant condition, using ‘Confident’ as the constant. All assumptions of regression analysis (linearity, homoscedasticity, independence of residuals) were met. Table 2 displays the results.

Table 2.
Multiple regression of self-reported accuracy and total judgement accuracy for each participant condition

	0 Fakes Condition [a]			2 Fakes Condition [b]			4 Fakes Condition [c]		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Self-Reported Accuracy	.076			.207			.118		
Constant		6.00	1.55		6.77	0.33		8.50	1.03
Unconfident		-1.20	1.70		-1.52*	0.68		-0.75	1.26
Moderately Unconfident		-0.25	1.65		0.23	0.76		-2.70*	1.22
Slightly Unconfident		0.25	1.74		0.41	0.49		-0.50	1.12
Neutral		-1.00	1.74		-7.70	0.76		-0.17	1.33
Slightly Confident		-0.57	1.61		0.08	0.54		-1.50	1.17
Moderately Confident		-0.50	1.74		0.09	0.56		-0.79	1.17

Note. [a] N = 40, *df* = 6, 39, [b] N = 41, *df* = 6, 40, [c] N = 39, *df* = 6, 38. * *p* = .05

Table 2 shows that of the three models there are only two significant coefficients; ‘Unconfident’ is a significant predictor of judgement accuracy in the *2*

Fakes condition (*B* = -1.52), and 'Moderately Unconfident' is a significant predictor of accuracy in the *4 Fakes* condition (*B* = -2.70). Both of these coefficients indicate that self-reporting as feeling either 'Unconfident' or 'Moderately Unconfident', respective to the conditions, is associated with a *decrease* in judgement accuracy scores. However, all three models were non-significant: *0 Fakes*, (*F*(6, 39) = 1.08, *p* = .840); 2 Fakes, (*F*(6, 40) = 1.83, *p* = .133); 4 Fakes, (*F*(6, 38) = 1.85, *p* = .121). As such, H7 that expected a relationship between self-reported accuracy and judgment accuracy, can only be partially accepted.

Personality

To test for effects of personality (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness to new Experiences, and social sensitivity) on accuracy, a multiple regression analysis was conducted using the personality traits from the TIPI and SS Scale as the predictors; Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness to New Experiences, and Social Sensitivity score. It is expected that participants' personality type (H3) and Social Sensitivity scores (H4) will have an effect on their judgement accuracy.

All assumptions of the multiple regression (linearity, homoscedasticity, independence of residuals) were tested prior to analysis, and all were met. Table 3 presents the results of the regression.

Table 3.
Multiple regression for personality predictors of overall judgement accuracy for each participant condition.

Predictors	0 Fakes Condition [a]			2 Fakes Condition [b]			4 Fakes Condition [c]		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
TIPI	.134			.105			.143		
Extraversion		0.08	0.19		0.14	0.16		0.32	0.18
Agreeableness		-0.34	0.22		0.16	0.19		-0.29	0.30
Conscientiousness		-0.16	0.19		0.06	0.18		0.03	0.28
Emotional Stability		-0.47*	0.21		-0.03	0.15		-0.06	0.28
Openness to New Experiences		-0.05	0.28		-0.09	0.22		-0.33	0.25
SS Scale		-0.04	0.02		-0.03	0.03		0.01	0.04

Note. [a] N = 40, *df* = 6, 39, [b] N = 41, *df* = 6, 40, [c] N = 39, *df* = 6, 38. * *p* = .05

Table 3 shows that the only significant personality predictor of participants' judgement accuracy was 'Emotional Stability' in the *0 Fakes* condition, where an increase in Emotional Stability is associated with a decrease in accuracy scores (*B* = -0.47). However, the overall *0 Fakes* condition model was not statistically significant: *F*(6, 39) = 2.01, *p* = .093. Additionally, neither model for the other conditions were statistically significant: *2 Fakes* , *F*(6, 40) = 0.67, *p* = .678; 4 Fakes, *F*(6, 38) = 0.89, *p* = .515. As such, it can be concluded that there is no significant relationship between personality variables and participants' accuracy scores in any of the three conditions, thus H3 and H4 cannot be accepted.

Social Media

Participants were asked a series of questions in relation to their use of social media to assess whether their usage or previous experience had an effect on their judgement accuracy. It is expected that time spent on social media per day will have an effect on judgement accuracy (H5), and there will be a positive relationship between previous experience creating a fake profile and judgement accuracy (H6).

Platforms

Participants were asked to select which social media platforms they use from a list of seven options: Facebook, Twitter, Instagram, Snapchat, TikTok, YouTube, and Other. After selections had been made, participants were then asked to rank the platforms they had previously selected in order of use, from used most often to least often. All participants selected at least one social media platform that they use regularly, but nine participants did not rank their selection, meaning only 111 participants both selected and ranked the social media platforms they use regularly. Of these 111 participants, only one participant selected and ranked all seven options. Table 5 outlines the social media platforms and the frequency of their individual rankings.

Table 5.
Participant rankings of social media platforms, based on the platforms they use most often (1st ranking), to the platforms they use the least (7th ranking).

Social Media Platform	Ranking Order							Total number of participants who use each platform
	1st	2nd	3rd	4th	5th	6th	7th	
Facebook	29	23	20	11	6	3	0	92
Twitter	6	15	14	14	4	1	1	55
Instagram	26	28	27	6	0	1	0	88
Snapchat	1	0	6	17	5	2	0	31
TikTok	4	11	7	4	7	0	0	33
YouTube	40	29	17	10	5	1	0	102
Other	5	3	6	2	1	2	0	19
Total number of rankings made	111	109	97	64	28	10	1	

It is clear from Table 5 that the most popular social media platform amongst participants is YouTube ($N = 102$), with 40 participants (39.2%) ranking it as their most used platform. Second to YouTube, is Facebook with 92 participants selecting it as a social media platform they use regularly, and 29 (31.5%) of those participants ranking it as the platform they use most often.

When selecting the option Other, participants had the opportunity to enter the other platforms they use. Such entries included Reddit, Pinterest, Discord, WhatsApp, LinkedIn, and Twitch. Of these Other entries, Reddit was the most popular platform with 10 entries out of the total 19.

Overall, 92 participants within this study are familiar with Facebook as a platform, and thus have an understanding of the way the platform works and as a result are exposed to Facebook profiles regularly.

Purposes

Participants were asked to indicate, from a list of 12 options, the specific purposes they use social media for. Of these 12 options, the most popular purpose was 'Watching videos (TV/Films/YouTube etc.)' which was selected by the overwhelming majority of participants ($N = 110$). Second, 'Socialising with friends/keeping in touch' was selected by 97 participants, followed by 'News (keeping up with current events)' which was selected by 78 participants. Interestingly, the option of 'Making friends/meeting new people', which could be considered as one of the main purposes of *social* media, was only selected by 24 participants.

Participants were provided with the option to select Other and record their specific purpose for using social media. Only 10 participants selected this option, inputting purposes such as "Learning", "browsing memes", "Look for funny content to share with my friends or partner", "Information purposes (Reddit)", "Communication", "General enterteinment [sic]", "…boredom scrolling with no real intention", "spying friends [sic]", "Have information about our studies etc.", and "IT Help groups…". With the exception of the comments in relation to learning, the remaining comments could be included under the umbrella of some the other 11 given options (See appendix A for full list of options).

Daily Usage

Of the 120 participants, only 2 (1.7%) reported that they were not regular users of social media. The majority of participants reported using social media for more than 1 hour per day (91.7%), with most using it for up to 3 hours daily ('2-3 hours per day' [*N* = 35, 29.2%], and '1-2 hours per day' [*N* = 34, 28.3%]). To investigate further any effects time spent on social media may have on judgement accuracy, multiple regressions were conducted using total accuracy scores for each condition. The predictor used as the constant within the model was 'Less than 1 hour'. Results are presented in Table 4.

Table 4.
Multiple regression for social media time predictors of overall judgement accuracy for each participant condition

Predictors	0 Fakes Condition [a]			2 Fakes Condition [b]			4 Fakes Condition [c]		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Hours spent on social media per day	.015			.106			.061		
Constant		5.60	0.70		6.00	0.72		7.50	1.07
1-2 Hours		-0.33	0.84		0.36	0.81		-0.25	1.15
2-3 Hours		-0.10	0.85		0.82	0.78		0.88	1.19
3-4 Hours		-0.27	0.94		2.00	1.14		0.67	1.23
4+ Hours		-0.15	0.89		0.88	0.84		-0.68	1.16

Note. [a] N = 40, *df* = 4, 39, [b] N = 41, *df* = 4, 40, [c] N = 39, *df* = 4, 38.

It is evident from Table 4 that there are no significant effects of time spent on social media and judgement accuracy, due to the lack of statistically significant coefficients. Additionally, each model was not statistically significant: 0 Fakes (*F*(4, 39) = 0.13, *p* = .970); 2 Fakes (*F*(4, 40) = 1.06, *p* = .389); 4 Fakes (*F*(4, 38) = 1.62, *p* = .193).

Previous Experience in Creating a Fake Profile

Nineteen participants (15.8%) indicated that they had previously created a fake social media profile, of which 11 were Males (57.9%), and 8 were Females (42.1%). The reason for the fake profiles included harmless ends such as self-amusement ("…just for a laugh", "For joke purposes…", "…a fake Harry Styles page"), gaming purposes ("…to roleplay with other people from the same community", "…to play Facebook games", "…an account for gaming purposes"), or anonymity / privacy purposes ("…to hide personal information", "…didn't want to share my personal info [sic]"). Some reported more malicious intent including investigative purposes ("…to know if my friends were hiding something from me", "…spying", "…check the stories of my ex without him knowing", "to stalk my friend…"), deceptive purposes ("To meet girls away from my partner", "…there were things I didn't want them to see about me", "to follow accounts that are sex based…if I followed them from my real profile that would cause problems"), and malicious purposes ("to convince someone to do something", "…to make fun of friends", "stalk my friend, for fun", "…to make fun of people who thought the account was real").

To test for a relationship between experience creating a fake profile and judgement accuracy, correlations were conducted. As the DV is continuous (accuracy scores) and the IV is categorical (Yes/No answers), a Spearman's rank correlation test was used. Each of the participant conditions were analysed separately. For those in the

0 Fakes group there was a significant medium positive correlation found between previous experience in creating a fake profile and judgement accuracy scores (r_s (40) = .418, p = .007). No significant correlations were found for those in the *2 Fakes* (r_s (41) = .199, p = .212) and *4 Fakes* (r_s (39) = -.305, p = .059) conditions.

To further investigate effects of previous experience creating a fake profile and judgement accuracy, a multiple regression was conducted to assess whether such experience can predict accuracy scores. Participants' total accuracy scores were used for each condition, with 'No' used as the constant in the models. Table 6 displays the results.

Table 6.
Multiple regression of previous experience creating a fake profile and total judgement accuracy for participant condition.

Predictors	0 Fakes Condition [a]			2 Fakes Condition [b]			4 Fakes Condition [c]		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Previous experience creating fake profile	.151*			.044			.109*		
Constant		5.24	0.24		6.60	0.21		7.75	0.26
Yes		1.60*	0.61		0.73	0.55		-1.32*	0.62

Note. [a] N = 40, *df* = 1, 39, [b] N = 41, *df* = 1, 40, [c] N = 39, *df* = 1, 38. * p = .05

Table 6 shows that for those in the 0 Fakes condition, previous experience creating a fake profile is associated with an increase in judgement accuracy scores (B = 1.60), and the model itself can explain 15.1% of the variance in judgement accuracy scores (F(1, 39) = 6.78, p = .013). A significant relationship was also found for those in the 4 Fakes condition, however in this case previous experience creating a fake profile is associated with a *decrease* in judgement accuracy scores (B = -1.32), with the model explaining 10.9% of the variance in accuracy scores (F(1, 38) = 4.54, p = .040).

However, a significant effect was not found in the 2 Fakes condition (*F*(1, 40) = 1.80, *p* = .188).

The analyses of the social media variables presented above show that no relationship was found between times spent on social media per day and judgement accuracy in any condition, and that a relationship between previous experience in creating a fake profile and judgement accuracy was found for two of the three conditions. As such, H5 cannot be accepted and H6 can be partially accepted.

Manipulated Characteristics of Profiles

To analyse if the manipulated characteristics of the profiles had an effect on both participants' judgements of the profiles and their accuracy of said judgements, several general linear mixed effects regressions (*glmer*) were run in R using the 'lme4' package (Bates, Machler, Bolker & Walker, 2015). The manipulated factors were entered as predictors and 'Prolific ID' and 'Profile Number' were random effects. For both the judgement and accuracy outcome measures, the addition of 'Profile Number' as a second random effect significantly improved model fit (*p* < .001). As such, it was retained in the model to reflect each individual profile used within the study, which is different to Prolific ID, which reflects the individual participants. Table 7 shows the results of Model 1 in which judgement (real or fake) is the outcome measure, and Model 2 in which accuracy (accurate or non-accurate judgement) is the outcome measure.

Table 7.
Results from 'glmer' Model's 1 & 2 where judgement and accuracy are regressed on the manipulated profile characteristics.

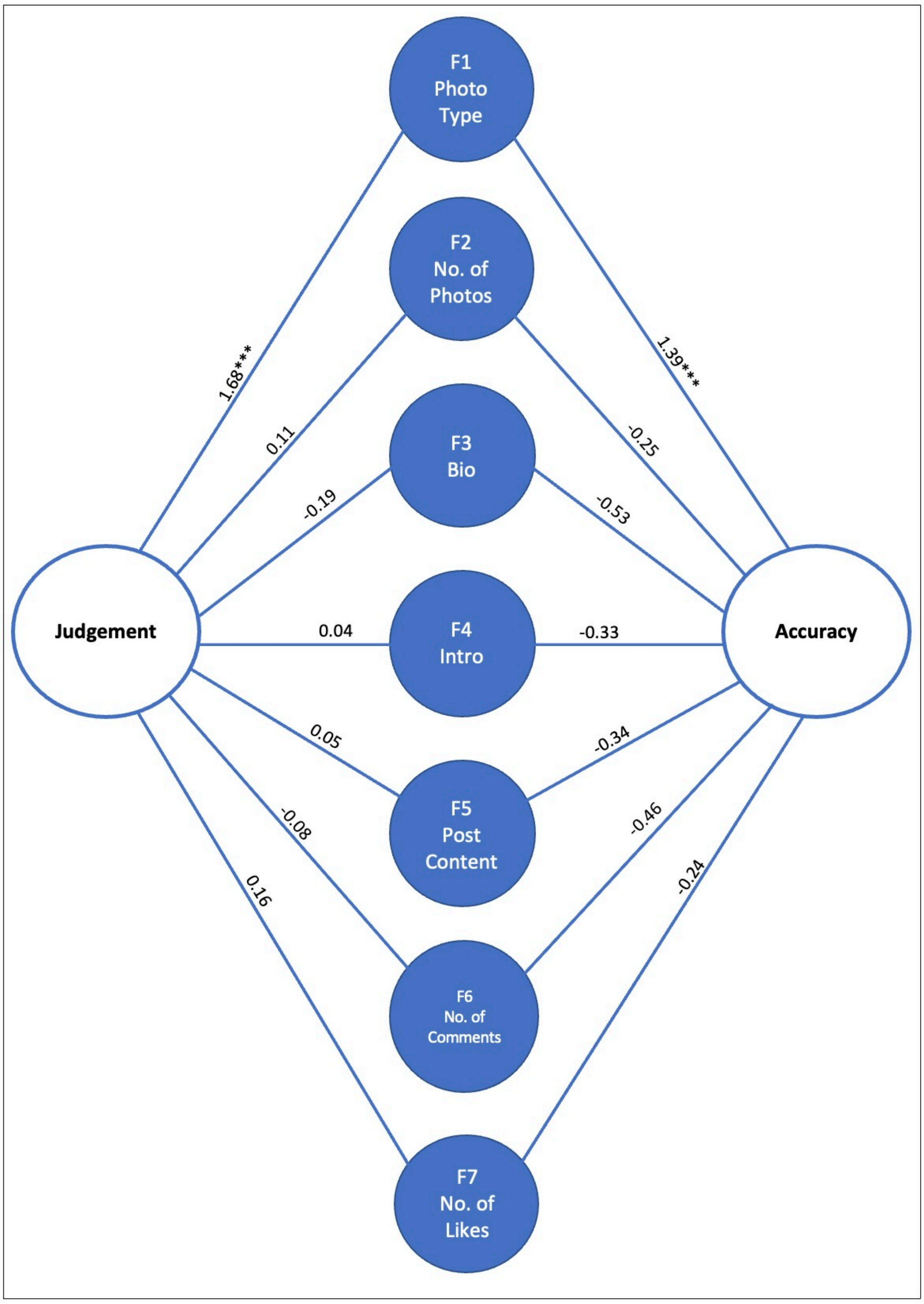
Predictors	Model 1 – Judgement [b]					Model 2 – Accuracy [c]				
	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*	
Fixed effects										
Intercept	-1.51	0.20	0.15	0.32	<.001***	-0.44	0.26	0.39	1.07	.088
Photo Type [a]	1.68	0.27	3.18	9.13	<.001***	1.39	0.34	2.05	7.92	< .001***
Number of Photos [a]	0.11	0.27	0.66	1.87	.689	-0.25	0.35	0.39	1.53	.462
Bio [a]	-0.19	0.27	0.48	1.41	.484	-0.53	0.35	0.30	1.17	.129
Intro [a]	0.04	0.27	0.61	1.78	.876	-0.33	0.35	0.36	1.43	.349
Post Content [a]	0.05	0.27	0.62	1.78	.848	-0.34	0.35	0.36	1.40	.321
Number of Comments [a]	-0.08	0.27	0.54	1.57	.769	-0.46	0.35	0.32	1.26	.193
Number of Likes [a]	0.16	0.27	0.69	1.99	.555	-0.24	0.35	0.40	1.56	.492
Random effects										
Residual Variance (σ^2)		3.29				3.29				
τ_{00} PROLIFICID		0.41				.04				
τ_{00} PROFILENUM		0.43				1.16				
Intraclass Correlation Coefficient		.20				.27				

Note. Number of Participants = 120, Number of Profiles = 74, Number of Observations = 1440. *p = .05, ** p = .01, *** p<.001.
[a] Model 1: 0 = Judgement of Real, 1 = Judgement of Fake; Model 2: 0 = Non-Accurate Judgement, 1 = Accurate Judgement. [b] Conditional R^2 = .081. [c] Conditional R^2 = .055.

Table 7 shows that the only significant predictor of both participants' judgement and accuracy is *Photo Type*. If *Photo Type* had been manipulated in the profile being judged, then participants are more likely to judge that profile as fake (B = 1.68) and said judgement of fake is likely to be accurate (B = 1.39), suggesting that participants over-rely on the visual aspects of the profiles. As an effect of manipulated characteristics has been found for both participant judgement and participant accuracy, hypotheses H8 and H9 can be accepted.

Figure 2 presents a lens model approach comparison based on Brunswik's lens model (as defined in Study 1), between Model's 1 and 2 with the estimates and statistical significance for each of the factors shown.

Figure 2.
Lens model diagram showing estimates and statistical significance of the manipulated characteristics for models 1 and 2.

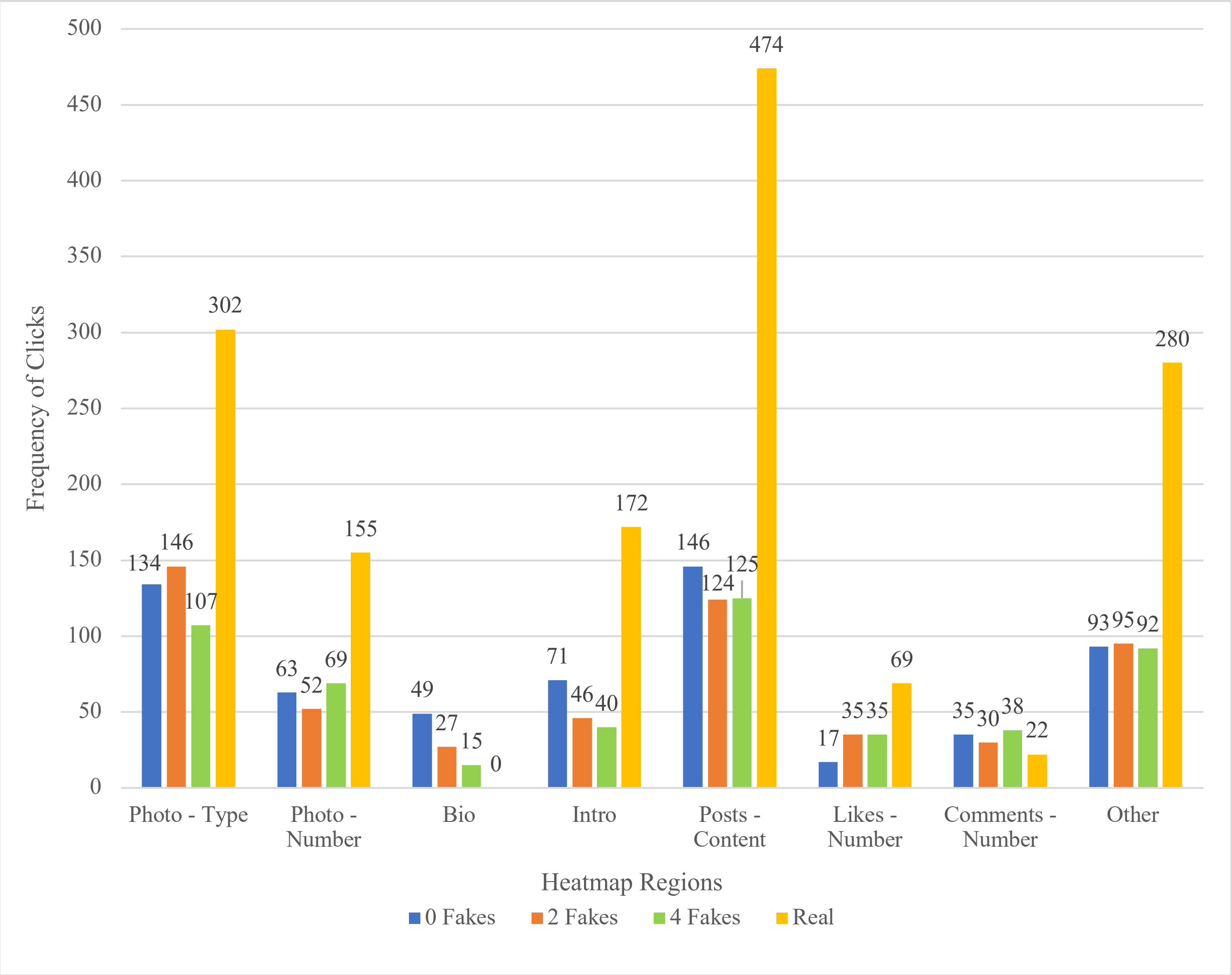


Post-Hoc Analysis

Heatmaps

Each profile within the study had a heatmap layer, meaning participants were able to click on the specific areas of the profile they used when making their judgement of the authenticity of the profile. Each participant was allowed a maximum of ten clicks per profile. The overall frequency of these clicks for all profile types are shown in Figure 3.

Figure 3.
Number of clicks within each heatmap region for all types of profiles



As evidenced in Figure 3, the manipulated profile characteristics relied upon most when making authenticity judgements across all three types of fake profiles and real profiles are *Photo Type* and *Posts Content*, whereas N*umber of Comments, Number* *of Likes,* and *Bio* were relied upon the least. Interestingly in regard to real profiles, *Bio* was not relied upon at all when making authenticity judgements of real profiles. This could be due to the fact that the real profiles were not manipulated, and as such none of them had information in the *Bio* section, thus meaning participants' attention was not drawn to that area.

Across all four types of profile, participants clicked on *Other* areas to inform authenticity judgements, and these were the third most popular region in all types of profiles. After manually looking through each heatmap image, it is evident that the majority of clicks under the term *Other* are general inaccurate clicks, i.e., they are just outside of the regions set and/or in the white spaces around posts, or they are clicks in the grey space area of the profiles, i.e., where there is no available profile information present. Very few clicks under the region of *Other* were in areas of the profile that contained information. Examples of which are clicks on; content of comments, likes on comments, replies on comments, friends, 'add friend' button, date of posts, 'see all photos' button, check-in locations, and the profile name.

To further understand specific areas participants use to inform their judgements, additional heatmap analyses were conducted whereby the number of clicks per area was split up by each type of judgement. Figure 4 displays these results.

Figure 4.

Graph showing the frequency of clicks for each judgement type per manipulated characteristics for all profile types.

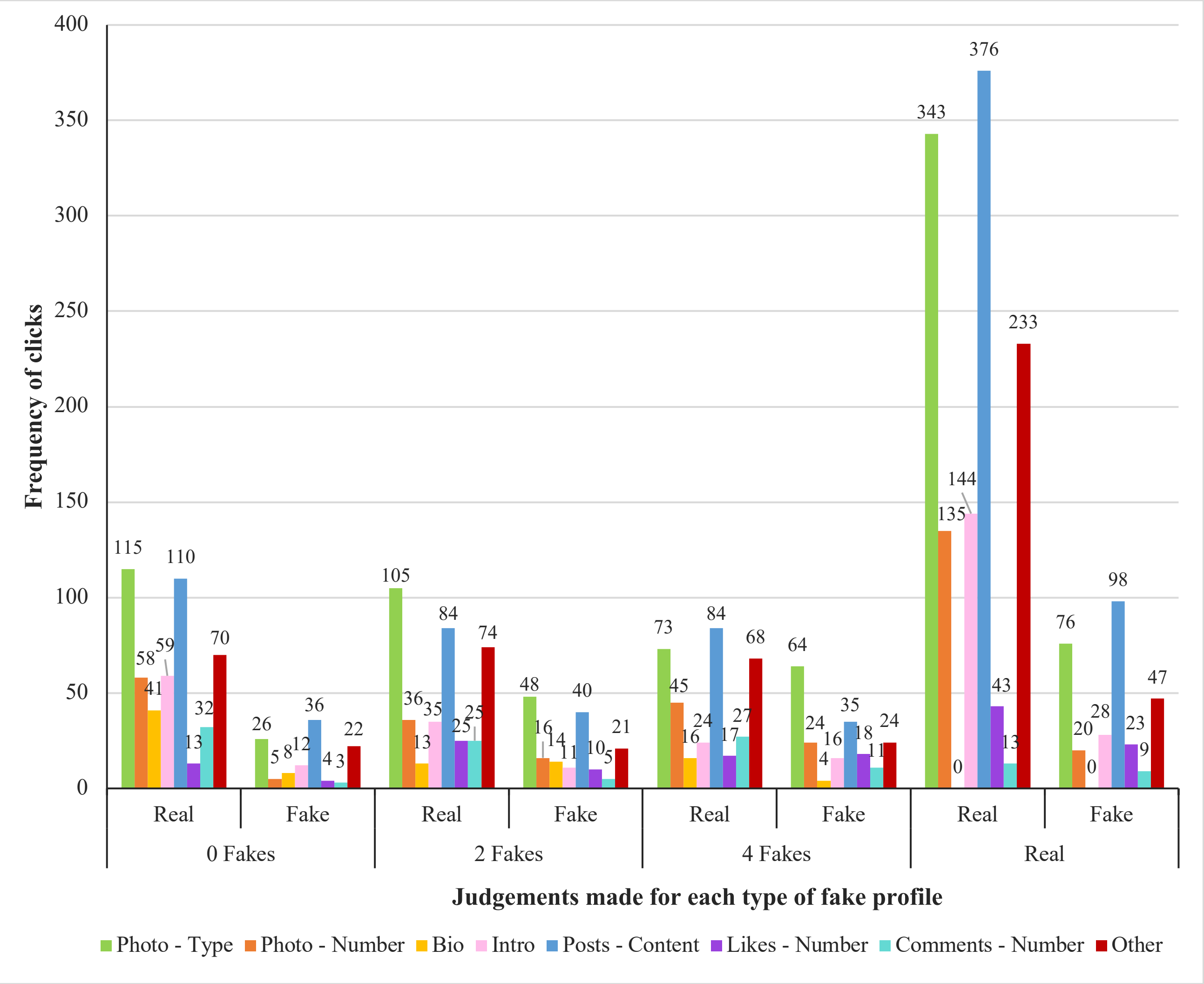
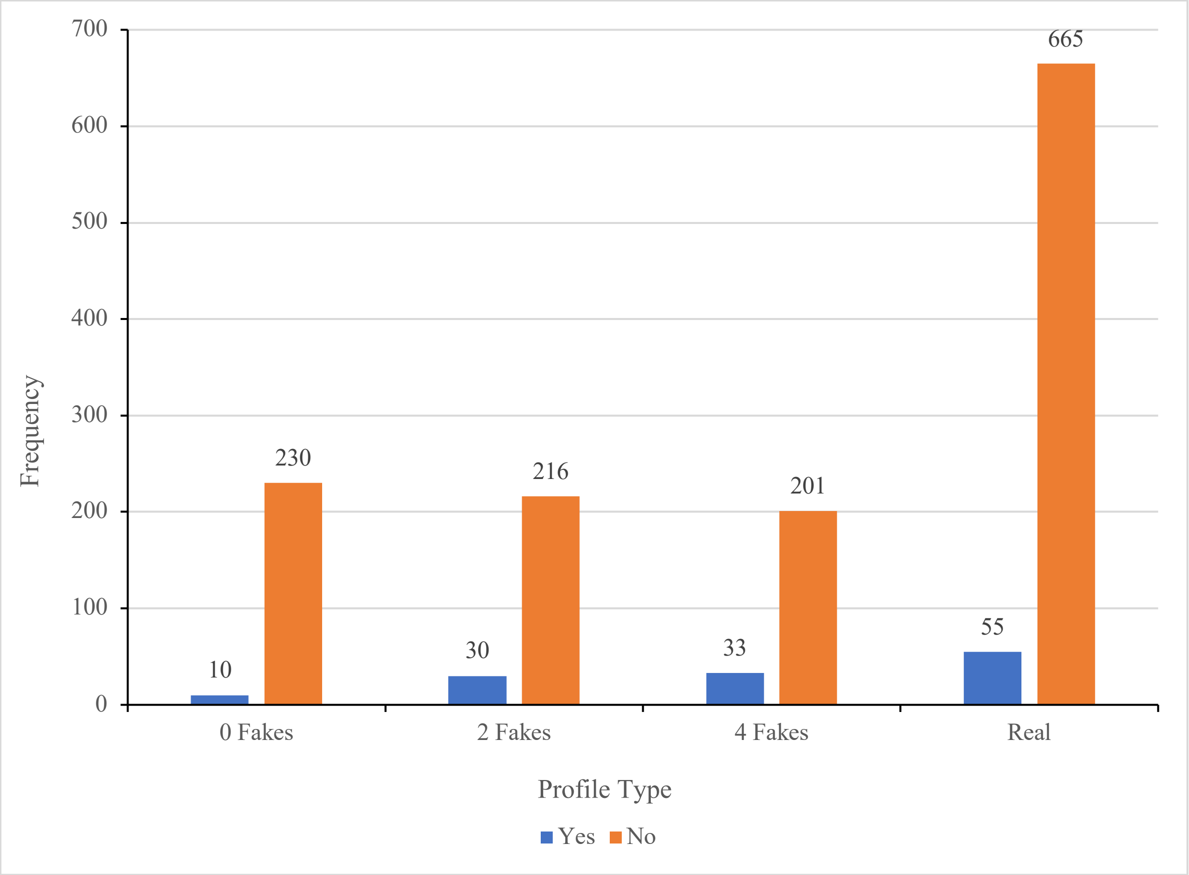


Figure 4 shows that overall, across all profile types, participants had a higher frequency of clicks when they were judging the profile as real. The disproportionate frequency of clicks for real profiles is related to participant condition – all participants (N = 120) judged real profiles whereas the fake profiles were split across all 120 participants in the conditions. With that being said the pattern described is still evident here – participants click more so on the real profiles when they are judging the profiles as real.

Deceptive Purposes of Profiles

After viewing and judging each profile, participants were asked to indicate whether they thought the profile was created with deceptive purposes or malicious intent. Figure 5 displays the frequencies of yes/no answers for each type of profile made in regard to the deceptive purposes of the profiles across all profile types.

Figure 5
Graph showing the frequency of yes/no answers to the question 'Do you think this profile was created for deceptive purposes?' for all profile types



Across all types of fake profiles, participants were of the view that the profiles were *not* created for deceptive purposes. Comparing the answers across conditions in regard to the fake profiles, Figure 5 shows that participants believed the profiles with 4 Fakes were more deceptive than those with *0 Fakes* or *2 Fakes*. In regard to the real profiles, a similar result was found whereby the vast majority of participants believed that the real profiles were *not* created for deceptive purposes (*N* = 665), and few believed that the profiles were made for deceptive purposes (*N* = 55). When asked for reasons as to why they think the profile was deceptive, participants reported: "catfishing to simply petty grudges by creating duplicate profile of someone else", "…to spread dangerous propaganda to those who might believe it", "posts seem to be inconsistent", "it has pictures of celebrities", to list but a few.

Discussion

Findings showed that overall, participants were more accurate at correctly identifying and judging real Facebook profiles as real, and least accurate at correctly identifying and judging fake Facebook profiles as fake. Accuracy of participant judgements for the fake profiles showed a linear trend, similar to that of Study 1 (Chapter 2), with accuracy being highest for fake profiles with four fake characteristics, lower for fake profiles with two fake characteristics, and lowest for fake profiles with zero fake characteristics.

The profiles with zero fake characteristics were created to mimic a realistic profile to 'fool' participants into thinking the profiles are real. This process involved zero direct manipulation of the seven different characteristics used within the other fake profiles, but rather the inclusion of all of these characteristics in a realistic way, by replicating a 'typical' active Facebook profile as closely as possible. The results reflect what was expected: the profiles with zero fake characteristics were the profiles with the highest number of inaccurate judgements (i.e., the majority of the participants in that condition judged the fake profiles to be real). This effect is most likely due to the fact these profiles looked most like an average real Facebook profile: the researcher aimed to create a profile that was a 'real fake', in that it was a fake profile that looked so convincingly real that participants could not detect that any elements were fake, thus participants always made the judgement that the profile was real. The poor ability of participants to detect these as fake suggests that the researcher did achieve this to some extent, and further supports findings from the deception detection literature that humans are not good at accurately detecting deception.

Further support for this notion is shown in the deceptive intentions data: the *0 Fakes* profiles had the lowest frequency of “Yes” answers to whether the profile was created with deceptive intentions. This result can be understood when considering that the majority of participants judged these profiles as real, and reinforces the suggestion made previously that perhaps these profiles are the hardest to accurately decipher as fake. It is therefore difficult for participants to make an informed judgement as to the deceptive nature of the profile when the profiles do not look ‘obviously fake’ or do not contain enough identifiable cues to base their judgement on. This could potentially be explained by Funder’s (1995) Realistic Accuracy Model (RAM), outlined fully in Study 1 (Chapter 2) which details the process of making accurate judgements of a person’s personality using cues in the environment. In relation to this study, this RAM model suggests that the limited amount of available information, or cues, in terms of fake manipulations on the *0 Fakes* profiles limited participants ability to detect any of the available cues therefore reducing the likelihood of an accurate judgement.

As of yet the RAM model has not been used in regard to accuracy of authenticity judgements of online profiles, however it has been shown to be applicable in online settings and social media profiles. Darbyshire et al. (2016) used the RAM model to test whether people can detect available cues and use these to form accurate judgements regarding personality traits on Facebook profiles. Researchers identified six different cue themes that participants used when forming their judgements; vocabulary, photographs, online interactions, occupational status, health status, and relationships with others, and concluded that judgements of personality through the context of Facebook do result in a degree of judgement accuracy. These findings show that the principles of RAM can be applied in a social media context and can help to explain the linear trend in accuracy based on the availability of cues on each of the different types of profiles.

Further results from this study show that the profiles with four fake characteristics were most accurately judged by participants. This could be because these profiles had less content on them overall due to the higher number of manipulations, for example one profile may have the following manipulations: one post, one photo, three friends, and no intro information, and so could be considered as the profiles that look the most 'obviously fake' and therefore identified accurately as fake. Additionally, these profiles had the highest number of cues available to the participants, a finding that can provide partial support for Funder's (1999) finding mentioned above, that detection of cues is reliant upon there being a large number of cues available to the judge. Further evidence for this is apparent when looking at the results of the profiles with two fake characteristics. These profiles were not the most accurately judged, nor were they the least, which suggests that they were more difficult to identify as fake than the profiles with four fake characteristics as they had less 'obvious' fake elements but were easier to identify as fake than the profiles with zero characteristics as they did have some 'obviously' fake elements.

A further point of interest garnered from the results of the study is that maximum judgement accuracy scores were only achieved by participants when judging real profiles. This linear pattern is similar to that of the mean accuracy scores discussed above, with the number of participants achieving 100% accuracy highest for the profiles with four fake characteristics. This could be explained by the participant condition, as the ability to detect real profiles increased as the number of fake characteristics increased, suggesting that perhaps the contrast between the two types of profile in the *4 Fakes* condition (*4 Fakes* and *Real*) was larger than in other condition, meaning participants were more able to accurately distinguish between the two profile types.

In regard to the specific areas of the profile participants used to make their judgements, the results showed that participants have a tendency to rely on the visual aspects of the profile, specifically the photographs/images, a finding which is also found when looking at the areas clicked on in the heatmaps. This reinforces the notion that participants rely on the visual aspects of the profile more in comparison to the more informative and person specific aspects of the profiles, such as *Intro* or *Bio*. One possible explanation for this is that images provide the judge with more information. For example, Lindsay et al. (2004) state that photographs are examples of a rich source of information and are perceived by people as evidence that the events in said photographs actually happened as depicted, and Darbyshire et al. (2016) found that participants widely reported they relied on the photographs of Facebook profiles when judging the personality of the profile user. Additionally, researchers Ivcevic and Ambady (2012) found that when participants were asked to evaluate the personality of an unknown Facebook user through their profile, their impressions were based on the users' profile pictures.

A further explanation draws from literature regarding cognitive processing speed. Whilst language processing is a quick process, where it has been found that 7.66 words can be read per second (Dyson & Haselgrove, 2000), it is widely reported that images are processed at a faster rate than text. In a recent ground-breaking addition to the processing literature, Potter, Wyble, Hagmann, & McCourt (2014) found that conceptual understanding of an image can occur in as little as 13ms. Whilst there is a lack of literature in relation to the processing speeds of images and text on social media, eye-tracking studies have been used within this context to analyse the areas of social media profiles that are fixated on. Nielson (2006) found that people typically fixate on the upper left-corner of the screen and then proceed to scan the rest of the page in an *F-shaped* pattern. This could provide an explanation as to why *Photo Type* was relied upon most, as the profile picture is in the upper left-hand corner and the cover photo is along the top of the screen. Both of these images were included under the heatmap region of *Photo Type*, and even with slight Facebook layout changes where the profile picture moves from the left-hand side to the middle of the page (as in Study 2 profiles), these images would still both be included under this *F-shaped* viewing pattern.

However, several researchers have found that this pattern only relates to text only web pages (Shrestha, Lenz, Chaparro, & Owens, 2007; Sutcliffe & Namoun, 2012). When images are introduced to the webpage, eye gaze and fixation changes. Beymer, Orton, and Russell (2007) found that when images were experimentally manipulated to appear on the right-hand side of the page, viewers fixated their attention to the right-hand side. In regard to Facebook, Scott and Hand (2016) found an *L-shaped* pattern when participants were viewing Facebook profiles with a professional motivation (e.g., looking at potential employees), and a *Z-shaped* pattern when viewing with a social motivation (e.g., looking at friends). Both of these shapes begin in the upper left-hand corner where the profile picture or cover photo are (dependent on Facebook layout), so in relation to this study viewers fixate first on the *Photo Type* area of the profile. However, when looking specifically at the regions of the Facebook profile that viewers fixated on for the longest periods of time, area that were the largest or most visually complex (i.e., posts, likes on posts) received the highest number of fixations. Profile pictures and the 'Info' section had a moderate number of fixations, and the name of the profile owner had the lowest number of fixations. These findings suggest that whilst viewers use a generic *L-* or *Z-shaped* pattern when viewing Facebook profiles, their fixations are relative to the context of the profile. In relation to this study, this suggests that *Post Content* is perhaps the area viewers fixate on, more so than *Photo Type*, and use this to form their judgement. This does not line up with the heatmap click frequencies reported as *Photo Type* is the highest across all profile types, suggesting that participants may fixate more on the post content when viewing the profile but use the profile pictures more when forming their judgement.

Further exploration into participants' judgements using Signal Detection Theory (Green & Swets, 1966) and criterion scores highlighted that the participants had a bias towards judging the profiles as fake. The presence of such a bias within the participants could be due to the instructions given to the participants at the start of the study: participants were aware they would be viewing social media profiles, of which some were fake, and some were real, and as such, the anticipation of this could have led to the response bias of judging the profiles as fake as the participants could have been more suspicious of the profiles overall.

A further limitation of this study is regarding the real profiles. The real profiles were gathered via word of mouth of the researcher, and as such were all known to the researcher. However, they were not a representative sample as all were of the same race and background, something which can be improved upon in further research. Additionally, the researcher did not manipulate the characteristics of these profiles, and so not all of the characteristics that were manipulated in the fake profiles were present in the real profiles. However, the researcher did include all of the same heatmap regions on every profile, whether that be fake or real, even when no information was present in these areas. For example, none of the real profiles had a *Bio*, but each heatmap overlay over the real profiles included a region around where the *Bio* could have been. This was to ensure that if any participants did click on this area it would be recorded as a click for *Bio* rather than *Other* .

A final limitation is in reference to said heatmaps. Heatmaps were included by the researcher in an effort to understand areas relied upon when making an authenticity judgement of a profile. Participants were allowed ten clicks per profile (the maximum number permitted by Qualtrics questionnaire software) to allow them to click on multiple areas of the profile that they used when making their judgements. This meant that across the sample there were hundreds of clicks in different areas, which gave the researcher an excellent overview of the areas used when making a judgement, but they did not give the researcher an idea of what each specific participant relied upon *most* when making their judgement. One way to counteract this effect would be for each of the clicks on the heatmap to be numbered in the order of the clicks. Doing so would give the researcher a much better understanding of the specific main areas used on the profiles, i.e., what areas the participants used first, and thus would mean the results garnered would provide a greater understanding of the areas used when making a judgement. However, at the time of writing, this option is not available using the Qualtrics software heatmap function, and so is an avenue future researchers should explore further.

The results outlined in this research take Psychology, particularly online deception, one step further to understanding the role humans play in online deception, by creating the basis for a ‘framework’ of fake characteristics of social media profiles. The purpose of the development of this framework is to assist those on social media to understand the key areas of social media profiles to look at when the authenticity of the profile is under question, i.e., if someone has been sent a friend request, how does one decipher from looking at the profile whether the profile is authentic and should therefore be accepted into their social media circle, or whether the profile is fake and should not be accepted? Further, this framework could be of assistance to the security services or police forces in specific reference to terrorist groups who recruit followers via online social media platforms, or child paedophiles who use social media to groom their victims. If there is a framework or outline of specific things to look for when seeing a social media profile for the first time and deciding the authenticity of that profile, then there could be less instances of users being scammed by fake accounts, or users being unknowingly exposed to terrorist propaganda.

Despite the limitations, this research has provided an initial understanding of the areas of Facebook profiles that are used when making judgements as to the authenticity of said profiles, something of which is yet to be studied from a psychological point of view. This research has also provided a better understanding of human deception detection in the online space, particularly on social networks, and also further evidence for the literature on human judgement accuracy of deception.

The judgement accuracy results overall not only support findings from Study 2, but also support findings from Study 1, whereby participants' judgement accuracy of fake profiles was highest for profiles with four fake characteristics and lowest for profiles with two fake characteristics. As such, it can be concluded from the linear trend in accuracy found in this study, that participants do have some level of ability to accurately judge a social media profile as fake by identifying the aspects of the profile that may be fake, and their ability to do so increases as the number of manipulations/cues increases. Further, the understanding of the types of areas used to inform judgements has also improved.

Chapter 4: Study 3

Introduction

The previous two studies showed a linear trend in judgement accuracy – participants are best at accurately judging a real profile, and worst for fake profiles, particularly those made to look as real as possible (i.e., with zero manipulated characteristics). This finding remained constant whether a person saw one type of fake profile (conditions) or multiple types of fake profiles (e.g., 2 fake characteristics only or these together with 4 fake profiles). Additionally, the finding showed that the type of photograph shown had the strongest impact on judgements.

To build upon these findings, the current study examines if the same pattern emerges when participants see all three types of fake profile, and the real profiles, in a repeated measures design. This study differs to Study 1 in that it introduces the *0 Fakes* profiles into the repeated measure design. This is to measure whether participants are still better at accurately judging the real profiles, or whether the exposure to all three different types of fake profiles (four different profile types overall) will affect their judgement accuracy.

Based on the results of the previous studies it was predicted that real profiles will be judged more accurately than fake profiles (Hypothesis 1), and profiles with the highest number of fake characteristics will be more accurately judged as fake, than profiles with fewer, or no, fake characteristics (Hypothesis 2). In regard to social media, it is expected that there will be a relationship between social media use (times spent per day) and judgement accuracy (Hypothesis 3), and between previous experience in creating a fake social media profile and judgement accuracy (Hypothesis 4). Due to partial acceptance of social media hypotheses in previous studies 1 and 2, it is difficult to predict the nature of the relationship hence why H3 and H4 are non-directional. It is also expected that a relationship will be found between participants' self-reported confidence in accuracy of their judgements and actual judgement accuracy (Hypothesis 5). Again, this is non-directional due to only receiving partial support in studies 1 and 2; as some relationships were observed it is warranted to predict further relationships to be observed within this study.

In relation to the manipulated characteristics of the profiles, it is expected that there will be a relationship between the manipulated characteristics of the Facebook profiles and participants' judgements of the profiles (Hypothesis 6), and that the manipulated characteristics will also have an effect on the accuracy of participants' judgements of the profiles (Hypothesis 7). As this study is the first time all participants were exposed to the 0 Fakes profile type, rather than having a specific 0 Fakes condition as in Study 2, it is difficult to predict the direction of these relationship, hence the non-directional nature of these final hypotheses.

Finally, as both previous studies have found very minimal effects, if any, of the individual difference's variables (personality type and social sensitivity) on judgement accuracy, the researcher cannot hypothesise that any relationship will be found in this study. As such, these variables will not be included within the hypotheses, but rather will be controlled for when analysing the data and results of such will be presented alongside the main analysis.

Method

Participants

An A-Priori power analysis of a repeated measures within subjects ANOVA, was conducted using G* Power (Faul et al., 2007) prior to data collection to determine the appropriate sample size for this study. The analysis indicated that a sample size of 24 participants would be sufficient to detect a medium effect size of f = 0.25, with an alpha level of α = 0.05 and a power of 1−β = 0.80. However, 200 participants were recruited. As outlined in Study 1 (Chapter 2) the reasoning behind this decision was to enhance the reliability and generalisability of the findings by reducing the errors associated with the estimates, improve the representativeness of the sample, and reduce the risk of Type 1 and Type 2 errors. Additionally, based on the results of Study 1 (Chapter 2) where a similar design was employed, a similar sample size of 200 participants was chosen to maintain consistency across the studies within this research and ensure results between each study were comparable.

A total of 202 participants were recruited online via Prolific through means of volunteer sampling. Two participants did not consent to participate in the study, thus the total number of complete entries used within data analysis was 200. Participants were aged between 18 and 67 years, with a mean age of 27.62 years. Of the 200 participants, 92 (46%) identified as Male, 104 (52%) identified as Female, 1 (0.5%) identified as Transgender, 1 (0.5%) identified as Non-Binary, 1 (0.5%) identified as Other. One participant (0.5%) selected the option of ‘Prefer not to say’. Participants were Asian or Asian British (N = 16, 8%), Black, African, Black British, or Caribbean (N = 9, 4.5%), Mixed or Multiple Ethnic Groups (N = 8, 4%), White, including any White backgrounds (N = 144, 72%), or Another Ethnic Group (N = 15, 7.5%). A total of six (3%) participants selected the option of ‘Prefer not to say’.

Participant birth locations were Poland (N = 34, 17.0%), Portugal (N = 29, 14.5%), Mexico (N = 21, 10.5%), or United Kingdom (N = 16, 8.0%). The most popular locations participants currently reside in were Portugal (N = 33, 16.5%), United Kingdom (N = 30, 15.0%), Poland (N = 29, 14.5%), and Mexico (N = 20, 10.0%). The overwhelming majority of participants reported that they were residing in the same country that they were born (N = 164, 82%). Thirty-six participants (18%) had moved locations; with 22 (61.11%) of these staying within the same continent, and thus remaining within the same culture.

Design

A 4 (Real profiles, 0 Fakes profiles, 2 Fakes profiles, and 4 Fakes profiles) x 2 (Accurate judgement vs. Non-accurate judgement) experimental Turing test design was used. The Dependent Variable (DV) is the accuracy of the judgements, and the Independent Variable (IV) is the Facebook profiles. Both variables are within subjects' measures following a repeated measures design.

Measures & Materials,

Measures

As per both previous studies, the same self-report measures of social media activity, the Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow & Swann, 2003), the Social Sensitivity (SS) scale (Riggio, 1986), and a follow-up questionnaire measuring participants' self-reported accuracy and previous experience in creating a fake profile.

Both personality questionnaires, the Ten-Item Personality Inventory (TIPI) and the Social Sensitivity Scale (SS Scale) were scored, in line with each respective manual. This included coding each Likert scale from 1-5 or 1-7 respective to each questionnaire, and reverse scoring some items using an SPSS syntax command. The TIPI questionnaire scoring provided a score on each of the Big-5 personality traits for each participant; Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to new experiences, and the Social Sensitivity Scale scoring provided one overall score.

Materials

A total of 74 Facebook profile screenshots were used; the same 68 fake profiles and six real profiles as used in Study 2. The 68 fake profiles created by the researcher contained different combinations of seven manipulated profile characteristics: *Photo Type, Photo Number, Bio, Intro, Posts Content, Comments Number, and Likes Number.* Each of these seven profile characteristics were identified in Study1 as needing further exploration in Study 2. Due to mixed results from both previous studies, these same profiles are being used again to investigate whether any results are replicated. The six real profiles were not manipulated in terms of profile characteristics, the only 'manipulations' that occurred were the omittance of identifying information for the purposes of ethical guidelines, such as the names of the profile owners, friends' photos, or names of friends that were commenting/interacting with the profile.

The 68 fake profiles consisted of a mixture of: 12 profiles with 0 fake characteristics, 21 profiles with 2 fake characteristics, and 35 profiles with 4 fake characteristics. As each type of profile has a different number of possible combinations of the seven fake characteristics, each type of profile therefore has a different total number, i.e., there were 21 possible combinations of 2 fake characteristics, and 35 combinations of 4 fake characteristics, hence the total number of profiles reflects this calculation (see Appendix P for characteristic framework). The profiles with *0 Fakes* technically had no manipulations of the fake characteristics, as these profiles were created specifically to fool participants into judging the profile as real (i.e., the profiles were created to look as real and authentic as possible). To allow for a level of variability and random assignment to participants, 12 profiles with *0 Fakes* were created.

The same six real Facebook profiles were used for this research as those used in the previous studies conducted by the researcher. These participants were well known to the researcher. This allowed for their profiles to be validated as real and authentic. Demographic information was not directly obtained from the participants, but as they are well known to the researcher it can be reported that; their ages ranged from 28-60 years (M = 41.67), all six participants identify as White British, and genders were split equally with three Females (50%) and three Males (50%).

Three accuracy scores were calculated for each participant: number of accurate 'true' judgements, number of accurate 'fake' judgements and 'total' number of accurate judgements (combination of fake and real accuracy). The accuracy score was split across the four different types of profiles: *0 Fakes, 2 Fakes, 4 Fakes,* and *Rea'*. Participants saw three of each type of profile thus can achieve a maximum score of 3 for each, or 12 overall (*Overall Accuracy*).

Procedure

The procedure near identically replicates that of Study 1. All participants were recruited via Prolific and redirected to the study on Qualtrics once informed consent was obtained. During the study, all participants were first required to complete three of the self-report measures, the social media questionnaire, TIPI, and SS Scale. Following this, participants were provided with a set of instructions in relation to the profile phase of the study, whereby they were informed that they would see 12 Facebook profile screenshots in a random order and asked to make a judgement as to the authenticity of the profile. Participants were also asked to identify the areas of the profile they used when making their authenticity judgement by clicking on the specific areas of the profile screenshot. Following the completion of the profile phase, all participants were asked to provide brief demographic details. As per both previous studies, participants were fully debriefed and provided with further information on the research and contact details of the researcher should they wish to withdraw or ask any further questions.

Procedurally where this study differs from that of Study 1 is in the profile phase. Participants were informed that they would see 12 profiles, as before, however in this study they were informed that there was not an equal split of real and fake profiles (i.e., they would not see six real profiles and six fake profiles (see Appendix R for updated instructions), but rather three random profiles from each of the profile types: three 0 Fakes, three 2 Fakes, three 4 Fakes, and three real profiles). During the debrief, participants were informed that they had seen nine fake profiles, and three real profiles (see Appendix S for updated debrief form).

Ethics

This research was fully approved by the ethics committee at Lancaster university on 14th May 2021, under an amendment to the same ethics submission as Studies 1 and 2.

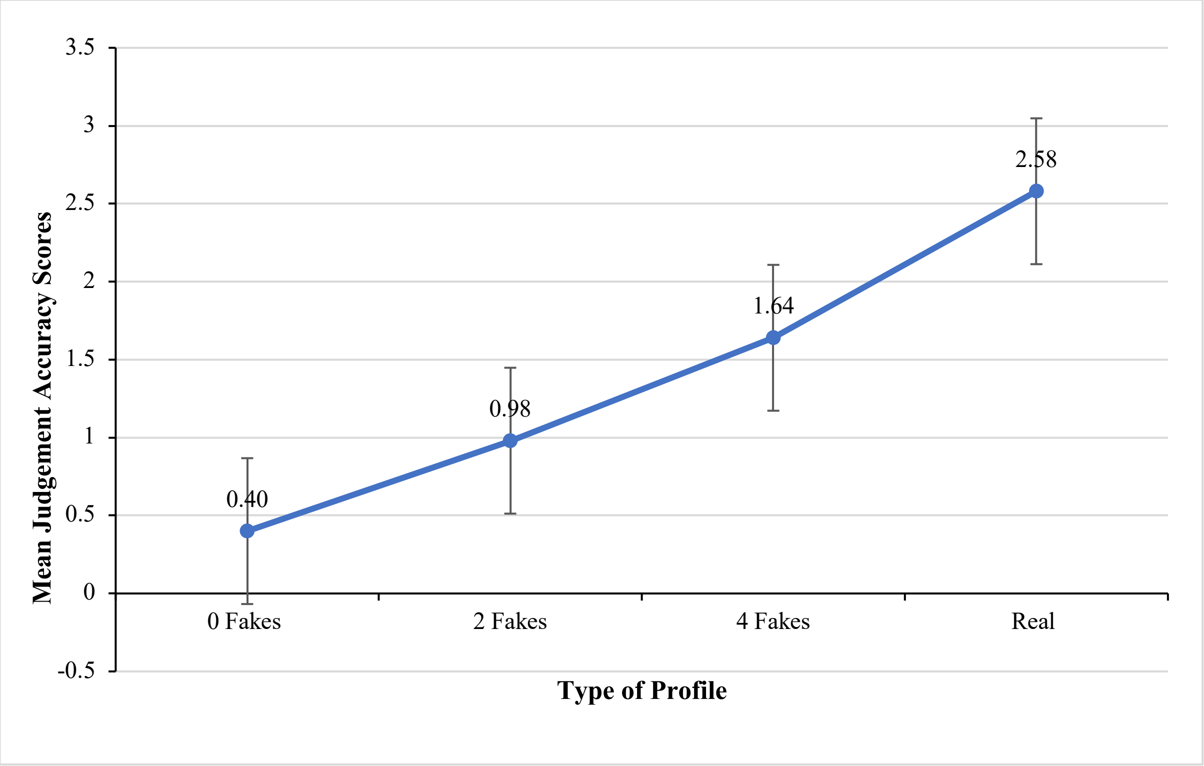
Results

Prior to conducting the statistical analyses, multiple normality tests were conducted to assess the appropriateness of the data for statistical analyses, namely t-tests, ANOVAs, and Regression models. The normality tests showed 8 outliers on judgements of the zero fakes profiles (3 outliers), and the real profiles (5 outliers). More specifically, these scores were lower than the distribution, namely participants who scored zero. A visual inspection of the histograms showed mainly normal distributions with only few with a slight positive or negative skew, and all Q-Q plots showed data of a linear pattern. Overall, the data were regarded as normally distributed, and as such, the outliers were not removed from analyses.

Profile Accuracy

Mean judgement accuracy scores of the real profiles and all three types of fake profiles were compared. Results show that when looking at all three types of fake profiles collectively, participants were better at accurately judging fake Facebook profiles (M = 3.02, SD = 1.61) than they were at judging real Facebook profiles (M = 2.58, SD = 0.70). However, when the mean scores for each type of fake profile are examined individually, participants' judgement accuracy is highest when judging real profiles. Figure 1 shows participants' mean accuracy scores for each type of profile.

Figure 1
Mean judgement accuracy scores of 200 participants for each type of profile.



As evidenced in Figure 1, there is a linear trend in mean accuracy scores across the different profiles; fake profile accuracy increases as the number of manipulated characteristics of the profile's increases, and overall, judgement accuracy is highest for real profiles.

Further analysis of participants' accuracy scores revealed that zero participants achieved a maximum judgement accuracy score of 12. However, a linear trend of maximum judgement accuracy was found in the individual total accuracy scores for each type of profile; 1% (N = 2) correctly judged all three *0 Fakes* profiles, 4% (N = 8) correctly judged all three *2 Fakes* profiles, 18% (N = 36) correctly judged all three *4 Fakes* profiles, and 67.5% (N = 135) correctly judged all three real profiles. Participants' overall mean accuracy score was 5.60 (SD = 1.58) , which is significantly lower than chance level of 6; $t(199) = -3.58$, $p < .001$, with a mean difference of -.400. Thus, participants' overall judgement accuracy is lower than that of chance.

To further investigate the mean differences in accuracy scores, a one-way repeated measures ANOVA was conducted in SPSS. The within subjects' factor was 'type of profile' with four levels: *0 Fakes, 2 Fakes, 4 Fakes*, and *Real*. Results showed that there was a violation of sphericity as Mauchly's test was highly significant, $x^2(5) = 19.61$*, p<.001.* To correct this, the Greenhouse-Geisser correction was used ($\varepsilon = .946$), as suggested by Maxwell and Delaney (2004). The results showed a significant main effect of profile type on accuracy; $F(2.84, 564.50) = 283.50$, $p <.001$, $n^2 = .588$. Pairwise comparisons show that participants were statistically less accurate at judging *0 Fakes* compared to *2 Fakes* profiles (0.58, 95% CI [.40, .76], $p < .001$), or *4 Fakes* (1.24, 95% CI [1.05, 1.43], $p < .001$), and at judging *2 Fakes* compared to *4 Fakes* (0.66, 95% CI [.44, .89], $p < .001$).

To understand participants' decision-making accuracy in regard to accurately judging a profile as either real or fake, Signal Detection Theory (SDT) was used. To test the ability of participants to decipher between fake and real profiles, fake profile accuracy and real profile accuracy scores were first transformed into hit rate and false alarm rate scores. From this, d-prime (*d'*) values and criterion (*c*) scores were calculated; *d*' is a measure of sensitivity used to indicate participants' abilities at discriminating between the signals (fake profiles) and the noise (real profiles) within the study, and *c* is a measure of participants' response bias (i.e., were participants biased more towards answering yes or no, or in this case, fake or real, when judging the profiles?).

Results indicate that participants' ability to distinguish signals (fake profiles) from noise (real profiles) was greater than zero (*d*' = 0.59, 95% CI [2.80, 3.24]), meaning they were able to identify fake profiles as fake. Additionally, participants did show a bias to responding 'yes' (*c* = -0.72), meaning they were biased to judging the profiles as fake. However, as reported, judgement accuracy was higher for real profiles when compared to each individual fake profile. This result, alongside the *d'* value and the *c* score, suggests that the fake profiles were deceptive enough to fool participants in to judging them as real: participants were statistically able to distinguish between the fake and real profiles *and* had a response bias towards judging the profiles to be fake, yet still judged the majority of fake profiles as real. However, these results should be interpreted alongside the knowledge that participants' judgement accuracy levels were not greater than that of chance, meaning that while participants were able to statistically distinguish between fake and real profiles, their overall accuracy of this ability was no better than random guessing.

The profile accuracy results presented above show that real profiles were more accurately judged than fake profiles of all types, and *4 Fakes* profiles were more accurately judged as fake than *2 Fakes* or *0 Fakes* profiles, thus H1 and H2 can be accepted.

Self-Reported Accuracy

To examine the expected relationship between self-reported accuracy and actual accuracy in judgements (H5), participants were asked how accurate they thought their judgements were. The majority of participants (N = 67, 33.5%) were 'Slightly

Confident' in the accuracy of their judgements. Very few participants were wholly 'Unconfident' (N = 6, 3%) or 'Confident' (N = 9, 4.5%).

To analyse whether the presence of such a relationship multiple regressions were conducted for accuracy scores for each type of profile. The variable of 'Self-Reported Accuracy' is a categorical variable with seven levels: 'Unconfident', 'Moderately Unconfident', 'Slightly Unconfident', 'Neutral', 'Slightly Confident', 'Moderately Confident', and 'Confident'. To allow for input into the regression, each level was coded into a dummy variable. The new dummy variable 'Confident' was used as the constant in the regression. The results are reported in Table 1.

Table 1.
Multiple regression of self-reported accuracy and judgement accuracy scores for each type of profile

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Self-Reported Accuracy[a]	.047			.047			.016			.018		
Constant		0.44	0.22		1.56	0.28		2.11	0.31		2.67	0.23
Unconfident		-0.28	0.35		-1.22**	0.44		-0.78	0.49		0.17	0.37
Moderately Unconfident		-0.06	0.27		-0.50	0.34		-0.44	0.38		0.11	0.29
Slightly Unconfident		0.23	0.25		-0.53	0.31		-0.44	0.35		-0.17	0.26
Neutral		0.06	0.28		-0.62	0.35		-0.49	0.39		-0.17	0.29
Slightly Confident		-0.16	0.23		-0.68*	0.30		-0.51	0.33		-0.14	0.25
Moderately Confident		-0.08	0.24		-0.52	0.30		-0.49	0.34		-0.05	0.25

Note. [a]*df* = 6,193. **p* < *.05,* ***p* <.01.

It is evident from Table 1 that there are minimal statistically significant relationships between participants' self-reported judgement accuracy and actual judgement accuracy across all four types of profiles. The statistically significant relationships that were found were both in the regression model analysing the judgement accuracy scores of profiles with two fake characteristics. Interestingly, both of the coefficients in these relationships are negative, meaning that when compared to the constant, judgement accuracy levels were statistically significantly *lower* when participants self-reported their confidence in their judgement accuracy as ‘Unconfident’ (B = -1.22) and ‘Slightly Confident’ (B = -.68). The result in relation to participants feeling ‘Unconfident’ in their judgements is reflected in their score being lower. In contrast, those who reported feeling ‘Slightly Confident’ also had lower accuracy scores, which suggests an overconfidence in participants’ own ability at accurately detecting fake and real profiles. However, all four regression models were non-significant: *0 Fakes*, F(6, 193) = 1.57 , p = .158; *2 Fakes*, F(6, 193) = 1.58 , p = .155; *4 Fakes*, F(6, 193) = 0.54, p = .779; *Real*, F(6, 193) = 0.60 , p = .728, meaning self-reported accuracy is not a good predictor of participants’ actual judgement accuracy, thus hypothesis H5 cannot be accepted.

Social Media

Participants were asked a series of questions in relation to social media use, specifically how much time they spend on social media per day, the platforms used, purpose for using social media, and whether they have previous experience creating a fake profile. It is expected that a relationship will be found between judgement accuracy and time spent on social media per day (H3), and previous experience creating a fake profile (H4).

Platforms

Prior to viewing the profiles, participants were provided with a list of seven social media platforms (Facebook, Twitter, Instagram, Snapchat, TikTok, YouTube, Other) and were asked to indicate the social media platforms they use and rank each of those platforms in order of use from most used to least used. All participants selected at least one platform that they use, however 29 participants did not rank their selections, meaning only 171 participants both selected *and* ranked the social media platforms they use. Of these 171 participants, only four selected and ranked all available platforms. Table 2 reports the most popular platforms and the frequencies of their individual rankings.

Table 2.
Participant rankings of social media platforms, based on the platforms they use most often (1st ranking), to the platforms they use the least (7th ranking).

Social Media Platform	Ranking Order							Total number of participants who use each platform
	1st	2nd	3rd	4th	5th	6th	7th	
Facebook	42	38	25	22	13	1	0	155
Twitter	15	21	23	18	6	6	0	96
Instagram	37	36	37	21	6	2	0	152
Snapchat	4	5	7	10	10	5	2	45
TikTok	13	17	16	14	5	1	0	71
YouTube	45	46	39	18	5	2	0	175
Other	15	6	5	3	3	0	2	34
Total number of rankings made	171	169	152	106	48	17	4	

As is shown in Table 2, the most popular social media platform amongst participants is YouTube (N = 175) with 45 participants (25.7%) ranking it as their most used platform. Facebook is the second most popular platform (N = 155) with 42 (27.1%) participants ranking it as the platform they use most often. The least selected option was Other (N = 34, 17%). Participants who selected the option *Other* were asked to enter the names of the platforms they used that were not included in the list. Such entries included the following platforms: Reddit, Strava, Weverse, Discord, Pinterest, Tumblr, WhatsApp, Twitch, Telegram, LinkedIn, and WordPress. Of these entries, Reddit was the most popular entry (N = 11). Overall, 155 participants (77.5%) are users of Facebook and thus are familiar with Facebook as a platform and have a basic understanding of the way the platform works and the components of a Facebook profile.

Purposes

Participants were asked to indicate the specific purposes they use social media from a list of 12 options (See Appendix A for full list). The two most popular purposes were 'Watching videos (TV/Films/YouTube etc.)' which was selected by 176 participants (88.0%), and 'Socialising with friends/keeping in touch' which was selected by 165 participants (82.5%). Interestingly, the option of 'Making friends/meeting new people', an option that could be considered as one of the main purposes of *social* media and an option where the accurate judgement of the authenticity of profiles would be the most important, was only selected by 53 participants (26.5%).

One of the 12 options given to participants was Other, whereby participants could record the specific purposes they use social media. In total, 12 participants selected this option, entering the following purposes: "communicate with my favourite artists", "finding inspiration", "staying up to date with colleagues", "self-help medical related videos", "entertainment", "memes", "watch videos for instructional purposes", "sharing my art, looking at other peoples' art", "discovering new music, films, books, etc", "watching fun content", "entertainment in form of viewing pictures of my areas of interest (travel, lifestyle…)", and "following artists and checking their work". With exception of the comments relating to finding inspiration, the remaining comments could be encompassed under the umbrella of the other given 11 options.

Daily Usage

Of the 200 participants, 198 (99%) indicated that they were regular users of social media. The most frequent duration spent on social media was ‘2-3 hours’ (N = 54, 27%), closely followed by ‘1-2 hours’ and ‘more than 4 hours’ (N = 50, 25%). Only 11 participants (5.5%) used social media for ‘Less than 1 hour’ .

To further investigate the effect of social media use on participants’ judgement accuracy, multiple regressions were conducted. The variable ‘Hours Spent on Social Media per day’ is a categorical variable with five different levels: ‘Less than one hour’, ‘1-2 hours’, ‘2-3 hours’, ‘3-4 hours’, and ‘More than four hours’. As such, dummy coding was used in which ‘Less than one hour’ was the reference category against which other categories were compared. Four models were tested – one for each type of profile (0 Fakes, 2 Fakes, 4 Fakes, and Real).

Prior to analysis, assumption testing was completed for the multiple regression tests. The data did not violate any of the assumptions. Results of the regressions are reported in Table 3.

Table 3.
Multiple regression for social media time predictors of overall judgement accuracy for each type of profile.

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Hours spent on social media per day	.014			.004			.009			.022		
Constant		0.55	-0.20		1.00	0.26		1.64	0.28		2.55	-0.21
1-2 Hours		-0.11	0.22		-0.06	0.28		0.10	0.31		-0.03	0.23
2-3 Hours		-0.08	0.22		0.02	0.28		0.03	0.31		-0.08	0.23
3-4 Hours		-0.17	0.23		0.06	0.29		0.02	0.32		0.11	0.24
4+ Hours		-0.27	0.22		0.08	0.28		-0.14	0.31		0.18	0.23

Note. df = 4,195.

Due to the lack of statistically significant coefficients in Table 3, it is evident that time spent on social media per day does not predict judgement accuracy for any type of profile. Additionally, each of the models were non-significant: *0 Fakes*, *F*(4, 195) = 0.72 , *p* = .851; *2 Fakes*, *F*(4, 195) = 0.19 , *p* = .942; *4 Fakes*, *F*(4, 195) = 0.45 , *p* = .774; ‘Real’, *F*(4, 195) = 1.09 , *p* = .361.

Previous Experience in Creating a Fake Profile

Participants were asked if they had any previous experience in creating a fake social media profile on any social media platform. A total of 31 (15.5%) participants indicated that they had previous experience in creating such profiles, 16 Females (51.6%), 14 Males (45.2%), and one participant (3.2%) who did not disclose their gender. These participants were asked to indicate why they had created a fake profile. Participants reported *investigative reasons* (“…I wanted to test my boyfriend for cheating”, “I have in the past to spy on ex’s [sic]”, “I wanted to view some profiles without they knowing, like ex-girlfriends of boys I had crushed on, or some person I don’t get along with [sic]”, “…to help my friend catch her boyfriend in lies”); *personal reasons* (“I didn’t want friends and family finding me on the social media accounts…”, "I just wanted to have another Facebook…..I think maybe one day it will help me", "…I needed it to look for information like university exams or events", "I wanted to participate in groups/communities that I was too ashamed to join with my personal account"); or *anonymity/privacy reasons* ("I created a fake profile to join a Facebook group anonymously", "…I wanted to express my feelings anonymous to my Facebook friends [sic]", "Just for anonymity reasons", "I didn't want to associate certain things that I like with my real account or name"). Only a few participants cited that they created the profile for *malicious reasons* ("…to prank a friend", "For joking purposes"). More participants reported creating the profiles for *gaming purposes/for fun* ("For roleplaying games", "…I liked a lot of games on fb [sic] so I created another profile to get more rewards", "use in game groups and chats", "Just for fun").

Overall, the vast majority of the participants that have created fake social media profiles in the past have done so without malicious intent, and more so for personal reasons or personal gain in terms of gaming. However, it is evident that there is still an element of deception in regard to creating fake profiles, as participants did report using the profiles to spy on others, or catch others out in lies etc, suggesting that there is still an intention to deceive others by use of fake social media profiles.

To investigate whether having previous experience of creating a fake profile can predict accuracy scores, a multiple regression was conducted using accuracy scores for each type of profile. As the variable is categorical it was dummy coded prior to analysis to allow for input into the regression model. 'No' was used as the constant. Table 4 presents the results of the regression analysis.

Table 4.
Multiple regression of previous experience creating a fake profile and accuracy scores for each type of profile.

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	*R^2*	*B*	*SE*	*R^2*	*B*	*SE*	*R^2*	*B*	*SE*
Experience in creating a fake social media profile	.001			.005			.005			.006		
Constant		0.41	0.05		1.01	0.06		1.67	0.07		2.56	0.05
Yes		-0.05	0.13		-0.17	0.16		-0.19	0.18		0.15	0.14

Note. df = 1,198

Table 4 shows no significant coefficients, meaning no relationship was found between previous experience creating a fake profile and accuracy for all profile types. Additionally, each of the models were non-significant: 0 Fakes, *F*(1, 198) = 0.17 , *p* = .682; 2 Fakes, *F*(1, 198) = 1.04 , *p* = .309; 4 Fakes, *F*(1, 198) = 1.06 , *p* = .305; 'Real', *F*(1, 198) = 1.27 , *p* = .261.

As a result of analysis of the social media variables, H3 cannot be accepted as no significant relationships were found between the number of hours participants spend on social media per day and their judgement accuracy of each type of profile. Similarly, H4 cannot be accepted as no relationship between previous experience creating a social media profile and judgement accuracy was found for any profile type.

Manipulated Characteristics of Profiles

To further understand the fake social media profiles, several general linear mixed effects models (*'glmer')* were conducted in R using the 'lme4' package (Bates, et al., 2015), to analyse whether the manipulated characteristics of the profiles influenced participants' judgements and accuracy of said judgements. Both judgement and judgement accuracy models were conducted with the manipulated factors of the profiles as predictors (*Photo Type*, *Number of Photos, Bio, Intro, Posts Content, Number of Comments,* and *Number of Likes*) with 'Prolific ID' and 'Profile Number' as random effects. The addition of 'Profile Number' as a random effect led to a significant improvement in the model fit over Prolific ID alone (p < .001), and thus the models reported below in Table 5 include both 'Prolific ID' and 'Profile Number'. It is predicted that there will be a relationship between the manipulated characteristics and both participants' judgement of the profiles (H6) and accuracy of judgements (H7).

Table 5.
Results from 'glmer' Model's 1 & 2 where judgement and accuracy are regressed on the manipulated profile characteristics.

Predictors	Model 1 – Judgement[b]					Model 2 – Accuracy[c]				
	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*	
Fixed effects										
Intercept	-2.20	0.18	0.08	.16	<.001***	-1.01	0.27	0.21	.62	<.001***
Photo Type[a]	2.17	0.22	5.68	13.38	<.001***	1.70	0.33	2.85	10.56	<.001***
Number of Photos[a]	0.37	0.21	0.95	2.19	.086	0.02	0.33	0.53	1.97	.940
Bio[a]	0.32	0.21	0.91	2.10	.130	0.04	0.33	0.54	2.00	.900
Intro[a]	0.30	0.21	0.89	2.05	.159	0.02	0.33	0.53	1.96	.951
Post Content[a]	0.20	0.21	0.80	1.86	.347	-0.13	0.33	0.46	1.69	.706
Number of Comments[a]	0.40	0.21	0.98	2.28	.059	0.02	0.33	0.53	1.97	.948
Number of Likes[a]	0.54	0.21	1.13	2.60	.011*	0.23	0.33	0.66	2.41	.487
Random effects										
Residual Variance (σ^2)	3.29					3.29				
τ_{00} PROLIFICID	0.70					0.12				
τ_{00} PROFILENUM	0.44					1.50				
Intraclass Correlation Coefficient	0.26					0.33				

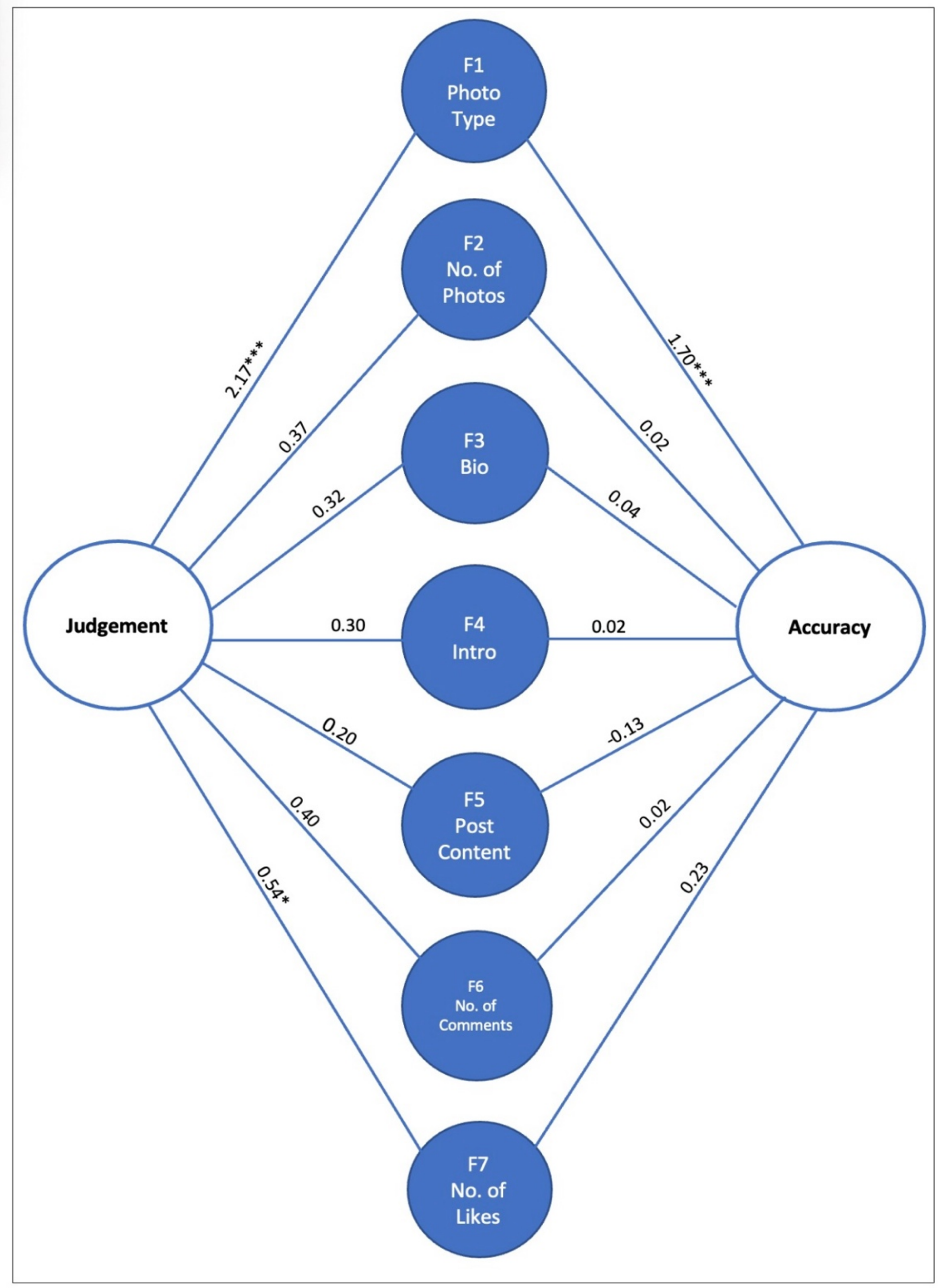
Note. N = 200, Number of Profiles = 74, total N = 2400. *p =.05, ** p =.01, *** p<.001.
[a] 0 = Judgement of 'Real', 1 = Judgement of 'Fake'. Model 2: 0 = Non-Accurate Judgement, 1 = Accurate Judgement. [b] Conditional R^2 = .24. [c] Conditional R^2 = .098.

It is evident from looking at Table 5 that the only manipulated factors that are statistically significant predictors of participants' judgements are *Photo Type* and Number of Likes, and *Photo Type* only for participants' judgement accuracy. When *Photo Type* was manipulated on the profiles, participants were significantly more likely to judge the profile as fake (*B* = 2.17), and that judgement of fake is significantly more likely to be accurate (*B* = 1.70). Additionally, if *Number of Likes* had been manipulated on the profiles, participants were more likely to judge the profile as fake (*B* = 0.54), however *Number of Likes* is not a significant predictor of accuracy. This finding suggests that participants may over rely on the *Number of Likes* to make their judgement, as the lack of accuracy in relation to these judgements suggests participants may be relying on this area when it has not been manipulated.

Figure 2 presents a lens model approach comparison based on Brunswik's lens model (as defined in Study 1 – Chapter 2), between Model's 1 and 2 with the estimates and statistical significance for each of the factors shown.

Figure 2.

Lens model diagram showing estimates and statistical significance of the manipulated characteristics for models 1 and 2.



Overall, Model's 1 and 2 provide evidence for both hypotheses 6 and 7 as an effect of manipulated characteristics has been found for both participant judgement and participant accuracy. As such H6 and H7 can be accepted.

Personality

As mentioned previously, both studies 1 and 2 found very minimal, if any, significant relationships between personality traits, social sensitivity scores and judgement accuracy. As such, the researcher did not hypothesise any relationships between these variables and judgement accuracy in this study, however, these variables were still controlled for. To investigate whether any effects of personality and social sensitivity on judgement accuracy were present, four multiple regression models were conducted – one for each type of profile (*0 Fakes* accuracy, *2 Fakes* accuracy, *4 Fakes* accuracy, and 'real' accuracy), with the TIPI trait scores and scores on the Social Sensitivity (SS) Scale as predictors. The SS scale was included to assess if a participants' individual ability to interpret the verbal communication of others, and their sensitivity and understanding of social norms, had a significant effect on the accuracy of their judgements.

Prior to data analysis, the data was checked that it met the assumptions of a multiple regression. Linearity was assessed via visual inspections of the scatterplots of each personality variable against each accuracy score. To allow for this assessment, scores from the personality variables and accuracy scores were transferred to excel for 'jittering' of the data. As each of the variables consisted of discrete data, the plots produced in SPSS were 'overplotted' meaning the resulting plot did not give a good indication of any trends in the data. Jittering the data added random noise into the data which effectively split the data points apart from one another, thus presenting a scatterplot without overplotting. The linear relationships that were found in each of the scatterplots were not strong, with all but one slope coefficient being below one, however all did have a linear relationship. Additionally, plot inspections of the standardised residuals showed homoscedasticity and the Durbin-Watson test showed that each personality variable had independence of residuals against each accuracy score. Further, the data met all the same assumptions as above for the multiple regression, including the additional assumptions of multicollinearity, leverage points, and Cooks' values. Table 6 presents the results of the multiple regressions with personality variables as predictors.

Table 6.
Multiple regression for personality predictors of overall judgement accuracy for each type of profile.

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
TIPI	.029			.049			.020			.026		
Extraversion		-0.02	0.48		-0.02	0.04		-0.03	0.05		0.02	0.04
Agreeableness		-0.01	0.03		-0.13*	0.06		0.05	0.06		-0.05	0.05
Conscientiousness		0.04	0.05		-0.03	0.05		0.07	0.05		-0.04	0.04
Emotional Stability		-0.01	0.04		0.01	0.05		-0.01	0.06		0.05	0.04
Openness to New Experiences		-0.07	0.04		0.10	0.06		-0.02	0.06		0.04	0.05
SS Scale		0.01	0.01		-0.01	0.01		-0.01	0.01		0.01	0.01

Note. df = 6, 193. **p < .05*

Table 6 shows that personality traits and SS scores are not good predictors of participants' judgement accuracy, a pattern found across all four different types of profile. The only significant relationship was found between 'Agreeableness' and accuracy scores for profiles with two fake characteristics. The negative coefficient indicates that the higher the levels of 'Agreeableness', the lower the accuracy score. However, the *2 Fakes* model overall was non-significant: *2 Fakes*, *F*(6, 193) = 1.65, *p* = .135. Additionally, the remaining three models were also non-significant: *0 Fakes*, *F*(6, 193) = 0.95, *p* = .462; *4 Fakes*, *F*(6, 193) = 0.65, *p* = .688; *Real*, *F*(6, 193) = 0.85, *p* = .536.

These results show that while the individual difference's variables were controlled for in this study, they did not have a significant effect on participants' judgement accuracy. Such findings support those from Studies 1 and 2 and further reinforce that these variables may not be key to predicting judgement accuracy.

Post-hoc Analysis

Heatmaps

Each of the profiles had a heatmap function layered on top to allow for obtaining data in relation to the specific areas of the profile used by participants to inform their judgement of the authenticity of the profile. Participants were instructed to click on the areas of the profile they used when making their judgement, and to be as accurate as possible when doing so. Each participant was given a maximum of 10 clicks per profile. The heatmap layer was non-visible to participants, only their individual clicks on the profile were visible, which appeared as a small dot. From this data it was possible to produce a single heatmap image that contained the heatmap clicks from all participants presented as colour splotches. The colour scale used to denote the number or clicks uses blue areas for the lower, or 'cooler', end of the scale, which represents few clicks, and red areas to indicate higher, or 'warmer', end of the scale, which represents a larger number of clicks. An example of a profile with a heatmap layer is presented in Appendix L.

For the purposes of analysis, 'regions' were added to each profile to indicate the manipulated characteristics of the profiles (*Photo Type*, *Photo Number, Bio, Intro, Posts Content, Number of Comments,* and *Number of Likes*). These regions encompassed each of the manipulated characteristics but did not overlap with any other regions to ensure each click for each region was recorded correctly (example shown in Appendix K). Any of the clicks within these regions would be recorded as a click for the respective manipulated factor listed above. The addition of these regions meant that the frequencies of clicks within each region was produced alongside the heatmap images as a second form of data. The frequencies for each region in each of the profile types are shown in Figure 3.

Figure 3

Frequency of clicks within each heatmap region for the four types of profile

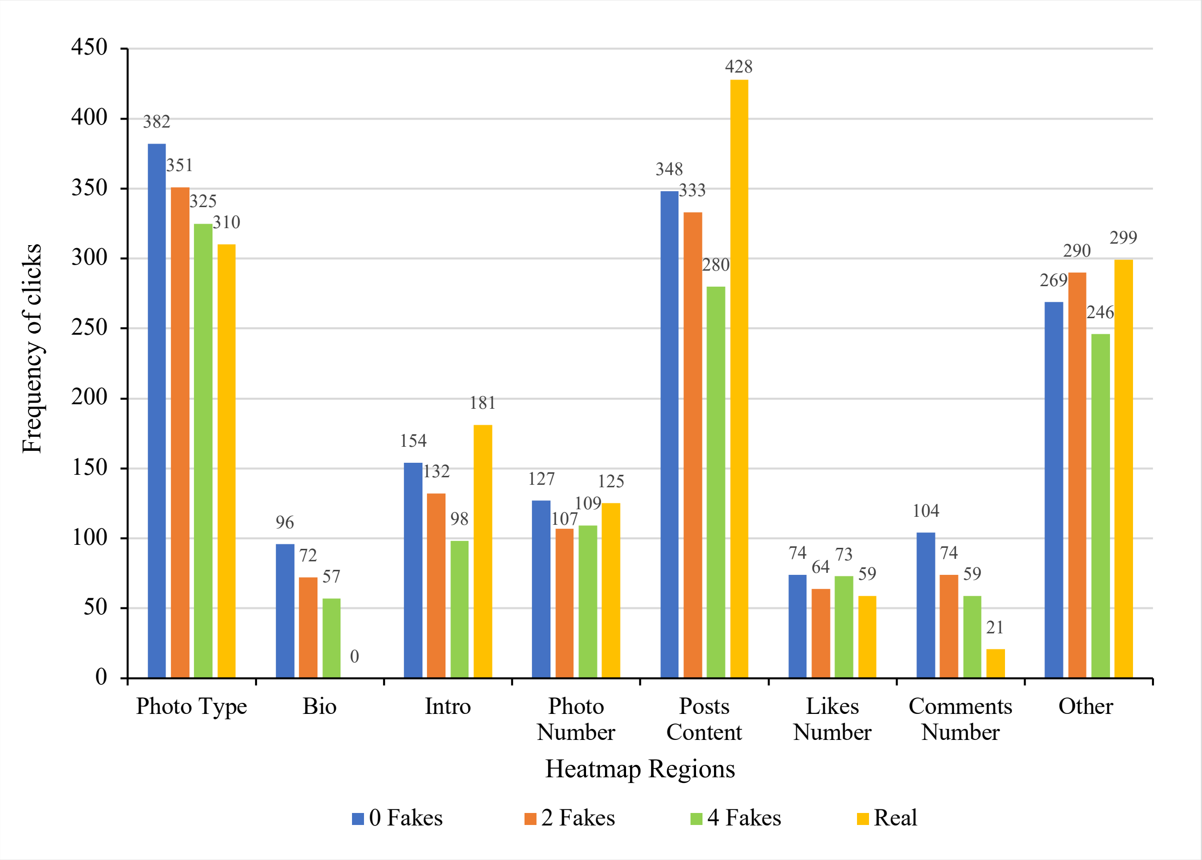


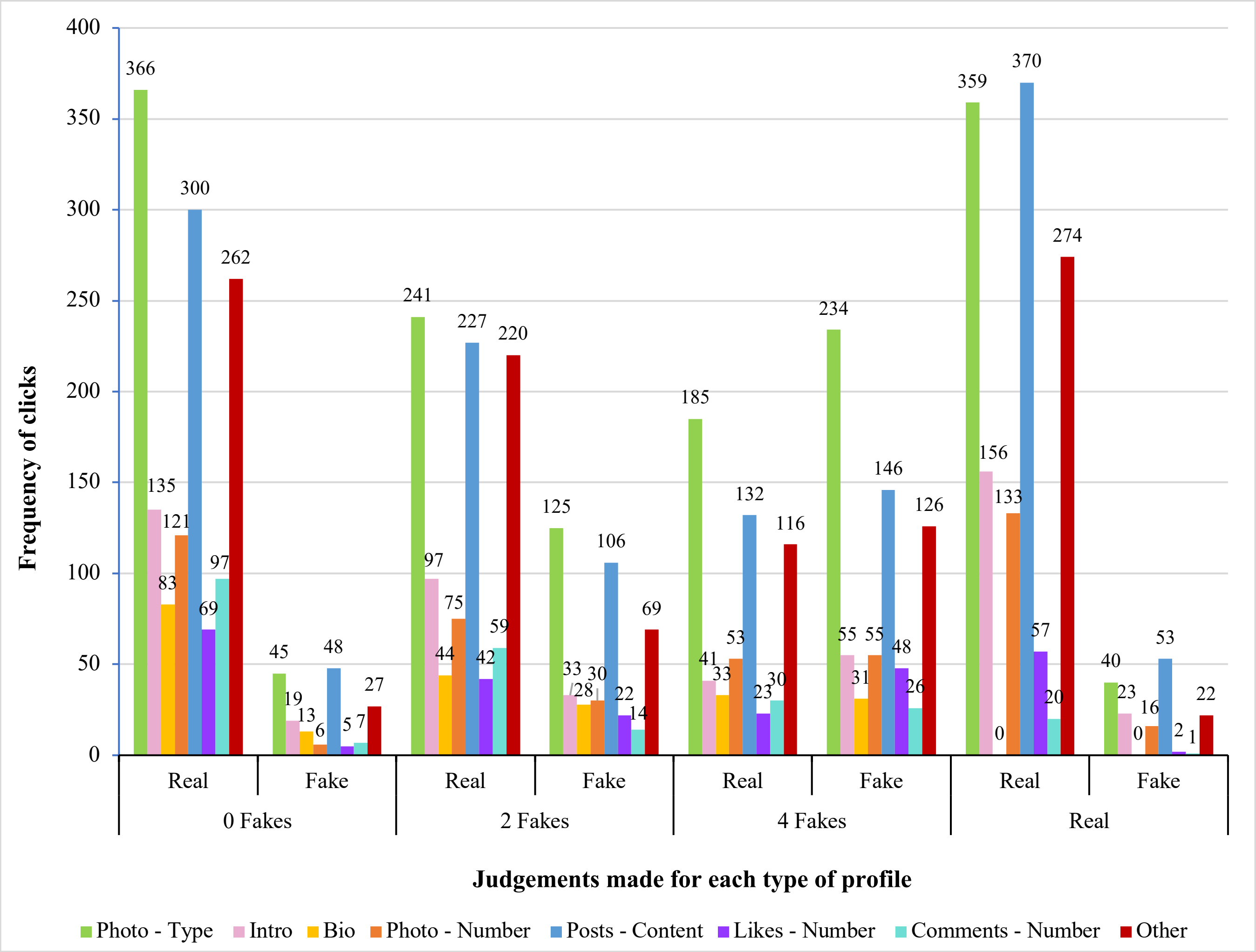
Figure 3 shows that *Photo Type* was relied upon most when judging the authenticity of fake profiles, mostly so for profiles with *0 Fakes*. Interestingly, real profiles were judged mainly by Posts Content, followed by *Photo Type*. These frequencies suggest that when judging fake profiles, participants rely heavily on the visual imagery of the profile, whereas when judging real profiles, participants rely heavily on the written content available on the profile and second most on the visual imagery.

The category of *Other* is the third most clicked area across all four types of profile. The region of *Other* is comprised of clicks that do not fall under the specified regions listed previously (i.e., parts of the profile that have not been directly manipulated). Through manually checking each heatmap image it is apparent that the vast majority of clicks in the *Other* region are general inaccurate clicks, (i.e., they are just outside of the specified regions set by the researcher), or they are clicks in the empty white/grey spaces around the posts and the profile. Very few clicks were (N = 21) were in areas of the profile that contained information that was not being measured. These fell on 'see all friends', 'see all photos', 'about' button, 'posts' button, and 'messenger' button. The most common *Other* clicks are the content of the comments on posts, and the replies to these comments. This was found across all profiles regardless of type (i.e., fake or real). The popularity of these *Other* clicks suggests that despite an obvious reliance on the visual stimuli (*Photo Type*) as evidenced in both this study and previous Studies 1 and 2, participants are looking at the social connections and conversations on the profile when making their judgements.

To further understand the areas participants used to inform their judgements additionally analyses were conducted regarding the frequency of heatmap clicks in each area for each type of judgement (real or fake). Results of which are displayed in Figure 4.

Figure 4.

Graph showing the frequency of clicks for each judgement type per manipulated characteristics for all profile types.



It is evident in Figure 4 that participants clicked more frequently on the profile when judging that profile as real, this was particularly so when the profile was a *real* profile and a *0 Fakes* profile. This finding reinforces the linear trend outlined in Figure 1 – participants were most accurate when judging real profiles, and least accurate when judging 0 Fakes. This suggests that participants may click more so on the profile when they believe their judgement to be accurate.

Deceptive Purposes of Profiles

After viewing and judging each profile, participants were asked to indicate whether they thought the profile they had viewed was created with deceptive purposes or malicious intent. Figure 5 displays the frequency of yes/no judgements made in regard to the deceptive purposes of the profiles across all three conditions.

Figure 5

Graph showing the frequency of yes/no answers to the question 'Do you think this profile was created for deceptive purposes?' for all profile types

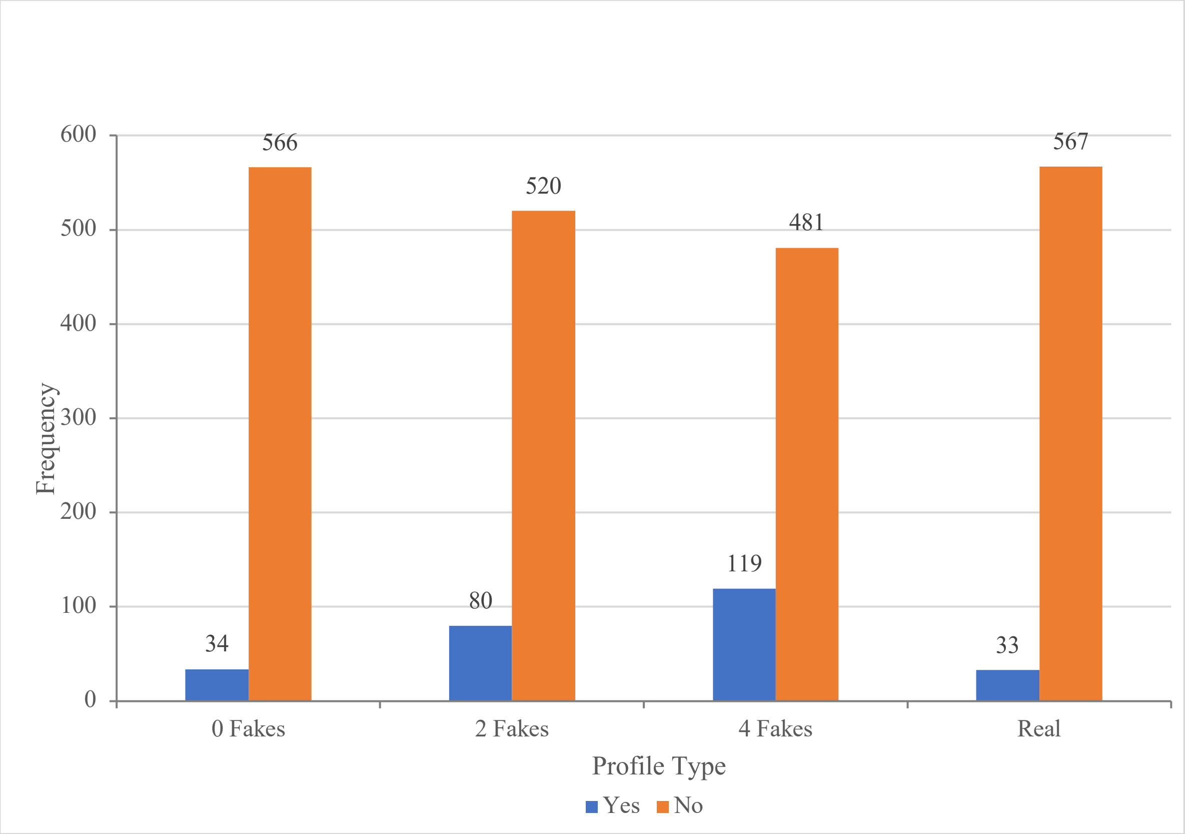


Figure 5 shows that participants predominately believed that the profiles were *not* created for deceptive purposes. Participants believed the real profiles to be the least deceptive, and in regard to the fake profiles believed profiles with four fake characteristics were most deceptive and the profiles with zero fake characteristics to be the least deceptive. It is interesting that the zero fakes characteristics profiles have the lowest rating of deceptive judgements, as this suggests these profiles are not only the hardest to accurately judge (as evidenced in Figure 1), but that they are also hardest to decipher as fake.

Discussion

This study addressed the question ‘can humans accurately detect online deception in the form of fake social media profiles?’. Overall, the results presented show that people are not good at accurately judging the authenticity of a Facebook profile. Interestingly, people are better at judging a real profile accurately than they are at accurately judging fake profiles. As has been found previously in Studies 1 and 2 (Chapters 2 and 3), the manipulated factor of *Photo Type* was relied upon heavily by participants when judging fake profiles, however, in contrast to previous findings, when judging real profiles, participants relied heavily on the Posts Content, with *Photo Type* being the second most relied upon factor. However, the factor of *Photo Type* was the only statistically significant predictor of both participants’ judgements and accuracy of their judgements, suggesting participants had an overreliance on the visual aspects of the profile. These findings may be due to several different reasons or theories. Such reasons will be outlined and analysed below.

Of great interest, the same linear trend in participants’ judgement accuracy was found, as in Studies 1 and 2 whereby participants achieved the highest mean accuracy score when judging real profiles, and the lowest accuracy score when judging fake profiles. Within the fake profiles, participants’ accuracy was highest when judging the profiles with four fake characteristics, followed by profiles with two fake characteristics, and lowest for profiles with zero fake characteristics. The replication of these results evidences the robustness of these findings, in that participants’ judgement accuracy is consistent for each type of profile regardless of whether they are presented with multiple types of different profiles (as in Study 1, Chapter 2), or only a single type of fake profile (as in Study 2, Chapter 3).

Consistent with the findings of Studies 1 and 2, *Photo-Type* determined participants' judgements and accuracy: when *Photo-Type* has been manipulated on the profile (i.e., photos of celebrities, landscapes, artwork, and cartoon characters) it increased the probability that participants would judge the profile as fake *and* increased the probability that said judgement would be accurate. The heatmap data also showed that this characteristic was also the most clicked on area when judging fake profiles, and second most when judging real profiles. The consistent finding of the effects of *Photo-Type* on participants' judgements and accuracy strengthens the conclusion of Studies 1 and 2, on the potential over-reliance on visual stimuli when making a judgement.

These results could be due to the fact the photos presented in the fake profiles are designed to be obviously manipulated (in some profiles) with images of celebrities, landscapes, artwork, and cartoon characters, suggesting that it may be more obvious to the participants which profiles are fake by looking at the type of photo alone. For example, a profile with a profile image of a celebrity would most likely be judged as fake as many would not expect a celebrity to have a normal Facebook profile, but rather a 'fan page' profile. Whereas, in the real profiles, the type of photo is either of the profile owner or a group photo including the profile owner (realistic, social photos), thus making them harder to accurately judge as real based on the type of photo alone. Hence, participants are potentially relying more on the verbal/written aspects of the profile when making their judgements for these real profiles.

Interestingly, and similar to the results of Study 2, *Posts Content* was the second-most clicked heatmap area of the profiles, and again was the most clicked on area when participants were judging real profiles. Several researchers have studied the psychology of linguistic cues, with some finding that a large quantity of words in an online environment increase levels of trustworthiness (Toma & D'Angelo, 2014), and others finding that liars tend to use more words when lying online than when lying in face-to-face settings (Zhou et al., 2004). One conclusion from this is that participants were relying on the content of the posts to make their judgement when the posts contained a significant amount of text. For example, a status update with a few sentences may be deemed as more trustworthy than a post that is sharing a link to a website or video and contains little to no text.

However, researchers have also found that users of online platforms expect that online self-descriptions may be deceptive, particularly on online dating websites where users are mainly concerned that potential romantic partners may be misrepresenting themselves (Cornwell & Lundgren, 2001). The expectation that this may occur suggests that participants would *not* rely on the content of the posts as much as they did to make their judgments. However, this research could relate more to the *Intro* and *Bio* sections of the profiles, as these are the more self-descriptive areas, in comparison to the daily thoughts and such like that tend to be present in the posts on the profile, thus providing support for the very few heatmap clicks found in these areas on any type of profile. In turn, this piece of research has provided further evidence as to the importance of the visual stimuli when making judgments of authenticity and may help explain further why participants rely more on the *Photo Type* characteristic than any others. So why is the opposite found in judgements of real profiles? Why is *Posts Content* relied upon most when judging real profiles?

This finding could be explained by the presence of a visible 'social network' on the real profiles (i.e., participants are using the connections in the posts, or the way other users are interacting on the posts), as evidence that the profile has a social network. For example, birthday messages posted on profile walls are hard to create or fake. This is because a whole network of fake profiles would need to be created, and actively used, to interact with each other to be able to post on each other's walls. Thus, the presence of these posts on the real profiles could be a clear indicator to some participants that the profile is real, as the participants are seeing a connection to multiple profiles or a social network. Those participants who use Facebook and so are familiar with it, of which there were many (N = 155, 77.5%), may have the understanding that such aspects of the profile would be very difficult to fake, meaning the presence of those on a profile offers strong evidence that a profile is real. The researcher recommends that more work is carried out to investigate the specific content contained within the posts and the comments that informs authenticity judgements rather than just attention, either with more capable heatmap software or eye-tracking technology.

Some researchers *have* used eye-tracking technology in the context of social networking sites, however these are not in relation to authenticity judgements but rather to understand attention, specifically the 'areas of interest' on a profile. For example, Vraga, Bode & Troller-Renfree (2016) simulated a Facebook newsfeed containing a plethora of different content (e.g., statuses, images, links, social posts, news) and found that richer content such as images and links, and the social content, enhanced attention. Scott & Hand (2016) found that different areas of a Facebook profile were focused on when the viewer had differing motivations; when viewing friends gaze was focused more so on the images content and areas that provide clues to the profile owners' personality, whereas gaze focus on the text content of the profile was related to viewing the profile of a potential employee. Whilst these studies do not provide direct evidence in relation to the profile areas on Facebook used to inform an authenticity judgement, they do provide some evidence that could explain the findings of this study in relation to an over-reliance on images (*Photo Type*). However, as mentioned, eye-tracking research could be expanded further with research in relation to authenticity judgements on Facebook.

This study also identified that participants are not overly concerned with the deceptive intentions of the profiles, a point which in itself could be considered as concerning: participants were aware they were viewing fake profiles and that they were potentially being deceived yet remained relatively trusting of said profiles. This could be explained by the truth default theory (Levine, 2014), in that people presume, as a default and without conscious reflection, that communication from others is honest. It could also be explained by the similar notion of *truth-bias* whereby research has found that people tend to believe others' communication as honest regardless of their actual honesty (Levine at al.,1999). It could be said that such bias is present here as participants self-reported that they believed the profiles to be trustworthy (answers of 'No' to deceptive intentions) even when they had been instructed that some of the profiles would be fake.

A revealing finding of this research is that no relationship was found between previous experience in creating fake profiles and participants ability to accurately identify fake profiles. This is an interesting finding in that it shows that even when participants had previous knowledge of creating a fake Facebook profile, and understood the mechanisms behind the profile and thus the areas that can be faked, they were still unable to accurately judge a profile as either real or fake, highlighting further the need for this research – what characteristics of a Facebook profile do human users look for and use to make their judgement? This also highlights an important issue; if previous experience has no effect on accuracy then what is needed to improve participants' accuracy? Perhaps training, or experimental exposure to a fake profile, could improve accuracy – a thought for future research.

One limitation of this study is that the heatmap clicks do not indicate the exact part of the text in the posts (*Posts Content*) that is influential, only the fact that this characteristic is used to inform judgements. For example, participants could be relying on the fact it is a "Happy Birthday" message, or they could be relying on a status update that contains information that corroborates with other information on the profile such as the profiles users' hometown or job shown in the Intro or Bio sections. This is a general limitation of using the heatmap software. To capture such specific data on 'Posts-Content', it would be necessary to categorise different types of posts, then create specific regions for each type to try and capture the clicks on the profile. However, the Qualtrics heatmap software does not allow for overlapping of regions, so a post with several of these distinguishable post content categories would not be accurately measured , thus bringing it back full-circle – the content of the posts can only be captured in the way in which it has been throughout this research. The results of this research so far do however suggest that the content of posts is an important feature of profiles when considering the authenticity of the profile, and thus is still a relevant contribution to literature despite its limitations.

A second, and possibly more significant, limitation is the lack of results evidencing the processes participants use when making their judgements, i.e., whether they make snap decisions on the authenticity, or whether they use more time to assess the whole profile before judgement. Kahneman (2011) referred to the mind as having two systems: System 1 and System 2. System 1 operates on an automatic and fast basis where little, or even no, effort is needed and there is no sense of voluntary control, whereas System 2 is a slower system that involves the use of concentration, choice, and agency to make decisions and choices. For example, System 1 can, amongst many other things, automatically detect hostility in another's voice, orient to the source of sudden noises, and understand simple sentences, or in other words System 1 is a set of innate skills. System 2 on the other hand is responsible for highly diverse and taxing operations such as filling out a form, focus on a sound of a particular thing in a noisy place, and checking the validity of an argument to name but a few. In relation to this research, it would be of interest to understand which of these systems are in play when judging the authenticity of social media profiles. Measuring for this will help further understand how judgements of online social media profiles are made, the cognitive processes used, and whether it really does matter what is written on a post if a snap decision is being made, perhaps one which is relying on the type of photo present on the profile, and area that has been evidenced to be highly influential over authenticity judgements. To further investigate the cognitive processes used when making judgments, the next study within this research will attempt to capture this by introducing a time-limit for a judgement of the profiles to be made in an effort to understand whether pressure of having to make a judgement in a short amount of time, thus using System 1, would have an effect on judgement accuracy.

Chapter 5: Study 4

This study aims to further understand the cognitive processes involved in humans' judgements of social media profiles; specifically, the factors they rely on when making authenticity judgements. In order to do this, a time-limit was introduced to measure whether accuracy was affected when an element of pressure was present, and whether participants use System 1 or System 2 when making their decision.

In his pioneering work on decision-making processes, Kahneman (2011) outlined a two-system approach of decision-making. According to Kahneman, the human brain uses two systems when processing information to form impressions and make decisions: System 1 and System 2. System 1 is fast, operating automatically with very minimal effort, and cannot be voluntarily controlled, whereas System 2 is slower, adopting a more considered approach to allocate attention to mental activities that require more effort (pg. 20-21). For example, tasks attributed to System 1 include detecting hostility in a voice, understanding sentences, orientation to sounds, and detecting distance between objects. When processing information using System 1, only the information immediately available is used, meaning the saliency of said information drives the decision-making process (de Castro Bellini-Leite, 2013). System 2 tasks are of a higher complexity, such as filling out a form, comparing values of two similar products, looking for a man with a white beard, or reciting a mobile number.

Both systems are continuously active during waking hours, with System 1 continuously producing suggestions for System 2 such as feelings, impressions, intentions, and intuitions. System 2 will either accept these, turning them into beliefs and/or actions or will be activated to increase cognitive effort when an error in System 1 is made (Kahneman, 2011, pg. 24). System 2 is labelled as the "lazy" system as it is effort averse meaning it will mostly adopt the suggestions of System 1 (Kahneman, 2011, pg. 64). Essentially, humans are predisposed to avoid using System 2, unless necessary, due to the amount of cognitive effort needed (Dennis & Minas, 2018). Researchers have also found that System 1 is most likely to influence human perceptions and judgements as our intuition, or immediate impression (information from System 1), is much easier to remember than the facts garnered from System 2 regarding the situation (Dennis & Minas, 2018).

In regard to social media, researchers have reported that social media use exacerbates our predisposition to utilise System 1 (Moravec, Kim, & Dennis, 2018) due to the hedonistic mindset of users (Johnson & Kaye, 2015). Hedonism is described as the "the doctrine that pleasure or happiness is the sole or chief good in life" (Merriam-Webster, n.d.), so when discussing hedonism in the context of social media, research has shown that users typically utilise social media for entertainment purposes (Johnson & Kaye, 2015) and are thus in a hedonistic mindset. When in such a mindset, System 1 is relied upon for intuitive judgements as users are not motivated to employ the cognitive effort needed for System 2 (Kahneman, 2003), and the ability to critically consider information is reduced (Cotte, Chowdury, Ratneshwar, & Ricci, 2006). Researchers have proffered that System 1 is utilised when on social media; 59% of Twitter users shared an article without opening the article and reading it first (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016), and Tony Haile (2014) reported that of the users who do open the article 55% close the external page down in under 15s, a time that has been argued to be too short for System 2 cognition to be activated.

Few researchers have investigated the amount of time needed for each System to activate during different tasks. In general, observed human behaviour should be considered as a combination of tasks that occur on different time scales (Kitajima &

Toyota, 2013). Newell (1990, p.122) considered such tasks in his early works on theories of cognition, positing four different bands that contain a number of timescales needed for different facets of human behaviour: biological, cognitive, rational, and social.

The biological band, with a timescale of 1ms – 10ms, is responsible for neurological processes such as the activation of neurons and the neural circuit. The cognitive band, ranging from 100ms – 10s, involves deliberate acts and operations such as understanding and acting accordingly upon the phrase "Please pass the salt" in one second, or solving a logic problem in 8 seconds (pg. 144). The rational band, associated with tasks to pursue goals, ranges from minutes to hours. It can be better understood in terms of rationality, specifically bounded rationality. Simon (1990) defined the term bounded rationality as the rational choice of the decision maker based on their cognitive limitations. In other words, when a human makes a decision, their rationality in relation to that decision is limited (bounded) by the knowledge they have at the time of the decision, meaning decision makers will often make a satisfactory decision that fulfils their criteria rather than an optimal decision whereby they undertake a cost-benefit analysis (Simon, 1956). Finally, the social band, with a timescale of days to months, can include significant social interaction tasks amongst humans. However, some researchers have criticised Newell's seminal work on cognition bands arguing that the social band takes place over a much shorter time span from minutes to hours. For example, Jackson (2018) argues that human communication in its entirety is a form of social action, and so the social band can include tasks such as text messaging, telephone calls, going for a date, or a wedding ceremony to name a few, all of which can occur in minutes or hours.

With specific relation to this study and System 1 and 2 processes, Kitajima and Toyota (2011) outline in their work on human processing with time constraints, that the biological and cognitive bands reside in System 1, and the rational, and social bands reside in System 2. Suggesting that any task that is automatic or takes under 10s can be referred to as System 1, and any task that ranges from more than 10s to hours, days, weeks, and months can be referred to as System 2. As mentioned previously, in relation to social media usage, users will utilise System 1 more often when online (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016; Tony Haile, 2014), suggesting perhaps that decision making in an online context is a process that takes 10s or less.

In relation to the time taken to make judgements, a possible explanation can be drawn from the thin-slicing literature. The term 'thin slice' is used to describe short snippets of social behaviour (of less than 5 minutes) that perceivers draw inferences from in regard to traits, states, or characteristics of persons or situations (Carney, Colvin, & Hall, 2007). A wealth of research into thin slicing as a concept has found that thin slicing judgements can be predictive of intelligence (Murphy, Hall, & Colvin, 2003), personality (Ambady & Rosenthal, 1993), and sexual orientation (Ambady, Hallahan, & Conner, 1999) amongst others. However, a general point of contention within the literature is whether the length of the slice (time) is related to accuracy. Some researchers have found that accuracy increases with exposure time (Blackman & Funder, 1998; Ambady et al., 1999), whereas in a meta-analysis of 38 studies on thin slice accuracy, Ambady and Rosenthal (1992) found no increase in correlations between thin slices of less than 30 seconds and thin slices of more than 300 seconds.

With a specific focus on social media, several researchers have found evidence for judgement accuracy based upon minimal information or short exposure times to stimuli; Stecher and Counts (2008) found that personality judgements of social media profiles made from condensed profiles (limited information) were equally as accurate as those made from full profiles. Similarly, Ivcevic and Amabady (2012) compared raters' personality judgements of Facebook users through either their full profile page or single pieces of information such as a profile picture alone. Their findings showed that rater accuracy was highest when making personality judgements based on the profile picture alone. Further, Turner and Hunt (2014) reported that even with very brief exposure times to the profile stimuli, a good consensus between raters when observing the personality traits of the Facebook profile owner was found. Each of these studies suggest that minimal information is needed to make an accurate personality judgement. However, there is a distinct lack of research regarding thin-slicing and authenticity judgements of social media profiles, meaning it is difficult to apply these findings to the context of this study.

Referring to economic literature on the effects of time pressures on decision-making, there are conflicting findings; some researchers have reported that participants perform worse in learning-based tasks when under time pressures (DeDonno & Demaree, 2008; Cella, Dymond, Cooper, & Turnbull, 2007), whereas others have found no effect of time constraints in learning tasks on participant's learning rates (Bowman, Evans, & Turnbull, 2005). Time limit interventions are also used methodologically within psychology. The literature reports that time pressures can reduce decision-making quality (Payne, Bettman, & Johnson, 1993), reduce the number of utilitarian judgements (judgements for the greater good) (Cummins & Cummins, 2012), and reduce accuracy of choice responses (Kocher & Sutter, 2006). However, in a meta-analysis of 26 studies, researchers Baron and Gürçay (2016) found that time is not a reliable predictor of moral judgements. Similarly, Tinghög et al. (2016) found no effect of time pressures on moral decision-making.

Based upon the literature discussed above it is expected that there will be a relationship between time taken to make a judgement and judgement accuracy (Hypothesis 1). However, due to the apparent conflictions in the literature and lack of consensus regarding time limits, System 1 and System 2 processing, and judgements of social media profiles, it is difficult to predict the nature of this relationship, hence the non-directional hypothesis. However, the analysis of the time limits, specifically time taken to make a judgement, can give us somewhat of an indication of whether participants use System 1 or System 2 processing, based on the literature regarding the 10s timescale discussed earlier.

Further, based on the consistent findings in the previous studies within this research, it is again expected that real profiles will be more accurately judged than fake profiles (Hypothesis 2), and profiles with the highest number of fake characteristics will be more accurately judged as fake, than profiles with fewer, or no, fake characteristics (Hypothesis 3). It is also expected that there will be a relationship between participants' self-reported confidence in accuracy of their judgements and actual judgement accuracy (Hypothesis 4), however due to the mixed results observed in previous studies this hypothesis remains non-directional.

In regard to the manipulated characteristics of the profiles, based upon the consistent findings from earlier studies it is expected that the manipulated characteristics will be significant predictors of whether participants judge the profiles to be real or fake (Hypothesis 5), and significant predictors of participants' judgement accuracy (Hypothesis 6). Specifically, it is expected that 'Photo-Type' will be the strongest predictor of both participant's judgements and accuracy of said judgements (Hypothesis 7).

Following suit from Study 3 (Chapter 4), the individual differences variables (personality type and social sensitivity) are not hypothesised to have an effect on participant's accuracy due to the minimal effects found thus far. As such, these variables will not be directly measured but rather controlled for when analysing the data. In addition, this study will also treat the social media variables in the same manner, as again, despite multiple methodological variations, minimal to no significant results have been found thus far. Results from both variables will be reported alongside the main analysis.

Method

Participants

An A-Priori power analysis of a repeated measures within subjects ANOVA, was conducted using G* Power (Faul et al., 2007) prior to data collection to determine the appropriate sample size for this study. The analysis indicated that a sample size of 16 participants would be sufficient to detect a medium effect size of $f = 0.25$, with an alpha level of $\alpha = 0.05$ and a power of $1-\beta = 0.80$. However, 200 participants were recruited. As outlined in Study 1 (Chapter 2) and Study 3 (Chapter 4) the reasoning behind this decision was to enhance the reliability and generalisability of the findings by reducing the errors associated with the estimates, improve the representativeness of the sample, and reduce the risk of Type 1 and Type 2 errors. Additionally, based on the results of both Study 1 (Chapter 2) and Study 3 (Chapter 4) where similar designs were employed, a similar sample size of 200 participants was chosen to maintain consistency across the studies and ensure results between each study were comparable.

A total of 203 participants were recruited via Prolific and completed the study using Qualtrics. These participants were aged 18 to 58 years (*M* = 26.21; SD = 7.03). Of the 203 participants, 106 identified as Female (52.2%), 95 identified as Male

(46.8%), and 2 identified as Non-Binary (1.0%). 111 participants (54.7%) identified their ethnicity as White, 38 (18.7%) identified as Black, African, Black British or Caribbean (Includes any Black background), 25 (12.3%) identified as Another Ethnic Group, 19 (9.4%) identified as Mixed or Multiple Ethnic Groups (Includes any mixed ethnic background), 9 (4.4%) identified as Asian or Asian British (Includes any Asian background, e.g. Bangladeshi, Chinese, Indian, Pakistani and any other Asian Background), and 1 (0.5%) selected Prefer not to say.

Design

To investigate the hypotheses, a 4 (Real profiles, 0 Fakes profiles, 2 Fakes profiles, and 4 Fakes profiles) x 2 (Accurate judgement vs. Non-accurate judgement) x 2 (Type 1 decision time vs. Type 2 decision time) repeated measures experimental Turing test design was used. The Dependent Variable (DV) is the accuracy of the judgements, and the Independent Variable (IV) is the Facebook profiles. Both variables are within subjects' measures following a repeated measures design.

Measures, Materials, Equipment

Measures

As per all previous studies, the same three self-report measures were administered to participants online using Qualtrics software; a social media questionnaire , the Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow & Swann, 2003), the Social Sensitivity (SS) scale (Riggio, 1986). In addition, a set of questions were included to measure the time limit given to participants. The first question read 'To what extent did you feel that the amount of time you were given to view the profiles and make your judgement was adequate?' and was answered on a 7-point Likert scale from 'Totally inadequate' (1) to 'Perfectly adequate' (7). The second question read 'To what extent did you feel rushed when viewing the profiles and making your judgement?', also answered on a 7-point Likert scale from 'Extremely rushed' (1) to 'Not rushed at all' (7). Additionally, the question: 'Do you think any of the profiles were made with deceptive intentions?' was included in the follow-up questionnaire, as in all previous studies 1-3, to capture participants' overall opinions of the profiles they viewed. Participants were shown this question throughout the questionnaire after they had viewed and judged each profile, however many participants chose not to enter qualitative data when asked. To try and capture this the same question was also added to the follow-up questionnaire to encourage participants to provide a response if they had not already done so.

Materials

As in Studies 2 & 3, 74 profiles were used. The 74 profiles consisted of: 6 real profiles, 12 fake profiles with '0 fake characteristics', 21 fake profiles with '2 fake characteristics', and 35 fake profiles with '4 fake characteristics'. The number of fake profiles in each category is based upon the number of different combinations of manipulated characteristics for each condition (*0 fakes, 2 fakes, 4 fakes*) (see Appendix P for characteristic framework).

In preparation for analysis, each participant was given an accuracy score for their profile judgements based on how many profiles they accurately judged as real or fake. The accuracy score was split across the four different types of profiles: *0 Fakes accuracy* (maximum score of 3), *2 Fakes accuracy* (maximum score of 3), *4 Fakes accuracy* (maximum score of 3), and *'Real' accuracy*(maximum score of 3). Additionally, an *overall accuracy* score for each participant was calculated across all types of profiles, giving a maximum score of 12.

Procedure

Prior to conducting this study, a pilot study was undertaken to get an idea of an appropriate time limit to be used in the main analyses and test the efficacy of the slightly different procedural changes. The procedure and results of the pilot will be presented below, followed by the procedure and results of the main analyses.

Pilot

The overall procedure of the pilot is near identical to that of Study 3 whereby participants completed the study on Qualtrics once they had been recruited from Prolific and redirected to the platform. Once consent had been given, participants completed three of the self-report measures (social media questionnaire, TIPI, SS Scale) at the beginning of the study. Following this, participants entered the 'profile phase' of the study. Participants were first given a set of instructions in relation to the procedure for viewing the Facebook profiles and informed that they were to view each profile and make a judgement as to its authenticity. Participants were also informed that each profile screenshot was a heatmap style question, and that they were to click on the areas of the image they used when making their judgement.

Where this study differs to that of our previous studies, is that during the profile phase, participants were given a time limit of 45 seconds per profile to complete their judgement. They were informed of the time limit in the instructions given prior to the 'profile phase', and were also able to see the timer, counting down from 45 seconds, underneath each profile. Within those 45 seconds, participants were required to view the profile, make their judgement, and then click on the areas of the profile they used/relied upon to make their judgement. The timer was introduced to measure participants' decision-making processes, i.e., whether they were Type 1 (fast, instinctive), or Type 2 (slow, deliberate, considered). The time was set at 45 seconds based upon analysis of the previous studies. The data from previous studies only provides 'total questionnaire completion times' rather than individual times for each profile. To obtain a rough baseline figure, calculations of the time taken across all three previous studies were conducted, finding that participants took 24 minutes 51 seconds on average to complete the study. Based upon estimations of how long the 'non-profile' sections of the study (i.e., the consent forms and questionnaires at the beginning and end of the study, and the demographics questions), would take to complete, and the overall average time taken, a total of 80 seconds per profile was calculated. To ensure there was an element of pressure for participants, and to measure their decision-making processes, 45 seconds was selected for this study. However, it was unknown if participants would find this too difficult and restrictive, thus effecting the accuracy of their judgements, hence the need for this pilot study using the 45 second limit.

Thirty-six participants were recruited for the pilot via Facebook, Instagram, Twitter, E-mail correspondence, and word of mouth. Eighteen participants failed to complete the task. These cases were removed, leaving 18 participants for the final analysis. The 18 participants were aged between 20 and 62 years (*M* = 35.61; SD = 12.56). Sixteen identified as Female (88.89%), and 2 identified as Male (11.11%), 1 participant identified their ethnicity as 'Asian or Asian British' (5.55%), and the remaining 17 participants identified as 'White, including any White backgrounds' (94.45%).

Findings showed the same linear trend in accuracy whereby *Real* profiles were the most accurately judged, and *0 Fakes* the least. In regard to time, mean time taken to make a judgement per profile type were; 41.51s (SD = 19.91) for *0 Fakes,* 50.36s (SD = 40.09) for *2 Fakes*, 41.93s (SD = 42.94) for *4 Fakes*, and 40.62s (SD = 16.45) for *Real*, with an overall average of 43.61s (SD = 4.53) across all profiles. To test whether time had a significant effect on judgement accuracy, a repeated measures ANCOVA (Analysis of Co-Variance) test was conducted with accuracy scores for each profile type as the within subjects' factor, and mean time (overall average across all profile types) as the co-variate. Results showed that no statistically significant two-way interaction between time and judgement accuracy was found; $F(2, 48) = 1.904$, $p = .141$, $n^2 = .106$.

When participants were asked whether they felt the time given was adequate to view the profiles and make a judgement, most participants (N = 6, 33.3%) selected 'Inadequate', and when asked whether they felt rushed most participants selected either 'Very rushed' (N = 4, 22.2%) or 'Rushed' (N = 4, 22.2%). These findings suggest that participants need more than the 45s-time limit to make their judgments. However, this may be due to the fact participants were required to make their judgements and then also click on the areas of the profile they used to make the judgement.

Main Analyses

The procedure follows that of the pilot. The only changes that have been made are based on the results of the pilot and in relation to the timed portion of the study. During the profile phase of the study in the pilot, participants were required to view the profile, make their judgement, *and* click on the areas of the profile they used to make their judgement within the 45 seconds time limit. Based on the mean time overall in the pilot (43.61s), the finding that participants felt they did not have enough time to make their judgements, and the lack of a statistically significant effect of time on judgement accuracy, this study reduced the time limit to 40 seconds to ensure that an element of pressure was applied. This was to allow for measurement of participants' decision-making process (Type 1/Type 2) – reducing the time limit slightly may lead to more reliance on fast, snappy, Type 1 decision-making and therefore less reliance on the slower, more deliberate decision-making process (Type 2). Further the reduction in 5s is not expected to be unachievable as it is only 3.61s less than the overall mean found in the pilot.

A further procedural change was to the instructions given to participants prior to the profile judgement phase of the study and to layout of the questions within the profile phase. In the pilot, participants were presented with a profile, underneath which was the timer and the question asking if they judge the profile to be real or fake, followed by the instructions to click on the areas of the profile they used when making their judgement. Once they had completed these tasks, the following page asked participants whether they thought the profile was deceptive, and whether they needed any more information to make their judgement (Appendix M). However, during this study, participants were first presented with a profile and the timer *only.* Once the 40 seconds was over, or when the participant had chosen to move forward, the participants were shown a separate page asking them to judge the authenticity of the profile they viewed on the previous page. Once a judgement had been made, participants were then shown the same profile again and were instructed to click on the image on the areas that they relied upon or used when making their judgement. Following this page, participants were asked two multiple choice questions of whether they thought the profile was deceptive, and whether they thought they needed more information to make their judgement (Appendix M).

The timed element was only included on the first page showing the profile. This was to ensure the timing data was only encapsulating the time taken to make the judgement rather than the time taken to make the judgement, *and* click on the areas of the profile, as it was in the pilot. Additionally, participants had no option to go back and view the profile again *before* making their judgement. They were only shown the profile for the second time *after* their judgement had been made.

The instructions given to participants were edited to reflect the above procedural changes, including informing them that there is a time limit, the order in which each judgement type question will be presented to them, and further details outlining that they will only be timed whilst viewing the profile and not whilst making their judgement or clicking on the heatmap image. See Appendix T for the updated instructions.

Ethics

This research was fully approved by the ethics committee at Lancaster University under a further amendment to the original ethics submission for studies 1,2 and 3. All participant data was stored on a secure hard drive, in line with GDPR guidelines, and only accessible to the researchers.

Results

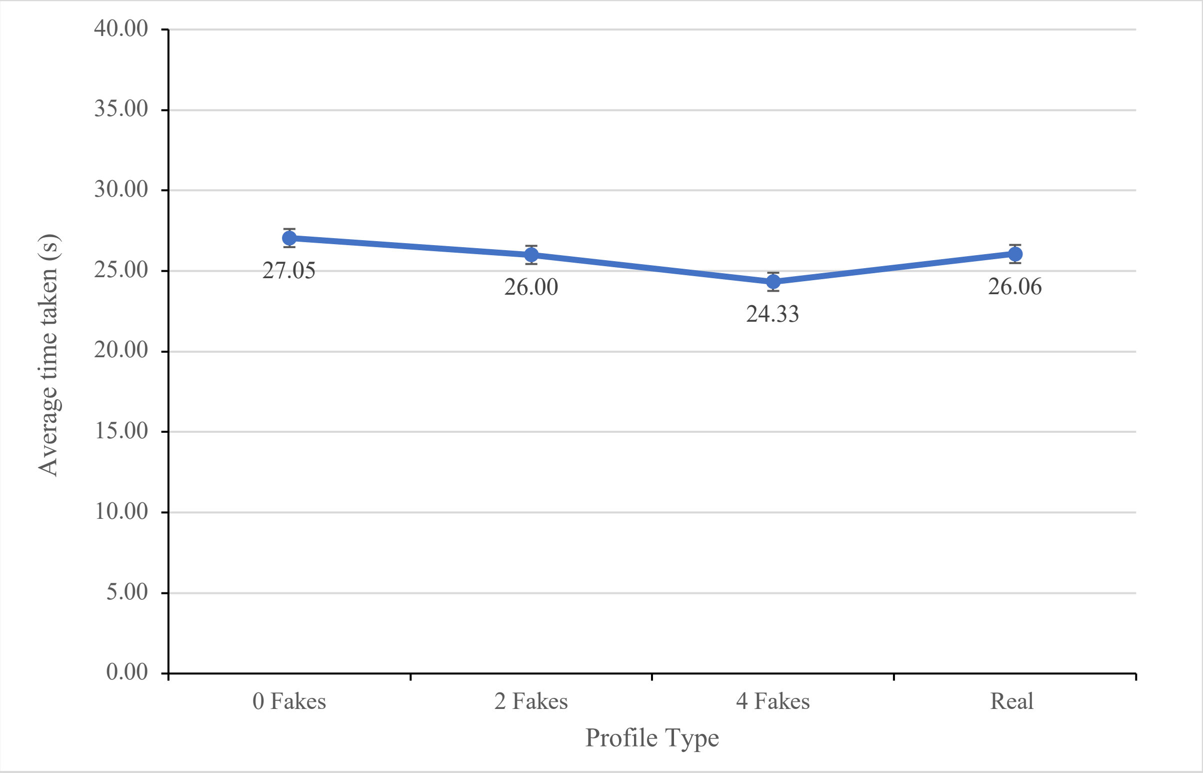
Raw data was exported from Qualtrics via Excel where it was sorted and coded prior to importing into IBM SPSS Version 27.0 for Mac and R Version 1.4.1564 (R Core Team, 2022) for analysis. There were no incomplete participant entries, so a total of 203 participants' data were analysed for this study.

Prior to conducting the statistical analyses, multiple normality tests were conducted to assess the appropriateness of the data. The normality tests showed the data had outliers, of which none were extreme, histograms showed mainly normal distributions with only few with a slight positive or negative skew, and all Q-Q plots showed data of a linear pattern. Overall, these tests concluded that the data is normally distributed and suitable for analysis.

Time Limit

On average, participants spent 25.86s (SD = 9.24) judging the authenticity of the Facebook profiles, with a wide range of 4.33s to 40.11s. Figure 1 displays the mean time taken to judge each type of profile.

Figure 1
Mean time taken, in seconds, for judgements of each type of profile (*N* = 203).



As evidenced in figure 1, participants spent the most time looking at the *0 Fakes* profiles when forming their judgements (*M* = 27.05, SD = 11.08), and the least time when looking at the *4 Fakes* profiles (*M* = 24.33, SD = 10.84). Interestingly, participants spent the same amount of time judging *2 Fakes* and *Real* profiles.

To test whether time had a significant mean effect on judgement accuracy, a repeated measures ANCOVA (Analysis of Co-Variance) test was conducted with accuracy scores for each profile type as the within subjects' factor, and mean time (overall average across all profile types) as the co-variate. When testing for sphericity, Mauchly's Test of Sphericity was violated ($X^2(5) = 15.41, p = .009$). To correct this the Greenhouse-Geisser correction was applied to the degrees of freedom and will be used when reporting the statistics. Results showed that no statistically significant two-way interaction between time and judgement accuracy was found; $F(2.857, 574.237) = 1.56$, $p = .199$, $n^2 = .008$.

Further, correlations were conducted to examine whether there were any relationships between average time taken to make a judgement and judgement accuracy for each profile type. Linear regressions were also conducted to analyse whether time spent judging the profiles can significantly predict judgement accuracy. An average time was given for all profiles per profile type, meaning participants had four separate average times (e.g., an average time to view the three *0 fakes* profiles is the *0 Fakes* time, the three *2 Fakes* profiles is *2 Fakes* time etc). Judgement accuracy for each different profile type was a score with a maximum of 3. Results of these analyses are presented in Table 1.

Table 1.
Correlations and linear regressions between average time taken to make a judgement and judgement accuracy, for each profile type (N = 203).

Variables	*M*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1. *0 Fakes* Accuracy [a]	*0.46*	*0.64*	-							
2. *2 Fakes* Accuracy [a]	*1.00*	*0.76*	*-.01*	-						
3. *4 Fakes* Accuracy [a]	*1.62*	*0.85*	*-<.001*	*.19**	-					
4. *Real* Accuracy [a]	*2.54*	*0.63*	*.05*	*-.09*	*-.09*	-				
5. *0 Fakes* Average Time [a]	*27.05*	*11.08*	*.17**	*-<.001*	*-.07*	*.05*	-			
6. *2 Fakes* Average Time [a]	*26.00*	*11.02*	*.08*	*-.04*	*-.07*	*.09*	*.68***	-		
7. *4 Fakes* Average Time [a]	*24.33*	*10.84*	*.04*	*-.11*	*-.16**	*.08*	*.59***	*.65***	-	
8. *Real* Average Time [a]	*26.06*	*10.47*	*.09*	*-.02*	*.03*	*-.08*	*.65***	*.61***	*.61***	-

Predictors	0 Fakes Accuracy [b]			2 Fakes Accuracy [c]			4 Fakes Accuracy [d]			Real Profile Accuracy [e]		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Average time per profile type	.030	0.01*	0.04	.002	-0.03	<.001	.026	-0.01*	<.001	.006	-<.001	<.001

Note. [a] *df* = 195, [b, c, d, e] *df* = 1, 201. **p < .05,* ***p* <.01.

Table 1 shows that there is a significant small positive relationship between average time spent judging *0 Fakes* profiles and judgement accuracy of *0 Fakes* profiles ($r(195) = .17$, $r^2 = .030$), meaning that time spent judging *0 Fakes* explained only 3% of the variability in judgement accuracy scores of *0 Fakes,* and an increase in time resulted in a significant increase in 0 Fakes accuracy scores of $B = 0.01$ ($F(1, 201) = 6.27$, $p = .013$).

Additionally, a significant small negative relationship was found between time taken judging *4 Fakes* and judgement accuracy of *4 Fakes* ($r(195) = -.16$, $r^2 =.025$), meaning that time spent judging *4* Fakes explained only 2.5% of the variability in judgement accuracy of *4* Fakes, and an increase in time taken to judge the profiles resulted in a significant decrease in *4 Fakes* accuracy scores of $B = -0.01$ ($F(1, 201) = 5.32$, $p = .022$). There were no significant relationships found between *2 Fakes* time and accuracy ($F(1, 201) = 0.41$, $p = .523$) or *Real* time and accuracy ($F(1, 201) = 1.20$, $p = .275$). These findings suggest that time taken to judge the profiles does have an effect on judgement accuracy, albeit a mixed effect.

Also shown in Table 1 are strong positive correlations between the average times taken for judgements of each type of profile. These findings suggest that time taken to judge profiles is consistent across each profile type, and that an increase in time for one profile type is likely to lead to an increase in time for other profile types.

Participants were asked two questions at the end of the study related to the time limit. The first question asked was '*To what extent did you feel that the amount of time you were given to view the profiles and make your judgement was adequate?'* and was recorded on a scale of *Totally inadequate* (1) to *Perfectly adequate* (7). The vast majority of participants reported that the time of 40s was either *Adequate* (N = 64, 31.5%) or *Perfectly adequate* (N = 50, 24.6%). Only one participant (0.5%) reported feeling that the time given was *Totally inadequate.* The second question asked was '*To what extent did you feel rushed when viewing the profiles and making your judgement?*' and was recorded on a scale of *Extremely rushed* (1) to *Not rushed at all* (7). Most participants reported that they were *Not rushed at all* (N = 77, 37.9%), or *Slightly rushed* (N = 44. 21.7%). Only 2 participants (1%) reported feeling *Extremely rushed.* Meaning, most participants made their judgements without the feeling of being pressured by time.

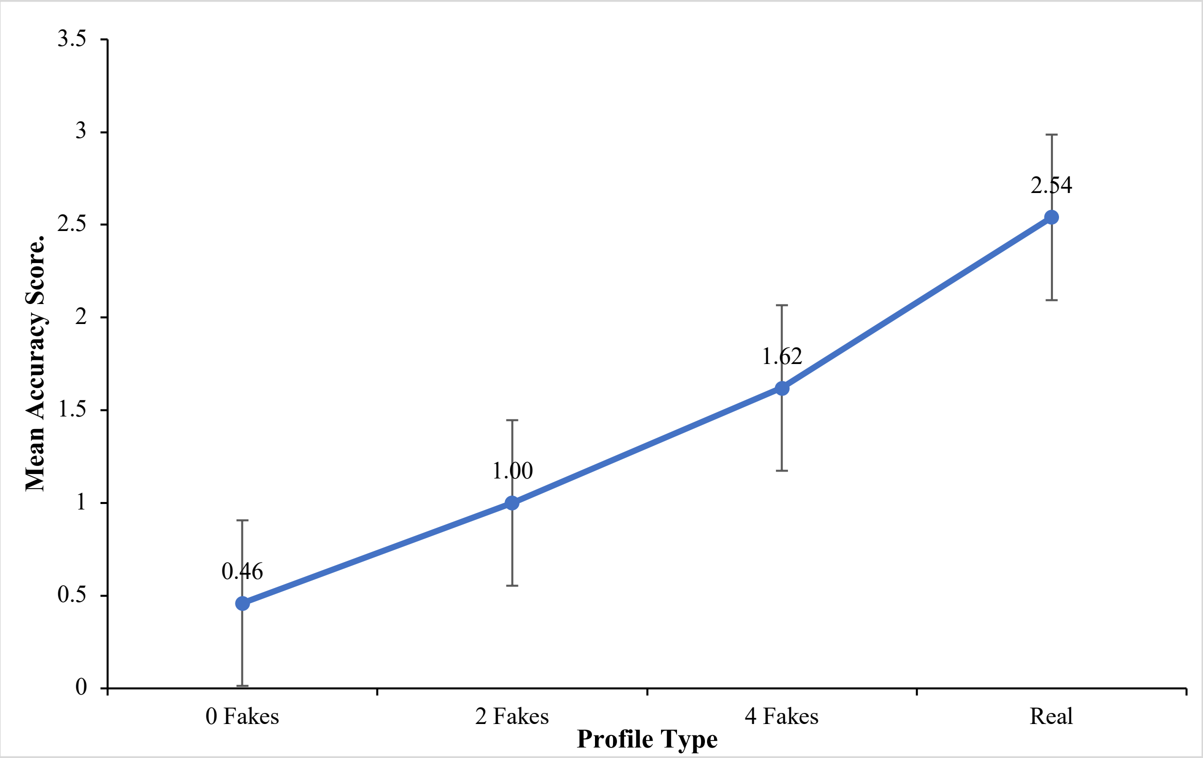
It has been evidenced from this analysis that are some relationships between time taken to make a judgement and judgement accuracy, Whilst some significant relationships were found between some of the variables, the results of the ANCOVA showed there was no significant main effect of time on judgement accuracy, therefore H1 can only be partially accepted.

Based upon the literature discussed regarding the 10s timescale used when on social media, exploratory analysis was conducted on participants whose average time was less than 10s to investigate this theory further. Of the 203 participants, only six (2.9%) had an average time lower than 10s. The total accuracy scores of these participants ranged from 3-6, which is less than or equal to the level of chance (6), and the majority of those were accurate judgements of fake profiles (69.0%). No significant effect of time spent on judgements and judgement accuracy was found for these participants; $F(3, 9) = 1.74$, $p = .229$, partial $n^2 = .367$. Additionally, these participants clicked on the profiles a total of 176 times, with 46 (26.1%) of those being on the *Real* profiles. Across all profile types *Posts Content* had the highest frequency of clicks (N = 53, 30.1%), followed by *Photo Type* (N = 49, 27.8%) and *Other* (N = 47, 26.7%).

Profile Accuracy

Mean judgement accuracy scores of fake and real profiles were analysed and compared. Overall, participant's mean judgement accuracy score across all profile types was 5.61 (SD = 1.48). Figure 2 displays mean scores for each type of profile.

Figure 2
Mean judgement accuracy scores for each type of profile (*N* = 203).



As shown in Figure 2 , a linear trend was found in participants' judgement accuracy: the more fake characteristics within the profile, the more accurate participants' judgements. Participants' accuracy was highest when judging real profiles (*M* = 2.54, SD = 0.63). In regard to the fake profiles, participants accuracy was lowest when judging 0 Fakes profiles (*M* = 0.46, SD = 0.64) and highest when judging profiles with 4 fake characteristics (*M* = 1.62, SD = 0.85). The linear trend found in accuracy, specifically accuracy of fake profiles, is mirrored in the average time taken to make a judgement (as shown in Figure 1). Participants spent more time viewing the profiles they were least accurate at judging correctly (*0 Fakes)* and less time viewing the profiles they were more accurate at judging (*4 Fakes)*.

To analyse whether participants' accuracy was better than that of chance, participants' scores for each type of profile were combined to give them an overall accuracy score out of 12. A one-sample t-test showed a highly statistically significant mean difference,-.39, $t(202) = -3.79$, $p < .001$, meaning that participants' accuracy ($M = 5.61$, SD = 1.48) was statistically lower than that of chance ($M = 6.0$).

Maximum judgement accuracy was also measured to analyse the number of participants who correctly judged all three profiles within each type of profile correctly, and whether any participants could be identified as a super-recogniser by scoring a maximum score of 12. A further linear trend was found for maximum judgement accuracy for each type of profile; 0.49% (N = 1) of participants accurately judged all 0 fakes profiles as fake, 2.96% (N = 6) accurately judged all 2 fakes profiles as fake, 15.27% (N = 31) accurately judged all 4 fakes profiles as fake, and 60.59% (N = 123) accurately judged all real profiles as real. Collectively, zero participants achieved the maximum score of 12 . Only one participant (0.49%) achieved a score close to maximum with a score of 10, which suggests, alongside the linear trends found, that participants' judgement accuracy is not consistently accurate across each type of profile.

To investigate further any interactions between the related means of the Independent Variable with four levels of profile; 0 Fakes, 2 Fakes, 4 Fakes, and Real, and the Dependent Variable with two levels of accuracy; Accurate, and Non-Accurate, a 4x2 repeated measures within-subjects ANOVA (Analysis of Variance) was conducted in SPSS. The results showed that Mauchly's Test of Sphericity was violated $x^2(5) = 16.07$, $p = .007$. Consequently, the Greenhouse-Geisser correction was used to correct the degrees of freedom. The results from the ANOVA showed a highly significant effect of the type of profile on participants' judgement accuracy, *F*(2.85, 575.91) = 310.89, p <.001, partial n^2 = .606. Additionally, pairwise comparisons show a statistically significant mean increase in accuracy from 0 Fakes to 2 Fakes (0.58, 95% CI [0.35, 0.72], $p < .001$), 2 Fakes to 4 Fakes (0.62, 95% CI [0.43, 0.81], $p < .001$), and 4 Fakes to 'Real' (0.92, 95% CI [0.71, 1.13], $p < .001$).

As in the previous studies, Signal Detection Theory (SDT) was again utilised to further understand participants' judgement process. Fake profile accuracy scores and real profile accuracy scores were transformed into hit rate and false alarm scores. Hit rate was calculated by dividing the number of hits (number of accurate judgements) by the number of signal trials (possible correct judgements), and the false alarm rate was calculated by the number of false alarms (inaccurate judgements) divided by the number of noise trials (the total number of signal trials incorrectly identified as noise trials). This calculation gave a hit rate and false alarm rate for both types of profile (fake and real). From these a d-prime (*d'*) value and criterion (*c*) scores were calculated - *d'* is a sensitivity measure used to indicate participants' abilities at distinguishing between fake profiles (signals) and real profiles (noise), and *c* is a measure of response bias, specifically whether participants had a stronger tendency to say yes or no (real or fake). Findings show that overall, participants were able to distinguish signals (fake profiles) from the noise (real profiles), *d'* = 0.57, 95% CI [2.88, 3.26], meaning participants were able to identify fake profiles as fake. Participants also showed a bias to judging the profiles as fake with a *c* score of -0.76.

Overall, the tests conducted in regard to profile accuracy show that real profiles were judged more accurately than fake profiles, and fake profiles with a higher number of fake characteristics were judged more accurately than those with fewer, or no, fake characteristics. As such, H2 and H3 can be accepted.

Self-Reported Accuracy

Participants were asked to self-report the level of confidence in the accuracy of their judgements on a scale from 1 (Unconfident) – 7 (Confident), with 'Neutral' in the middle. Most participants reported feeling either *Slightly Confident* in their judgements (*N* = 57, 28.1%) or *Moderately Confident* (*N* = 52, 25.6%. Only eight participants (3.9%) reported feeling *Unconfident*.

To understand whether there is a relationship between participant's self-reported accuracy and actual judgement accuracy a multiple regression was conducted with accuracy scores for each profile type. To test the appropriateness of the data for the regression, assumption tests was conducted prior to conducting the regression models. All assumptions (linearity, normality, independence of observations, homoscedasticity, and multicollinearity) were met. No significant coefficients were found between any of the self-reported accuracy predictors and judgement accuracy, for any profile type. Additionally, each regression model was also non-significant; *0 Fakes*, (*F*(6, 196) = 1.42, *p* = .209*), 2 Fakes, (F*(6, 196) = 0.86, *p* = .529), *4 Fakes, (F*(6, 196) = 0.56 , *p* = .766), and *Real (F*(6, 196) = 0.42, *p* = 868). Thus, it can be concluded self-reported accuracy is not a good predictor of participants' actual accuracy as no statistically significant relationships were found. Therefore, H4 cannot be accepted.

Manipulated Characteristics of Profiles

Each manipulated characteristic (*Photo Type, Photo Number, Bio, Intro, Posts Content, Number of Comments, Number of Likes*) was entered into a general linear model ('*glmer')* in R (utilising the '*lme4'* package), to understand whether the manipulated characteristics of the profiles had an effect on participants' judgements and accuracy of their judgments. Both judgement and accuracy models were first conducted using ‘Prolific ID’ as a random effect, and then conducted again with ‘Prolific ID’ and ‘Profile Number’ as a nested random effect. The addition of ‘Profile Number’ as a nested effect statistically significantly model improved the fit of the model at *p* <.001 level, and this was the case for all models conducted, thus the models reported below include both ‘Prolific ID’ with ‘Profile Number’.

Model’s 1 and 2, reported in Table 3 measured participants judgements and *Overall Accuracy* scores against the manipulated characteristics predictors.

Table 3.
Results from 'glmer' Model's 1 & 2 where judgement and accuracy are regressed on the manipulated profile characteristics.

Predictors	Model 1 – Judgement [b]					Model 2 – Accuracy [c]				
	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*	
Fixed effects										
Intercept	-1.95	0.17	-2.29	-1.61	<.001***	-0.88	0.28	-1.43	-0.34	.001**
Photo Type [a]	2.34	0.22	1.91	2.78	<.001***	2.09	0.34	1.42	2.76	**<.001*****
Number of Photos [a]	0.30	0.22	-0.12	0.72	.164	-0.03	0.34	-0.70	0.64	.929
Bio [a]	0.27	0.22	-0.15	0.70	.209	0.01	0.34	-0.66	0.68	.980
Intro [a]	0.24	0.22	-0.19	0.67	.266	-0.03	0.34	-0.70	0.64	.923
Post Content [a]	0.46	0.22	0.04	0.89	.033*	0.13	0.34	-0.54	0.80	.704
Number of Comments [a]	0.07	0.22	-0.36	0.50	.763	-0.29	0.34	-0.97	0.34	.399
Number of Likes [a]	0.09	0.22	-0.33	0.52	.670	-0.20	0.34	-0.87	0.47	.550
Random effects										
Residual Variance (σ^2)	3.29					3.29				
τ_{00} PROLIFICID	0.26					0.11				
τ_{00} PROFILENUM	0.47					1.55				
Intraclass Correlation Coefficient	0.18					0.33				

Note. Number of Participants = 203, Number of Profiles = 74, Number of Observations =2436 . **p* =.05, ** *p* =.01, *** *p*<.001.
[a] Model 1: 0 = Judgement of Real, 1 = Judgement of Fake; Model 2: 0 = Non-Accurate Judgement, 1 = Accurate Judgement. [b] Conditional R^2 = .383 [c] Conditional R^2 = .417

Table 3 shows that *Photo Type* is highly statistically significant predictor of both participants' profile judgements and accuracy of said judgements. When *Photo Type* had been manipulated on a profile, participants were more likely to judge that profile as fake (B = 2.34) and the judgement of fake is more likely to be an accurate judgement (B = 2.09). The manipulated characteristics *Posts Content* is also a significant predictor of participant judgement; when this characteristic is manipulated on the profiles participants are more likely to judge that profile as fake (*B* = 0.46). *Posts Content* is not a significant predictor of participants' judgement accuracy, meaning that even though participants are more likely to judge the profile as fake when *Posts Content* has been manipulated, this judgement is not necessarily an accurate judgement.

To further understand whether time taken to judge each different profile, and the different manipulated characteristics had an effect on participants' judgements and accuracy of judgements, the two models reported above were conducted again with the addition of *Time Taken Per Profile* as a fixed effect alongside the manipulated characteristics. Time taken to judge each profile was not a significant predictor of participants' judgements; *B* = 0.006, *SE* = 0.005*, p* = .202*, 95% CI* [-0.003, 0.015]*,* or judgement accuracy; *B* = -0.001*, SE* = 0.004*, p* = .727*, 95% CI* [-0.010, 0.007].

Figure 3.
Lens model diagram showing estimates and statistical significance of the manipulated characteristics for models 1 and 2.

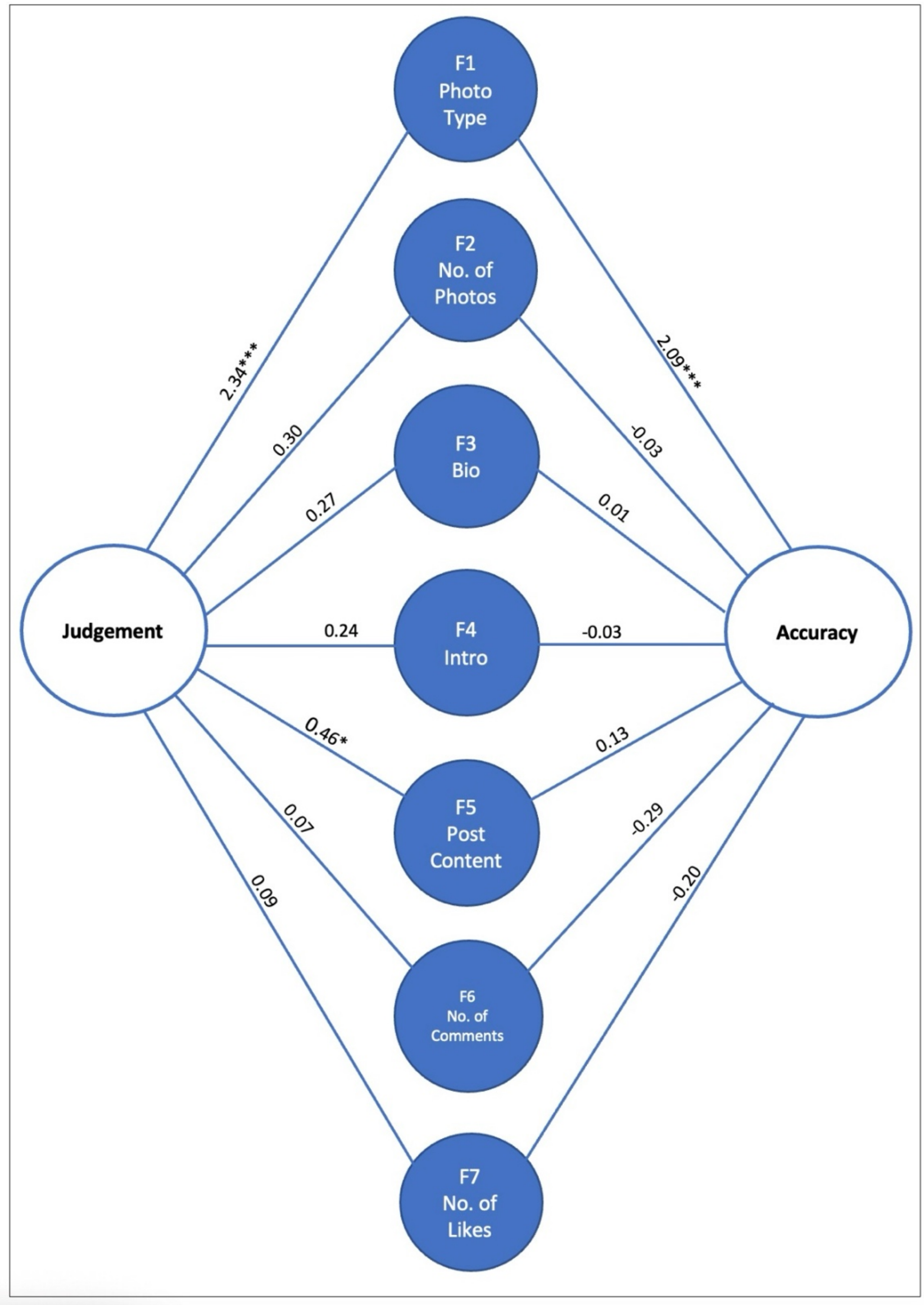


Figure 3 displays the estimates for both models in a Brunswik Lens Model diagram, whereby the manipulated characteristics are the available cues and the participants’ judgements (whether the profile is real or fake) are the observed decision, and the judgement accuracy is the correct decision.

The results from these models provide partial support for H5 and H6 as some of the manipulated characteristics were significant predictors of both participants' judgements and accuracy of said judgements, meaning H5 and H6 can be accepted. Hypothesis 7 stated that *Photo Type* will be the strongest predictor of both judgement and judgement accuracy. The results above show this to be true meaning H7 can be accepted. The additional models in regard to the effect of time taken on each profile and participants' judgement and judgement accuracy did not find any relationship, thus do not provide any additional support for H1.

As mentioned previously, this study did not hypothesise any relationships between individual differences variables (personality traits as measured by the TIPI and social sensitivity) or social media variables and judgement accuracy, due to the very minimal, or non-existent, relationships found in all previous studies in this research. These variables were still controlled for to ensure the analyses reported above were not confounded by these variables. The results of which are outlined below.

Personality

To investigate whether there were any significant relationships between personality and social sensitivity on judgement accuracy, four multiple regression models were conducted – one for each type of profile (*0 fakes* accuracy, *2 fakes* accuracy, *4 fakes* accuracy, and *real* accuracy), with the TIPI trait scores and scores on the Social Sensitivity (SS) Scale as predictors.

All assumptions of the multiple regression test (linearity, homoscedasticity, independence of residuals) were tested prior to analysis, and all were met. Table 4 presents the results of the multiple regressions between the individual differences variables and judgement accuracy scores for each type of profile.

Table 4.
Multiple regression for personality predictors of judgement accuracy for each type of profile

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
TIPI	.011			.026			.006			.019		
Extraversion		0.03	0.03		0.03	0.04		0.03	0.04		-0.05	0.03
Agreeableness		-0.01	0.04		-0.10*	0.05		<.001	0.06		0.01	0.04
Conscientiousness		-<.001	0.04		0.04	0.05		-<.001	0.05		-0.03	0.04
Emotional Stability		0.03	0.04		-0.01	0.05		-<.001	0.05		-<.001	0.04
Openness to New Experiences		-0.01	0.04		0.03	0.05		-0.01	0.06		0.02	0.04
SS Scale		<.001	<.001		<.001	<.001		<.001	<.001		<.001	<.001

Note. df = 6, 196. *p < .05

Table 4 shows that the individual differences variables are not good predictors of judgement accuracy for any profile type. One predictor, *Agreeableness,* was a significant predictor of participants' judgements of *2 Fakes* profiles, meaning that an increase in the *Agreeableness* trait (levels of kindness, cooperativeness, friendliness and politeness) is associated with a decrease in accuracy score (*B* = -0.10). However, the overall regression model for *2 Fakes* judgement accuracy was non-significant, *F*(6, 196) = *0.86, p* = .528. Similarly, the regression models for remaining profile types were also non-significant; *0 Fakes F*(6, 196) = 0.35, *p* = .910; *4 Fakes, F*(6, 196) = 0.18*, p* = .981; *Real, F*(6, 196) = 0.62*, p* = .715. It is evident from these results that there is no relationship between individual differences variables and participants' judgement accuracy.

Social Media

Participants were asked a series of questions in relation to their use of social media to assess whether their usage had an effect on their judgement accuracy.

Platforms

Participants were asked to select the social media platforms they use from a list of seven; *Facebook, Twitter, Instagram, Snapchat, TikTok, YouTube, and Other*, and rank these from most to least used. Of the participants who selected at least one platform (N = 190, 93.6%), *YouTube* was selected most often with 171 selections (84.2%), followed by *Instagram* with 163 selections (80.3%) and *Facebook* with 149 selections (73.4%). This means that 149 participants were active users of Facebook at the time of participation in the study, and so are familiar with the platform. Facebook was ranked as the most used platform by 33 (22.1%) of these 149 participants. *Other* was selected by 32 participants (15.8%), with only 9 (28.1%) of these ranking it as their most used platform. When asked to detail which platforms they use, participants reported: Twitch, Reddit, Discord, WhatsApp, Twitter, Pinterest, Tumblr, LinkedIn, Good Reads, and Telegram.

Purposes

Participants were given a list of 12 different purposed for using social media and were asked to select one or more that relate to their usage (Appendix A). The option of *Watching videos (TV/Films/YouTube etc.)* was selected by most participants (N = 185, 91.1%), followed by *Socialising with friends/keeping in touch* (N = 168, 82.8%). Six participants selected the option of *Other* and reported using social media for the following purposes: "learning", "find ideas or free pics/documents to download [sic]", "relaxing and reading interesting content", "memes, education", and "read information, memes".

Daily Usage

Of the 203 participants, 200 (98.5%) reported being a regular user of social media. When asked how much time they spend on social media per day, the majority of participants selected *4+ hours* (N = 57, 28.1%) and *2-3 hours* (N = 53, 26.1%). *Less than 1 hour* was selected by the fewest participants (*N = 13,* 6.4%).

To investigate if there is a relationship between the time spent on social media and participants' judgement accuracy, multiple regressions were conducted using accuracy scores for each profile type. The predictor *Less than one hour* was used as the constant. The multiple regressions for hours spent on social media are presented below in Table 5.

Table 5.
Multiple regression for social media time predictors of judgement accuracy for each type of profile.

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Hours spent on social media per day	.030			.012			.004			.011		
Constant		0.54	0.18		0.85	0.21		1.54	0.24		2.69	0.18
1-2 Hours		0.09	0.20		0.06	0.24		0.11	0.27		-0.20	0.20
2-3 Hours		-.010	0.20		0.23	0.24		0.03	0.27		-0.22	0.20
3-4 Hours		-0.23	0.21		0.26	0.26		0.19	0.29		-0.14	0.21
4+ Hours		-0.15	0.20		0.14	0.24		0.06	0.26		-0.09	0.20

Note. df = 4,198.

Table 5 shows that none of the coefficients or regression models are significant, meaning that no relationship has been found - time spent on social media per day is not a good predictor of participant's judgement accuracy for any profile type: *0 Fakes, F*(4, 198) = 1.55, *p* = .188, *2 Fakes, F*(4, 198) = 0.61, *p* = .656, *4 Fakes, F*(4, 198) = 0.21, *p* = .932, or *Real, F*(4, 198) = 0.53, *p* = .711.

Previous Experience in Creating a Fake Profile

Participants were asked to report whether they had any previous experience in creating a fake social media profile on any social media platform, and if so, provide reasons as to why they had done so. Of the 203 participants, 29 (14.3%) reported that they had previously created a fake profile. The reasons given include *investigative purposes;* "…to spy on other people", "to catch my girl cheating on me", *security/anonymity purposes;* "…speaking to people I wasn't supposed to be speaking to because my partner didn't want me to be speaking to them", "I wanted to get into a group to comment anonymously", "…I had just started using social media and was scared of using my personal details", and *malicious reasons;* "…to prank my friends", "…wanted to teach the class bully a lesson".

To investigate whether there was a relationship between previous experience creating a fake profile and judgement accuracy scores, multiple regressions were conducted for each profile type, using 'No' as the constant. The results of these regressions are reported in Table 6 below.

Table 6.
Multiple regression of previous experience creating a fake profile and accuracy scores for each type of profile.

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Experience in creating a fake social media profile	.009			.002			.001			.001		
Constant		0.48	0.05		0.98	0.06		1.61	0.07		2.53	0.05
Yes		-0.17	0.13		0.09	0.15		0.05	0.17		0.06	0.13

Note. df = 1, 201

As is evident from Table 6 there are no relationships between previous experience creating a fake profile and judgement accuracy for any type of profile due to the lack of significant coefficients. Additionally, each regression model was non-significant; *0 Fakes, F*(1, 201) = 1.82, *p* = .179; *2 Fakes, F*(1, 201) = 0.32, *p* = .574; *4 Fakes, F*(1, 201) = 0.07, *p* = .788; *Real*, *F*(1, 201) = 0.21, *p* = .651.

It has been evidenced that there are no relationships between social media predictors (hours spent on social media, or previous experience creating a fake profile) and judgement accuracy scores for any profile types.

Post-Hoc Analysis

Heatmaps

Heatmap layers were used over each profile to measure the areas participants used when making their judgements. Participants were asked to click on the profile on the areas they relied upon and were given a maximum of ten clicks to do so, a maximum set by the Qualtrics software used. These clicks were captured under heatmap regions, defined by the researcher, that covered each of the manipulated characteristics (Appendix K), and were visible to the researcher only. The regions were made as wide as was possible without any overlapping between regions representing different manipulated characteristics, and participants were instructed to be as accurate as possible. The frequency of these clicks across all profiles, for each of the manipulated characteristics per profile type are shown in figure 4.

Figure 4.

Graph showing the frequency of clicks per manipulated characteristics for all profile types

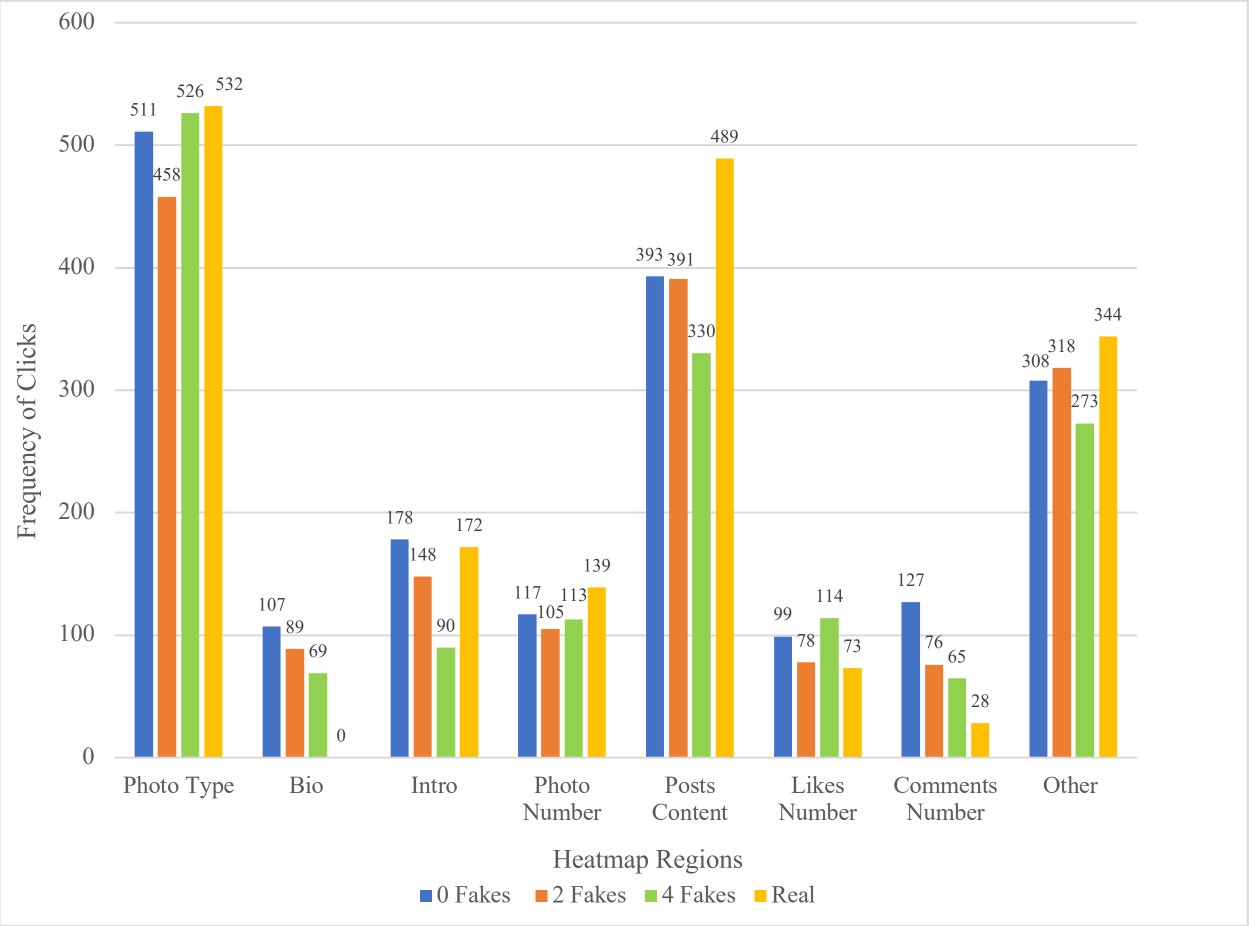
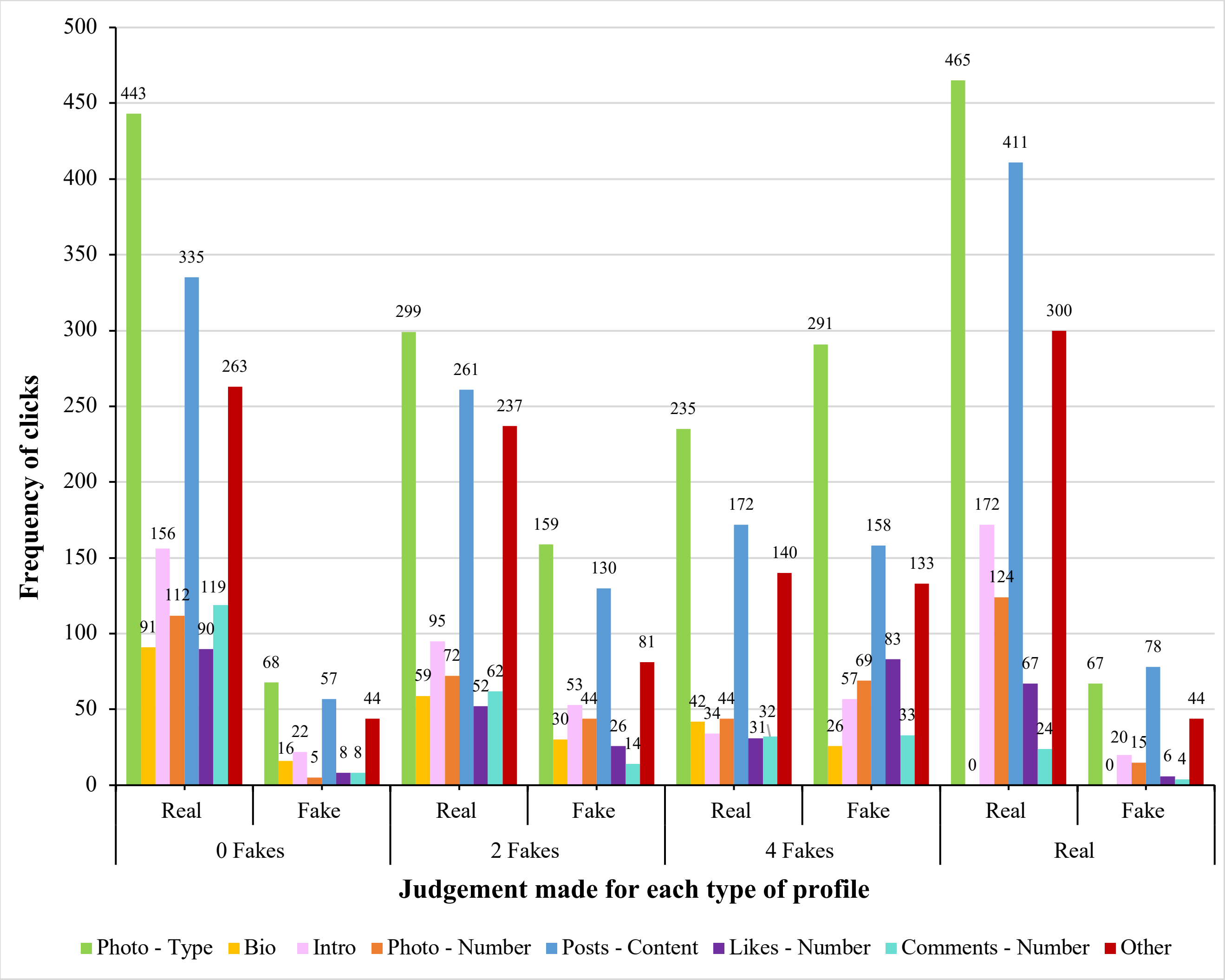


Figure 4 shows that across each profile type, *Photo Type* is the most clicked on area that participants used to inform their judgements, closely followed by *Posts Content.* Again, as in previous studies, the region of *Other* had a high frequency of clicks, which after manual checking of each, represented inaccurate clicks. An inaccurate click is any of those not encapsulated by the region surrounding the manipulated characteristics, for example a click not directly on the number of photos but rather just off to one side. None of the clicks under *Other* were on areas of the profile that had not been manipulated such as *Number of Friends*.

To further understand the areas of the profile used to inform judgements, further heatmap analysis was conducted on the frequency of clicks per type of judgement, i.e., did participants click on different areas when judging the profile as fake or real? Figure 5 displays the further analyses.

Figure 5.
Graph showing the frequency of clicks for each judgement type per manipulated characteristics for all profile types.

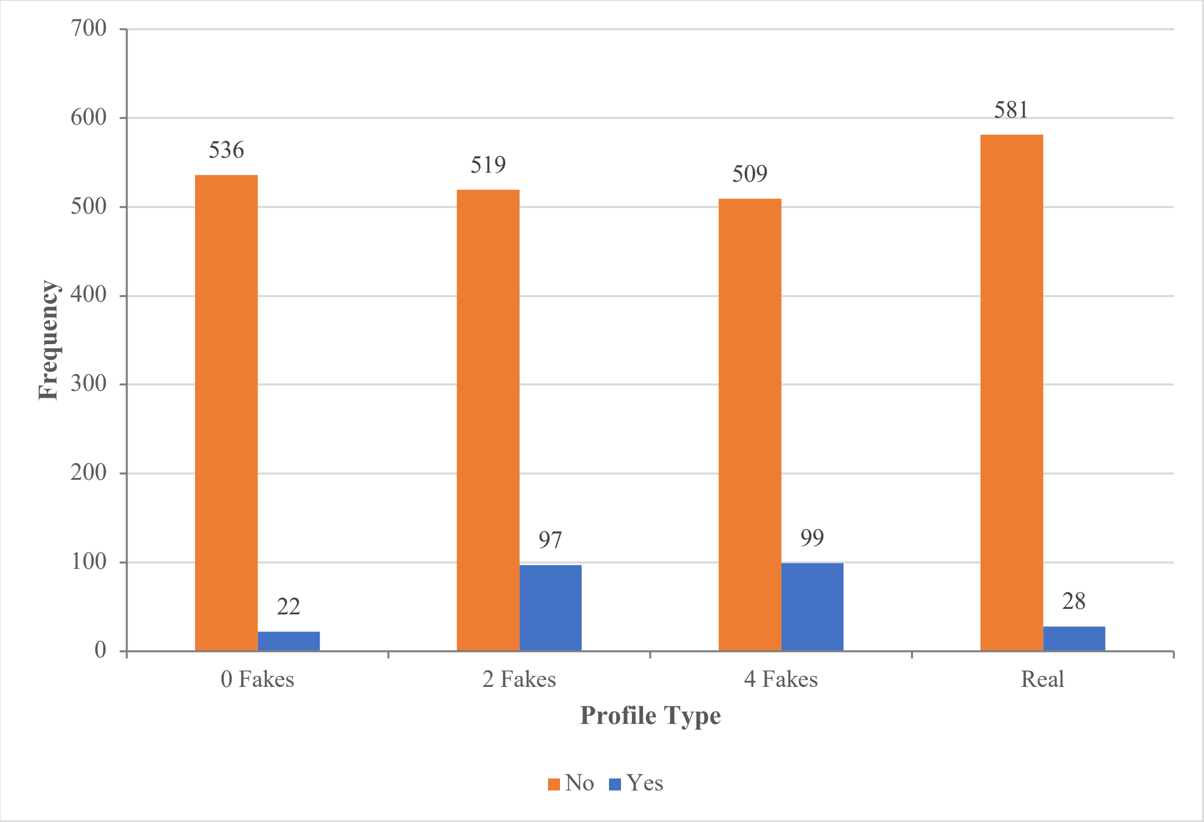


Overall, Figure 5 shows that participants clicked more frequently on the profile when they were judging the profile as real, regardless of the type of profile. This is particularly evident for *0 Fakes* profiles and *Real* profiles. Additionally, *Photo Type, Posts Content,* and *Other* are the most frequently clicked on areas, as outlined above in Figure 4, however it is evident here that this is the case whether the profile is judged as fake or real. This suggests that these characteristics are used more as a tool for judgement of the profiles, indicating that the authenticity of a profile is not wholly dependent on the profiles content but rather the mere presence of these areas on the profile, i.e. actually have a photo uploaded rather than a blank space where the photo should be.

Deceptive Purposes of Profiles

After judging each profile, participants were asked to answer yes or no as to whether they believe the profile was created with deceptive intentions/malicious intent. The data in relation to this question is displayed in figure 6 below.

Figure 6.
Graph showing the frequency of yes/no answers to the question 'Do you think this profile was created for deceptive purposes?' for all profile types.



As is evident in Figure 6, the majority of participants did not think the profiles, of any type, were made with deceptive intentions. The *4 Fakes* profiles were reported as being the most deceptive (N = 99) and *0 Fakes* the least (N = 22). These reports mirror the judgement accuracy findings reported in figure 1, whereby *4 Fakes* were judged more accurately as fake than *0 Fakes,* meaning the profiles that are seen as fake more often (*4 Fakes)* are judged more so as deceptive than those that are seen as real more often (*0 Fakes).*

Once participants had answered *Yes* they were asked to outline why they think the profile was deceptive. An example of the reasons outlined are: "Catfish and illegal activities", "… didn't have posts related to the profile shown", "Use of fake pictures", "…maybe to get photos or money", "maybe to scam people", "when the pictures are not very personal or profiles with few likes".

Discussion

Overall, findings from this study showed that there was no significant main effect of time taken to judge a profile on judgement accuracy of profiles. As found previously, participants showed a linear trend in accuracy scores, judging *Real* profiles most accurately and *0 Fakes* profiles least accurately.

Prior to conducting this study, a pilot study was carried out to test both the method and a time limit of 45 seconds. The results from the pilot indicated that the time taken to judge a profile did not have a significant effect on judgement accuracy, and participants reported that they did not have adequate time to make their judgements and overall felt rushed. This could have been due to the fact that, during the pilot, participants were required to view the profile, make their judgement, *and* click on the profile indicating the areas they used to form their judgement all within the 45 second time limit.

Based upon the results of the pilot, two amendments were made to the method and procedure of the main study. Firstly, in the main study participants were only required to view the profile under timed conditions. After the timer has reached the specified limit, participants were then asked to make their judgement. Following this, the same profile was shown to them again where they were instructed to click on the areas they used to inform their judgement. These changes were made to reduce any tasks that may distract from the judgement process and apply unnecessary pressures, and to ensure that the time limit was strictly related to forming a judgement of the profile from the stimuli available. The requirement to fulfil all of the steps in the pilot study within the time limit may have overwhelmed participants and had a detrimental effect on their judgement accuracy.

Speed-Accuracy Trade-off theory (SAT) posits that when performing a task either speed or accuracy has to be sacrificed in order to perform the task (American Psychological Association, n.d.), effectively meaning that time pressures/constraints can reduce response accuracy. Several researchers have found this to be true in regard to decision making processes - time pressures have reduced the quality of decision-making (Payne, Bettman, & Johnson, 1993), and induced judgements of a lesser extreme nature (Kaplan, Wanshula & Zanna, 1993). It has been suggested that the pressure of a time constraint prevents in-depth detailed processing of the information available, resulting in a 'closure of the mind' effect (Kruglanski & Freund, 1983). People rather rely on general rules of thumb otherwise known as heuristics (Tversky & Kahneman, 1974) when making decisions.

Heuristics are mental shortcuts used to make fast judgements, and arise from System 1 (Kahneman, 2011, pg. 98). The heuristic of most relevance here is the fluency heuristic which states that if one cue is processed at a faster speed, or more fluently, than another, then the mind infers this cue is of higher value (Schooler & Hertwig, 2005). Essentially, the fluency heuristic leads us to favour information that flows smoothly – the faster we can process the information the more likely we are to believe it. Researchers have found that when making snap judgements people are prone to relying on the frequency heuristic (Alter & Oppenheimer, 2009; Lewandowsky, Ecker, Seifert, Schwarz & Cook, 2012). In a social media context, information can be viewed at the pace of the user or in a rapid manner, as Dunaway, Searles, Sui and Paul (2018) found to be the case when studying the use of social media and the mediums on which the platforms are viewed. The rapid use of social media suggests that the fluency heuristic is replied upon when processing information. Consistent with this notion, Fazio, Brashier, Payne and Marsh's (2015) *Fluency-Conditional Model* suggests that people choose a fluency heuristic strategy to process information when judging the truth of said information rather than employing cognitive effort to search stored knowledge that could provide evidence to support the truthfulness of the information. As such, when considering the SAT theory, reliance on heuristics when under time pressure suggests that accuracy of judgements or decisions is affected due to the speed at which the task is undertaken.

The second methodological amendment based upon the pilot results was reduction of the time limit from 45 seconds to 40 seconds. Despite participants reporting being rushed and not having adequate time to complete their judgements, with the changes in the procedure outlined above, the researcher deemed the five second reduction to be fair and achievable when only one task was required to be completed in that timescale. Further, the reduction in time limit was made in an effort to measure which processing system, System 1 or System 2, is used when judging the profiles.

Results from the main study replicate those in the pilot study as no significant effect of time on judgement accuracy was found. There were some significant relationships between time taken to form a judgement and accuracy of said judgements, however these were only found in relation to profiles with *0 Fakes* and *4 Fakes.* The direction of the relationship differed between the two; *0 Fakes* was positively related with time and *4 Fakes* negatively related. Meaning that the longer the time taken to judge a *0 Fakes* profile accuracy is likely to increase, whereas the longer the time taken to judge a profile with *4 Fakes*, profile accuracy is likely to decrease.

This could potentially be explained by the fact that *4 Fakes* profiles have the most extreme manipulations, i.e., may contain only one photo, two friends, one post, and no intro information. As well as having the most manipulations however they are also the profiles that display the least amount of information. From this it could be suggested that these profiles invoke more of a 'snap-decision' process due to the minimal amount of content or cues being available, thus less time is needed to reach a decision. Having more time for these profiles brings no additional benefits due to limited content available to judge. By contrast, *0 Fakes* profiles have less extreme manipulations (and more profile information). Even with a suspicion of deception, participants find it harder to pinpoint anything immediately wrong. In these instances, more time might allow for a more critical analysis of any slight variations which weren't apparent with a cursory inspection. Although at this stage this is only loosely evidenced and more research is required to further investigate this possible inference.

Participants' average time to make their judgements was well within the 40 second time limit (25.86s), and they reported having adequate time and did not feel rushed at all, suggesting the methodological changes made for this study were appropriate. To understand the decision-making system used when making their judgements, further exploratory analysis was conducted in regard to the six participants who had an average time of less than ten seconds to make their judgements. As outlined in the introduction, researchers have suggested that when using social media people use

System 1 (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016; Haile, 2014), which is said to involve processes under ten seconds (Kitajima & Toyota, 2011). Findings show that these six participants did not achieve high judgement accuracy scores and time spent judging the profiles did not have a significant effect on their judgement accuracy. This suggests that System 1 processes within the 10 second time limit, as suggested in the literature, do not produce high levels of accuracy in regard to authenticity judgements. A concerning finding when considering that System 1 is most likely to influence human judgements (Dennis & Minas, 2018) due to the laziness of System 2 (Kahneman, 2011, pg. 64).

Of interest, the six participants clicked most on *Posts Content* as the area of the profile used to inform their judgements, rather than *Photo Type* as in the main study. *Photo Type* was the second most clicked on area, followed by *Other*. Literature on thin-slicing and stimuli processing speeds reports that accurate judgements can be made in very short spaces of time, whether that be judgements of faces in under 100ms (Willis & Todorov, 2006) or judgements of IQ in between 1s and 10s (Murphy et al., 2003). Therefore, this literature suggests that these six participants could make accurate judgements of fake profiles in under 10s. However, it should be taken into consideration here that there is the possibility that these participants could be anomalies within the sample in that they rushed through to completion and did not complete the study in the expected way.

As in previous studies, the linear trend in judgement accuracy for each of the different profile types was found once again, with *0 Fakes* receiving the lowest accuracy score and *Real* profiles receiving the highest accuracy score. Additionally, the manipulated characteristics of *Photo Type* was once again a significant predictor of participants' judgement and judgement accuracy – if *Photo Type* has been manipulated on the profile, then the likelihood participants would judge the profile as fake increases, and the likelihood that this judgement would be accurate increases. Such findings highlight the robust nature of this running thread through this research and show that, even with the added manipulation of time pressure, *Photo Type* is a reliable cue to rely upon that indicates a fake profile.

The individual differences and social media variables were not hypothesised to have an effect on participant's judgement accuracy so were not included within the main analysis. However, they were still controlled for and results of such were reported. The results showed that there were no significant effects of individual differences (personality and social sensitivity) or social media (time spent on social media and previous experience creating a fake profile) on participant's judgement accuracy.

The lack of results in regard to the time limit and judgement accuracy could be due to the fact the participants had autonomy over the page moving on from the profile and on to the next page. Participants were able to select the arrow to move on at any point within the 40 seconds time limit. This feature could perhaps reduce the amount of pressure experienced by the participants, meaning some could have spent longer on the profile to make sure they gather as much information possible to make a "correct" judgement. To invoke more of a sense of pressure, future researchers could enforce the page to move forwards after the maximum allotted time had been reached.

Additionally, future research should manipulate the duration of the time limit further. Despite the fact no significant effect of time on accuracy was found, a few significant correlations were found, suggesting that time taken to judge the profile had some influence over the accuracy of the judgements. By reducing the time limit, not only may an effect of time on accuracy be found, but it may also aid in providing a better understanding of the decision-making processes, i.e., which systems (System 1 or System 2) are invoked when judging the authenticity of social media profiles. Overall, further research is needed into the specific timings of System 1 and System 2 processes to truly understand when System 2 is activated when forming judgements or making decisions.

Chapter 6: Study 5

Introduction

To gain a deeper understanding of the accuracy of judgements regarding fake social media profiles, it is crucial to consider the impact of culture, as social media is a global phenomenon, that transcends boundaries and cultural norms. Therefore, this study follows similar methodology as in previous studies within this thesis while introducing the variable of culture.

Culture is notoriously difficult to define (Spencer-Oatey, 2012), so much so that a global consensus on the definition of culture has yet to have been achieved (Apte, 1994). The notion of 'culture' is said to be 'one of the two or three most complex words in the English language' (Eagleton, 2000, p.7). It is unknown how many different definitions there are currently, however in the 1950's anthropologists Kroeber and Kluckhon (1952) reported there were 164 different definitions of culture, so it can only be expected that this number has grown significantly in the 72 years since. Additionally, dictionaries offer more than one definition; Merriam-Webster dictionary offers six different definitions and Oxford English dictionary offers 12 different definitions.

The widely credited, and first known definition of culture, by Tylor (1871) states that "Culture or Civilization, taken in its wide ethnographic sense, is that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society" (p.1). However, the definition used within this research, outlined by Kashima and Gelfand (2011), differentiates culture from society; "a set of meanings or information that is non-genetically transmitted from one individual to another, which is more or less shared within a population (or group) and endures for some generations." (p. 499).

Culture and cultural differences have been studied by many across a broad range of spectrums from anthropology and management to sociology and psychology (Fatehi, Priestley, & Taasoobshirazi, 2020). However, despite the wealth of research, little progress has been made in actually understanding cultural differences and the underlying primal template of culture (Gelfand, 2018, pg. 2). Culture is often referred to as being either *individualist* or *collectivist* (Hofstede, 1980; Triandis, 1995). *Individualism* describes those where the ties between individuals are loose (Hofstede, 1991, p.28), where people tend to be more autonomous and prioritise their personal goals over collective goals (Triandis, 1989). Such countries considered as having an individualist culture are United Kingdom, Australia, and the United States of America. *Collectivism* describes cultures where members are integrated into cohesive, strong groups (Hofstede, 1991, p.28) and are strongly guided by group norms (Triandis, 1995). Such countries include Mexico, Hong Kong, and Singapore. Members of collectivist cultures tend to disclose less information, with any information shared being low-stakes information, and individualist cultures tend to disclose much more information which is usually of a more personal nature (Hofstede, 1980).

A more recent development in the cultural research domain is the concept of *tight* and *loose* cultures (Gelfand, Nishii, & Raver, 2006). Although first theorised by anthropologist Pelto (1968), and further developed by Berry (1966) and Triandis (1989), progress on the theory had been at a standstill with no attempts to apply the theory to modern-day societies until the reintroduction of the theory in the early 2000's by Gelfand, Nishii, & Raver (2006). The *Tightness-Looseness* (TL) theory differs from the *Individualism-Collectivism* (IC) theory in that TL captures cultural variance in relation to the influence of social-norms and sanctions, and IC relates to how a member's family or ingroup can influence behaviour (Gelfand, Nishii, & Raver, 2006).

Tight cultures are ones that can be considered as a *rule makers* in that they have strong social norms and very little tolerance for any deviations from such norms, and *loose* cultures, or *rule breakers,* are those who are permissive and have weak social norms (Gelfand, 2018, p.3). Cultures tend to be a combination of both TL and IC theories, for example, Germany is considered as *Individualistic – Tight,* New Zealand & UK as *Individualistic – Loose,* Singapore, China, and India as *Collectivist – Tight,* and Brazil, and Hong Kong as *Collectivist – Loose.*

Cross-cultural study within Psychology, and many other fields, is an ever-growing avenue of investigation. Over the past 15 years, researchers have been applying cross-cultural methodology and theory to how different cultures use social media. As social media continues to reach every corner of the world with a multitude of global platforms and billions of users, it only makes sense to investigate the cultural differences in how and why people use social media. However, in the years following the early conception of social media, many researchers highlighted that little is known about social media use in collectivist cultures (Nadkarni & Hofmann, 2012; Wang, Norcie, & Cranor, 2011), and this appears to still be the case. Of those that have researched both individualist and collectivist cultures and their relationships with social media have used the medium of Facebook, perhaps due to the fact it is a global platform and consistently sits atop the chart as the most used platform in the world, with 3.049 billion active users as of January 2024 (Statista, 2024c).

Na, Kosinski, & Stillwell (2015) found that social networks on Facebook in individualist cultures were more egocentric than collectivist cultures, and Huang and Park (2013) found that Facebook profiles in collectivist cultures include more contextual and relational information than profiles in individualistic cultures. Additionally, Zhao and Jiang (2011) found profile images differ between cultures;

members of different collectivist cultures post more neutral images of themselves, or no photos at all, whereas members of individualist cultures post photos of a much less discrete nature. These studies demonstrate that a persons' culture determines both the features of persons' profile online and the patterns of behaviour exhibited online.

In relation to this body of research on fake Facebook profiles, human judgement, and the characteristics used to inform authenticity judgements, the researcher felt it important to understand whether judgement accuracy of fake profiles differs across and/or between collectivist and individualistic cultures. The individualist cultures used are a combination of cultures in the west, sometimes referred to as *Western*, and consist of mainly European and American countries. The culture of comparison focused on within this research is the Indian culture. India is the seventh largest country in the world, by area, with a population of 1.44 billion (Statista, 2023), and culturally is *collectivist-tight* culture; third *tightest* nation in a study based on social norms conducted by Gelfand et al. (2011). Characteristics of *tight* culture in India, outlined by Gelfand (2018), include low alcohol consumption rates (p.45), low crime rates (p.27), and less creative problem-solving due to sticking to the norms (p.130). When considering the suitability of selecting Indian culture for the study, the relationship between Facebook and India was investigated. On a global usage scale, India is the country with the most users. As of January 2024, India has just under 367 million users – a figure almost double that of the United States of America which is currently the country with the second highest number of users (190 million) (Statista, 2024d). It has also been reported that India has a significant issue with fake profiles, and a few years ago was one of the top countries with fake Facebook accounts (Bilge, Strufe, Balzarotti & Kirda, 2009). Considering the above information collectively, focusing on Indian culture seemed the most appropriate when considering the cultural differences of Facebook use and fake profiles.

Despite the fact that India uses Facebook more than any other country or culture, little is known about how Indian users actually use Facebook, including the ways in which they present themselves and what is considered 'normal' Facebook use. Of the little research there is, the main focus has been on Facebook addiction among Indian students (Masthi, Cadabam & Sonaksi, 2015), the impact of Facebook on school grades (Maheshwari & Mukherjee, 2020; Stollak, Vandeburg, Burkland, & Weiss, 2011), and basic demographic details such as time spent on Facebook per day (Manjunatha, 2013). One distinct feature of Facebook in India is the introduction of the 'lock profile' feature in 2020, whereby a user now has the ability to block their profile content from anyone on the platform whom they are not friends with. Facebook introduced this feature specifically to protect female users in India – women were reporting numerous cases of identity theft whereby the photos from their profiles were used on fake profiles or for emotional blackmail (Andrew Hutchinson, 2020). This feature is not available worldwide. Whilst Facebook have not released an official list of the countries this feature is available to, it has been reported that as of 2024 the following countries can lock their profile: United Arab Emirates, Afghanistan, Iraq, Myanmar, Ukraine, Pakistan, Saudi Arabia, Turkey, Morocco, Egypt, Sudan, and India (Naveh Ben Dror, 2024). Of interest here is that the majority of these countries can be considered as *collectivist* or *tight* cultures, a notion that is not surprising when considering *tight* countries tend to place restrictions on what can be said in public and will often censor the media and monitor use of social media platforms (Gelfand, p.53).

After a lengthy literature search the researcher did not find any studies that have analysed fake profiles in India - this research appears to be the first to do so. As such, literature from judgement and decision-making in different cultures, an area that has been widely researched, will be drawn upon to understand how culture may have an effect on judgements of fake profiles.

Decisions, defined as making one's mind up, result from judgements, which are ideas and opinions of an object, person, situation, phenomenon etc. (Bonner, 1999). Decision-making has been shown to differ across cultures. For example, the Danish have a functional approach to decision-making by evaluating all available solutions before landing on a decision (Schramm-Nielsen, 2001). German decision makers rely on their status within a hierarchy to guide their decisions, whereas in China and India, a clear, unambiguous top-down decision-making is favoured, whereby natives often adhere to their superiors' formal authority (Khairullah & Khairullah, 2013).

Judgements, of many different kinds, have also been shown to differ across cultures. Probability judgements have been found to differ between Asian and British students (Wright & Phillips, 1980): Asian students were more overconfident in the accuracy of their judgements, giving more extreme and unrealistic judgements than British students. Judgements of risk and risk-taking behaviour tasks have found a cultural difference; Asian participants are less risk-averse than Western (Americans, New Zealanders, Dutch) participants when judging gains and losses in hypothetical situations (Marshall, Huan, Xu, & Nam, 2011). In specific relation to this study, it has been found that first impression judgements of others' personality on social media are not overly accurate, even when the judgement is made of someone from their own culture (Turner & Chin, 2017). Additionally, it has been reported that cultures differ in their judgement and categorisation of truthful and deceptive statements (Fu, Lee, Cameron, & Xu, 2001).

Due to the limited research on authenticity judgements across cultures, specifically authenticity of social media profiles, it is difficult to hypothesise a specific relationship or interaction based on previous literature in this field. However, when considering the literature on perception of familiar stimuli, findings show that there are cultural differences in how different stimuli are perceived. For example, East Asians have been found to allocate more attention to facial information of stimuli than Americans (Miyamoto, Yoshikawa, & Kitayama, 2011), and Americans focus more on focal objects rather than East Asians who focus more on contextual stimuli (Millar, Serbun, Vadalia & Gutchess, 2013). When perceiving facial expressions in images, Ekman (1972) and Izard (1971) found that participants could correctly identify the emotion displayed in the facial expression, and so could do so even when the participant was from a different culture to that of the person in the image stimuli. However, Elfenbein and Ambady (2002) report that accuracy for categorising facial expressions was reliably higher when both the perceiver and the stimuli are from the same culture. These examples show that there are differences in how cultures perceive information regarding faces and emotional expressions, and this perhaps may also be the case in regard to online stimuli. As such, this study hypothesises that there will be a difference in accuracy scores between cultures, specifically when a culture is judging profiles from their own culture versus profiles from a different culture (Hypothesis 1).

Based upon previous findings in this current research it is also hypothesised that there will be a difference between the accuracy of profile judgements for real profiles and fake profiles (Hypothesis 2). Additionally, profiles with the highest number of fake characteristics will be more accurately judged as fake, than profiles with fewer, or no, fake characteristics (Hypothesis 3). Again, due to the lack of literature regarding judgements of fake profiles, specifically judgements across cultures, it cannot be hypothesised whether an effect will be found between the cultures within this study. However, this will be measured when reporting the results.

It is also expected that there will be a relationship between participants' self-reported confidence in accuracy of their judgements and actual judgement accuracy (Hypothesis 4). Even though the study mentioned above conducted by Wright and Phillips (1980) found higher levels of overconfidence in judgements by Asian students when compared to American students, this study was conducted using probability judgements rather than authenticity judgements and different cultures to those used within this study. As such, the hypothesis remains non-directional, but an effect is expected nonetheless.

Again, based upon the consistent findings from previous studies within this research thesis it is expected that the manipulated characteristics will be significant predictors of whether participants judge the profiles to be real or fake (Hypothesis 5), and significant predictors of participants' judgement accuracy (Hypothesis 6). Specifically, it is expected that 'Photo-Type' will be the strongest predictor of both participant's judgements and accuracy of said judgements (Hypothesis 7). The lack of research into characteristics of fake profiles overall, and specifically in relation to different cultures, again means that the researcher cannot hypothesise any specific cross-cultural effects. However, any effects of culture on judgements and judgement accuracy related to the profile characteristics will be measured and the findings of such reported.

In contrast to hypotheses in Study 3 and 4 (Chapters 4 and 5), it is expected that the social media variable *time spent on social media per day* will have an effect on the judgement accuracy of Indian participants (Hypothesis 8). As mentioned previously, there have been several studies on the issues of Facebook addiction in India (Masthi, Cadabam & Sonaksi, 2015). Additionally, it has been reported in a technology policy report that users in India spend on average 194 minutes per day on Facebook (Sharma & Gautam, 2023). As such, it is expected that an effect will be found amongst the Indian participants.

Finally, the social media variable of *previous experience creating a fake profile*, and the individual difference variables (personality types and social sensitivity) are not included in the hypotheses due to the lack of findings in the previous studies within this thesis and the lack of research in general on these variables in relation to judgements of fake profiles and the effects of culture. As such, these variables will not be directly measured but rather controlled for and reported alongside the main findings of this study.

Method

Participants

The participants in the first phase of this study were required to provide their real Facebook profile for use within the study. As with the previous studies, the same six real Facebook profiles obtained from the researchers' friends and family were used for this research. These participants ranged in age from 29-62 years (*M* = 43.33 years, SD = 14.47), three identify as Female and three as Male, and all identify as White British.

In addition to these six real profiles, a further six real profiles were obtained from an Indian student population to allow for cross-cultural comparisons. These participants were recruited via an advert on Facebook and Instagram via Indian society accounts and UK University Freshers accounts (see Appendix Y). A recruitment snowball effect was present amongst peers as some participants shared the advert with their peers. As stated in the advert, participants were required to be Indian and have just started either their first or second year of their University course, have a Facebook account, and be users of Facebook. These stipulations were necessary to ensure that the participants not only used Facebook, but that their Facebook accounts were as authentic to their culture as possible and not influenced by British culture, hence why they were required to have just started their course and be new to the UK. The second-year stipulation was included to encompass the year of 2020 during the COVID-19 pandemic whereby participants may have started their course but still have been residing in India. To capture the demographics of these students and assess whether these stipulations were met, a short demographic questionnaire was created on Qualtrics (Appendix N) and circulated to those who responded to the advert.

Once initial consent was obtained, participants were given access to an instructional video showing them how to screenshot their Facebook profile. The researcher created two videos that were tailored for either Mac users or Windows users - participants were sent the link to both videos and asked to open the appropriate one based on their computer software. The videos instructed participants to capture the six most recent posts on their timeline, as well as the other standard aspects of their profile, including their: cover photo and profile photo, bio, introduction, photos etc. Participants were then asked to email the researcher their screenshots. The researcher then created an overall screenshot of their profile by using Photoshop software to piece all the individual screenshots together. The only edits made to these screenshots was the removal of any identifying information such as names, profile pictures of friends etc. that were outlined in the initial consent form (Appendix G). Once completed, the screenshot was emailed back to the respective participant along with a second consent form (Appendix H). Participants were asked to look at the screenshot carefully, and if happy with the screenshot asked to read through the consent form and sign it confirming that they are happy for the profile to be used in the next phase of the study.

All six participants that provided their Facebook profile all identified as ‘Asian or Asian British (Includes any Asian background, e.g., Bangladeshi, Chinese, Indian, Pakistani and other Asian Background)’, all were born in India, all have lived in the UK for less than one year as all are in their first year of study. The ages ranged from 18-35 years (*M* = 26.83 years; SD = 5.19), two participants identified as Female (33.3%) and four as Male (66.6%). These six participants were paid a £20 Amazon voucher for their time and participation.

Participants were recruited online through Prolific through means of volunteer sampling. Two identical studies were ran separately to allow for capture of the two different cultures within this cross-cultural study: Indian Prolific users and non-Indian Prolific users. The first study was run with India excluded from the participant pool, and the second study had two stipulations to capture Indian participants; ‘Country of birth’, and ‘Place where most time has been spent before turning 18’. Both of these categories were set to India.

An A-Priori power analysis of a repeated measures within subjects ANOVA, was conducted using G* Power (Faul et al., 2007) prior to data collection to determine the appropriate sample size for this study. The analysis indicated that a sample size of 11 participants per data set (Indian participants and Non-Indian participants) would be sufficient to detect a medium effect size of f = 0.25, with an alpha level of α = 0.05 and a power of 1−β = 0.80. However, 200 participants were recruited (100 per data set). As outlined in Study 1 (Chapter 2), Study 3 (Chapter 4), and Study 4 (Chapter 5) the reasoning behind this decision was to enhance the reliability and generalisability of the findings by reducing the errors associated with the estimates, improve the representativeness of the sample, and reduce the risk of Type 1 and Type 2 errors. Additionally, based on the results of studies 1, 3, and 4 where a similar design was employed, a similar sample size of 200 participants was chosen to maintain consistency across the studies within this research and ensure results between each study were comparable. A total of 210 participants were recruited on Prolific.co across both studies, however 10 were removed due to incomplete entries, leaving 200 in total, 100 per dataset.

In the Non-Indian dataset, 38 identified as Male, 57 as Female, and 5 as Non-Binary. The ages ranged from 19 – 54 years with a mean age of 27.67 years. When asked to state their ethnicity, 26 participants identified as 'Black, African, Black British or Caribbean (Includes any Black background)', 11 participants identified as 'Mixed or Multiple Ethnic Groups (Includes any mixed ethnic background)', 55 as 'White (Includes British, English, Scottish, Welsh, Northern Irish, Irish, Irish Traveller or Gypsy and any other white backgrounds)', and 8 as 'Another Ethnic Group'. There were no participants in the Non-Indian dataset that identified as 'Asian or Asian British (Includes any Asian background, e.g., Bangladeshi, Chinese, Indian, Pakistani and other Asian Background)'.

In the Indian dataset, 54 participants identified as Male, and 46 identified as Female. The ages ranged from 21-67 years with a mean age of 32.14 years. When asked to state their ethnicity, 94 participants identified as 'Asian or Asian British (Includes any Asian background, e.g. Bangladeshi, Chinese, Indian, Pakistani and other Asian Background)', one as 'Black, African, Black British or Caribbean (Includes any Black background)', two as 'White (Includes British, English, Scottish, Welsh, Northern Irish, Irish, Irish Traveller or Gypsy and any other white backgrounds)', and three participants selected 'Prefer not to say'.

At the end of the demographic's questionnaire, participants were asked to enter cultural details, including the country they were born in, the country they currently live in, and how long, in years, they have lived in their current country. In the Non-Indian dataset, the most popular countries participants were born in and currently live in were South Africa, Poland, Portugal, and Mexico. Whereas, in the Indian dataset, the vast majority of participants were born in India (95%) as expected with the study parameters, however none of these participants currently still live in India. The most popular countries of current residence are United Kingdom, Germany, Canada, and United States of America.

Design

A 4 (Real profiles, 0 Fakes profiles, 2 Fakes profiles, '4 fake' profiles) x 2 (Accurate judgement vs. Inaccurate judgement) x 2 (Indian vs. Non-Indian) experimental Turing test design will be used to investigate the study hypotheses. The Dependent Variable (DV) is the accuracy of the judgements and the cultures, and the Independent Variable (IV) is the Facebook profiles. Both variables are within subjects' measures following a repeated measures design.

Measures, Materials, Equipment

Measures

As this study employed a very similar method to the previous studies within this research thesis, the same four self-report questionnaires were administered to participants online using Qualtrics. These being the social media questionnaire created specifically for this research, the Ten Item Personality Inventory (TIPI) (Gosling et al., 2003), the Social Sensitivity (SS) scale (Riggio, 1986), and a follow up questionnaire containing two short questions to measure participants' self-reported accuracy and previous experience in creating a fake profile. (See Appendices A-D for details on each).

Materials

A total of 148 profile screenshots were used within this research: 68 Non-Indian fake profiles, 68 Indian fake profiles, six Non-Indian real profiles, and six real profiles from Indian university students.

The fake profiles created by the researcher contained different combinations of seven manipulated profile characteristics: *Photo Type, Photo Number, Bio, Intro, Posts Content, Comments Number,* and *Likes Number*, regardless of whether they were Non-Indian or Indian cultural profiles. These are the same manipulations as used in Studies 2, 3, and 4, to allow for investigation as to whether any results are reproduced with the addition of a cultural variable. The real profiles were not manipulated in terms of profile characteristics, the only 'manipulations' that occurred were the omittance of identifying information for the purposes of ethical guidelines, such as the names of the profile owners, friends' photos, or names of friends that were commenting/interacting with the profile.

Both the 68 fake Non-Indian profiles and the 68 fake Indian profiles consisted of a mixture of; 12 profiles with '0 fake characteristics', 21 profiles with '2 fake characteristics', and 35 profiles with '4 fake characteristics'. As each type of profile has a different number of possible combinations of the seven fake characteristics, each type of profile therefore has a different total number, i.e., there were 21 possible combinations of 2 fake characteristics, and 35 combinations of 4 fake characteristics, hence the total number of profiles reflects this calculation (see Appendix P for Non-Indian characteristic framework, and Appendix T for Indian characteristic framework). The profiles with 0 Fakes technically had no manipulations of the fake characteristics, as these profiles were created specifically to ‘fool’ participants into judging the profile as real, i.e., the profiles were created to look as real and authentic as possible. To allow for a level of variability and random assignment to participants, 12 profiles with 0 Fakes were created.

These characteristics were garnered from available literature and statistical reports regarding Facebook use. This remained true for the Non-Indian profiles however in terms of the Indian profiles, there was a distinct lack of research available into the typical profile characteristics found on Indian profiles. Coupled with the ‘locked’ profiles of a lot of Indian users, it proved difficult to find the norms within the profiles to use as a baseline. So, in order to know what to manipulate on the profiles to denote them as fake, the researcher created a profile characteristic database using convenience sampling. This process involved using the names created for the fake profiles (See Appendix T for the table) and searching these names on Facebook. The researcher took the first profile in the search results that had their profile on public and did not use the ‘lock feature’. From these profiles, the research recorded the presence of same characteristics as the manipulated ones in the Non-Indian fake profiles (*Photo - Type, Photo Number, Bio* (Yes or No), *Intro* (List of things included), *Posts Content*, *Comments Number*, and *Likes Number*). In relation to the posts, comments, and likes, the researcher took this information from the first five posts on their timeline. The same information was gathered from the six real Indian profiles provided by the participants from Lancaster University to add a further level of authenticity to the characteristics. Descriptive statistical data of these profiles was analysed, including the range, median, mode, and mean for each characteristic. To gain the range for each characteristic of what constitutes a ‘real profile’ the upper and lower bounds were calculated. For example, the average number of photos was 182, with an upper bound of 267, and a lower bound of 98, hence a range from 98-267 was used as a ‘real’ manipulation and anything outside of this was used as a ‘fake’ manipulation. The specific ranges for each characteristic are detailed in Appendix T.

Further to this, the researcher investigated the most popular things in India to include within the content of the posts to ensure the profiles created were representative of an Indian person living in India and not assumptive or culturally biased. For example, the researcher gathered information regarding the top 10 most popular celebrities (M & F), games, companies/brands, artists/musicians etc., and used these as the topic or content of the posts on the profile.

To ensure the fake profiles used in the previous studies were all in fact non-Indian profiles, the researcher manually checked through each and identified three profiles that could be deemed as Indian and would therefore need changing to ensure the Non-Indian profiles were Non-Indian only. Of the three identified profiles, one was deemed to be suitable as a non-Indian profile as it had no obvious identifying factors that would explicitly show the profile was from an Indian culture, i.e., there was no location detailed in the information section. The only feature that raised a question in the researcher’s mind was that of the photos being photos of an individual in distinctive cultural dress, however, to assume that this means the profile cannot be that of a ‘non-Indian’ person is both reductionist and culturally biased. Thus, the profile remained unchanged.

The two profiles that were edited both explicitly detailed a location that is not that of a ‘Non-Western’ culture. The profile ‘RJ’ showed: the location as being in Zimbabwe, the *Bio* as being born and living in Zimbabwe, and the second post on their timeline as a graduation ceremony in Zimbabwe (See Appendix U for the original). As a result, these three aspects were edited to be a Western location and a more suitable post based on the location of the profile user (See Appendix V for edited version). The profile 'MI' showed the location as being in India (See Appendix W for the original), so this was edited to read a non-Indian location (See Appendix X for edited version).

The non-Indian fake profiles were edited further by changing a few of the locations to match the locations of the universities - three of the real Indian profiles obtained from the students at university stated that they were living in the same city and studying at the same university. The edits were made to the Non-Indian fake profiles to avoid these profiles standing out and looking 'obviously real' due to the location of these participants being in the same location. Ten of the fake profiles were edited to have either the 'location or university the same as those in the real profiles. The same location and university information was also included in the Indian fake profiles created for this research for consistency. This method was chosen in favour of removing the location/university data from the real profiles, as doing so would mean the researcher would be directly manipulating the real profile thus resulting in it losing it authenticity and could therefore be categorised as 'fake'. Doing so not only controlled for any potential biases that may skew the data, but it also meant that the real profiles remained unchanged and were therefore accurate representations of a real profile.

The final edit to the profiles was to the layout of the Facebook profiles. All profiles used within Study 3 and 4 were changed to the new Facebook layout to match the new Indian fake and real profiles to ensure consistency between the two different cultures.

Procedure

This study followed a similar procedure to Study 3, with minor additions due to the cross-cultural variables present within this study.

All participants for the profile judgement phase of the study from both cultures were recruited online via Prolific All participants were redirected from their respective participation website through to the same study on Qualtrics.

Firstly, all participants were provided with information regarding the study and were asked to provide their informed consent to participate. Once consent was given, participants were required to complete three short self-report measures: the social media questionnaire, TIPI questionnaire, and SS Scale. Following this, participants were given a set of instructions regarding the profile viewing phase of the study, informing them that they would be viewing a total of 16 Facebook profile screenshots in a random order and would be required to make a judgement as to their authenticity. The participants were not told how many real or fake profile they would be viewing, but rather that 'some are fake, and some are real', to ensure that their decision-making process and judgements were as authentic as possible and not influenced by any erroneous details. Further, participants were asked to click on the areas of the profile that they used when making their judgement as to the authenticity of the profile (See Appendix R for instructions).

To answer the research question of whether people are more accurate at judging Facebook profiles from their own culture or from a culture different to their own, all participants were shown 12 random fake profiles (four *0 Fakes*, four *2 Fakes*, and four *4 Fakes*), and four random real profiles, totalling 16 profiles. Each of the two cultures saw a mixture of profiles: eight from their own culture and eight from the different culture. For example, the participants were shown; 8 Non-Indian profiles (two *0 Fakes*, two *2 Fakes*, two *4 Fakes*, and two *real* profiles), and 8 Indian profiles (two *0 Fakes*, two *2 Fakes*, two *4 Fakes*, and to *real* profiles). The total number of profiles shown was changed from 12, as in the previous studies, to 16 to ensure an equal number of profiles from each culture were shown to participants, thus allowing for clearer reproducibility of data.

Following completion of the profile viewing phase, participants were asked to complete a short follow-up questionnaire and asked to enter some demographic details. Participants were then debriefed about their participation and provided with further information regarding the study itself, the research overall, and contact details of the researcher and supervisors should they wish to ask any questions or withdraw from the research (See Appendix Z for ‘updated debrief form – 16 profiles’).

Ethics

This research was fully approved by the ethics committee at Lancaster university, under an amendment to the same ethics submission as Studies 1, 2, and 3. All participant data was stored on a secure hard drive, in line with GDPR guidelines, only accessible to the researchers, and destroyed after the appropriate amount of time.

Results

Data was split across two separate files due to the implementation of the inclusion criteria in Qualtrics to capture participants from India, thus one dataset contains the data from participants from India, and the second is data from participants not from India, otherwise referred to herein as the “Non-Indian” participants. The raw datasets were both exported from Qualtrics and imported to separate Excel workbooks whereby the descriptive statistics were analysed, and the remainder of the data was coded and sorted into the appropriate formats for input into IBM SPSS and R for further analysis. A total of 10 entries were removed from the analyses; nine from the Western dataset due to incomplete or timed-out entries, and one from the Indian dataset as consent to participate was not given. As such, each dataset had a total of 100 participant responses used in the analyses.

Both datasets were treated the same, using the same methods of analysis to produce the necessary output. The total accuracy score was calculated by summing the total scores achieved for each type of profile in each culture. Each participant viewed two profiles from each type of profile, in each culture; two Western '0 fake' characteristics, two Indian '0 fake' characteristics, two Western '2 fake' characteristics, two Indian '2 fake' characteristics, and so on. As such, each participant was given an accuracy score, with a maximum total of 16, for their judgements of the profiles based on how many profiles they correctly judged.

To assess the appropriateness of the data, several normality tests were conducted prior to inferential statistical analyses. Initially the accuracy score data had to be jittered within Excel to allow for plotting against other variables measured within the research, as the data points were stacked upon one another. Upon visual inspection of the histograms produced, the data in both datasets showed normal distributions, and all Q-Q plots and scatterplots showed the accuracy data was of a linear pattern against the other measured variables. Additionally, both datasets had 10 outliers each, of which were genuine high or low scores so were of no concern.

Profile Accuracy

Overall, Indian participants had a mean total judgement accuracy score of 7.79 (SD = 2.04) and Non-Indian participants had a mean total judgement accuracy score of 7.67 (SD = 1.50), showing that Indian participants are more accurate overall than Non-Indian participants. Analyses of overall accuracy scores for both sets of participants showed that zero participants in both datasets achieved a maximum judgement accuracy score of 16. The highest score achieved was for participants in the Non-Indian data set was 11 (N = 2, 2.0%), and for participants in the Indian data set the highest score was 13 (N = 3, 3.0%). Judgement accuracy scores for each type of profile and each culture were analysed, and the results of which are shown in Figure 1.

Figure 1
Mean judgement accuracy scores for each type of profile and each culture type (N = 200).

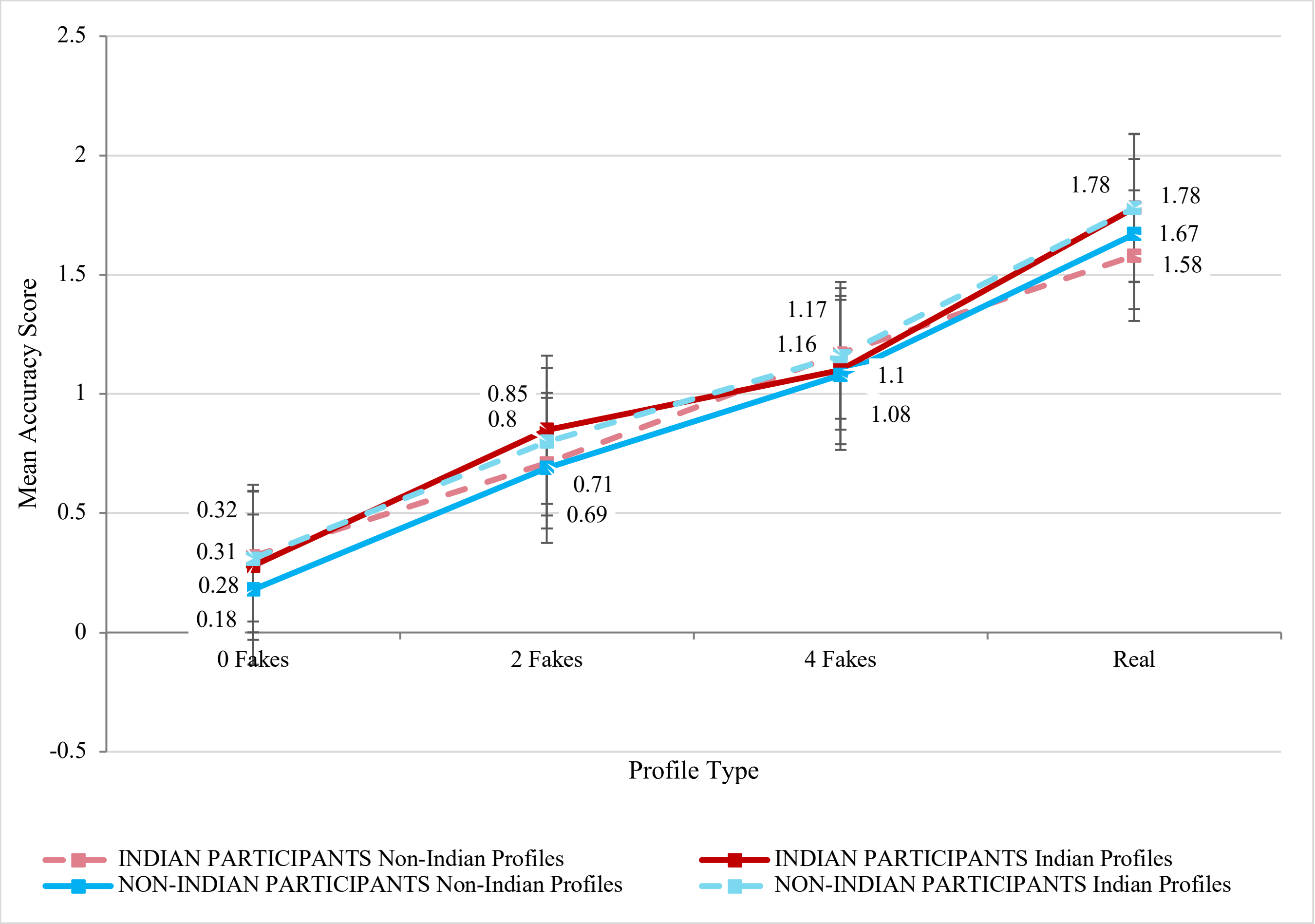


Figure 1 shows an overall linear trend in judgement accuracy with *Real* profiles judged most accurately, and in reference to the fake profiles, *0 Fakes* were judged least accurately and *4 Fakes* most accurately. These findings replicate all previous studies whereby the same linear trend was found, and in this case the trend was also found across both Indian and Non-Indian participants.

Further, figure 1, shows that the Indian participants (represented by the red lines), were more accurate at judging profiles from their own culture (Indian profiles) when those profiles contained *2 Fake* characteristics or were *Real*, and more accurate judging profiles from outside their culture (Non-Indian profiles) when those profiles contained *0 Fake* and *4 Fake* characteristics. Whereas, the Non-Indian participants (represented by the blue lines) were more accurate, across all profile types, at judging profiles from outside their culture (Indian profiles) than profiles from their culture (Non-Indian profiles). The Non-Indian participants achieved the lowest accuracy scores of all participants, for all of the three fake profile types from their own culture. Paired samples t-tests were conducted to measure the mean differences in accuracy across cultures; Indian participants were more accurate at judging Indian profiles (*M* = 4.01, SD = 1.35) than Non-Indian profiles (*M* = 3.78, SD = 0.15), however the mean difference of 0.23 was not statistically significant; *t*(99) = 1.19, *p* =.119, CI [-.15, .61]. Non-Indian participants were more accurate at judging Indian profiles (*M* = 4.05, SD = 0.98) than Non-Indian profiles (*M* = 3.62, *SD* = 1.18), a statistically significant mean difference of 0.43, *t*(99) = 2.76, *p* =.003, CI [.12, .74]. These mean scores suggest that cultural familiarity has an influence on judgement accuracy, but the influence of such varies between Indian and Non-Indian participants.

To analyse whether overall judgement accuracy scores were better than that of chance a series of t-tests were conducted. Focusing first on the participants in the Indian data set, their overall mean judgement accuracy score of 7.79 (*SD* = 2.01), was statistically significantly lower than the chance level of 8; *t*(99) = -1.03, *p* = .153, CI [-.62, .20], with a mean difference of -.21. Results for participants in the Non-Indian data set, their overall mean judgement accuracy score of 7.67 (*SD* = 1.50), was statistically significantly lower than the chance level of 8; *t*(99) = -2.19, *p* = .015, CI [-.63, .03], with a mean difference of -.33. Each of these test's evidence that participants, in either culture, did not perform better than chance when judging profiles.

To understand further the mean differences in accuracy scores, repeated measures ANOVAs were completed. The within subjects' factor was Profile Type which consisted of four levels: *0 Fakes, 2 Fakes, 4 Fakes,* and *Real.* As the data was split across two different data sets, one for Indian participants and one for Non-Indian participants, two separate AVOVAs were conducted for each data set.

In regard to the Indian participants, when testing that the data fit the assumptions of the ANOVA, results of Mauchly's test of sphericity showed a violation, $X^2(27) = 68.93$, p <.001. To correct this, the Greenhouse-Geisser correction was used in following ANOVA analysis. A statistically significant large effect of profile type on judgement accuracy was found; $F(5.97, 591.08) = 75.09$, p <.001, partial $n^2 = .43$. Similarly, when looking at the Non-Indian participants, Mauchly's test of sphericity showed a violation, $X^2(27) = 86.28$, p <.001, as such the Greenhouse-Geisser correction on the degrees of freedom was used to report the results of the ANOVA. A statistically significant large effect of profile type on judgement accuracy was found; $F(5.71, 565.48) = 102.67$, p <.001, partial $n^2 = .51$.

Pairwise comparisons were conducted to analyse the mean differences for each profile type in each culture, within each dataset (Indian participants and Non-Indian participants). Results of which are outlined in Table 1.

Table 1.
Pairwise comparisons of mean accuracy scores for each profile type in both cultures

	Indian Participants			Non-Indian Participants		
Profiles	*M*	*SE*	*95% CI*	*M*	*SE*	*95% CI*
0 Fakes Indian vs. 0 Fakes Non-Indian	-0.04	0.07	[-0.25, 0.17]	0.13	0.06	[-0.06, 0.32]
2 Fakes Indian vs. 2 Fakes Non-Indian	0.14	0.09	[-0.15, 0.43]	0.11	1.00	[-0.21, 0.43]
4 Fakes Indian vs. 4 Fakes Non-Indian	-0.07	0.09	[-0.38, 0.24]	0.08	0.08	[-0.19, 0.35]
Real Indian vs. Real Non-Indian	0.20	0.06	[-0.02, 0.42]	0.11	0.07	[-0.10, 0.32]

Table 1 shows that overall, there were no significant differences in mean judgement accuracy scores between each type of profile in each culture and the culture of the participants (Indian or Non-Indian). Whilst the differences between the specific profile types for each culture were not significant, the overall effect of profile type (as evidenced in the ANOVA's above) is significant, and large for both Indian participants (partial n^2 = .43), and Non-Indian participants (partial n^2 = .51), meaning the profile type can explain a substantial amount of variance in participants' judgement accuracy.

For a more in-depth analysis into the effects of the profiles on participants' decision-making process, specifically when judging a profile as either real or fake and their response bias, Signal Detection Theory (SDT) was used. Overall accuracy for all fake and real profiles, and accuracy for Non-Indian and Indian fake and real profiles, were transformed into hit rates and false alarm scores. Hit rate was calculated by dividing the number of hits (number of accurate judgements) by the number of signal trials (possible correct judgements). The false alarm rate was calculated by the number of false alarms (inaccurate judgements) divided by the number of noise trials (the total number of signal trials incorrectly identified as noise trials). Both types of profile (including those in both cultures), had a hit rate and a false alarm rate, whereby from these a d-prime (*d’*) value and criterion (*c*) score were calculated - *d’* is a sensitivity measure used to indicate participants’ abilities at distinguishing between fake profiles (signals) and real profiles (noise), and *c* is a measure of response bias, specifically whether participants had a stronger tendency to say yes or no (real or fake).

Indian participants overall were able to distinguish signals (fake profiles) from the noise (real profiles), *d’* = 0.93, 95% CI [4.02, 4.84], meaning participants could identify fake profiles as fake. Participants showed a bias to judging the profiles as fake (*c* = -1.17). When Indian participants were judging Non-Indian profiles, participants were able to distinguish between the signals and the noise, *d’* = 0.59, CI [1.94, 2.46], and were biased to judging the profiles as fake (*c* = -0.76). When judging Indian profiles, Indian participants were able to distinguish between the signals and the noise, *d’* = 1.40, CI [1.97, 2.49], and were biased to judge the profiles as fake, more so than Non-Indian profiles, with a more liberal *c* score of -0.95.

Non-Indian participants overall were able to distinguish the fake profiles (signals) from the real profiles (noise), *d’* = 0.90, 95% CI [3.94, 4.50], and displayed a bias towards judging the profiles as fake (*c* = -1.03). When Non-Indian participants were judging Non-Indian profiles, participants were still able to distinguish between the fake profiles and the real profiles, *d’* = 0.27, 95% CI [1.57, 1.77], and showed a bias towards judging the profiles as fake (*c* = -0.41). When judging Indian profiles, the Non-Indian participants again were able to distinguish between the fake and real profiles, *d’* = 0.48, 95% CI [1.69, 1.87], and again showed a bias towards judging the profiles as fake (*c* = -0.43), more so than the Non-Indian profiles.

In summary, both sets of participants (Indian and Non-Indian) were able to distinguish between the signals (fake profiles) and the noise (real profiles) in the study.

Both consistently showed a bias to judging the profile as fake, regardless of the culture of the profile, and this bias was stronger when judging profiles from their *own* culture.

Judgement accuracy findings overall show that accuracy differed between the two different cultures (Indian and Non-Indian); Indian participants overall had higher judgement accuracy scores than Non-Indian participants, as such, H1 can be accepted. Additionally, Indian participants had higher levels of accuracy when judging profiles from their own culture, however Non-Indian participants had lower levels of accuracy when judging profiles from their own culture and were most accurate when judging Indian profiles. These results also find evidence for H2 as a difference in accuracy scores for profile judgements of real and fake profiles was found. Hypothesis 3 is also supported as again the same linear trend in judgement accuracy across the profile types was found; *0 Fakes* were judged least accurately of the fake profiles, and *4 Fakes* the most, with *Real* profiles judged most accurately overall. Interestingly, this trend was also replicated across cultures.

Self-Reported Accuracy

Participants were asked after completing all profile judgements to report how confident they felt about the accuracy of their judgements by selecting a rating on a scale from 1 (Unconfident) – 7 (Confident), with 'Neutral' in the middle (4). Indian participants (N = 100) mostly reported feeling *Moderately Confident* (N = 38), or *Slightly Confident* (N = 28). Only two of the Indian participants reported feeling *Unconfident.* Similarly, Non-Indian participants mostly reported feeling *Slightly Confident* (N = 33), or *Moderately Confident* (N = 30), and again only two reported feeling *Unconfident*. Indian participants overall reported a higher level of confidence in their judgement accuracy than Non-Indian participants.

To understand whether particular levels of self-reported accuracy have a relationship with actual judgement accuracy scores for both cultures, multiple regressions were conducted with overall fake accuracy scores and real accuracy scores for each culture's profiles. Results from each model shows that there are no significant coefficients (*B),* meaning participants self-reported accuracy is not predictive of actual accuracy, a finding consistent across both cultures. Additionally, each of the models for Indian participants were non-significant (Fake accuracy, *F(*6, 93) = 1.15, *p* = .338; Real accuracy, *F(*6, 93) = 0.38, *p* = .890), and each of the models for Non-Indian participants were non-significant (Fake accuracy, *F(*6, 93) = 0.67 , *p* = 0.67; Real accuracy, *F(*6, 93) = 1.06 , *p* = 0.39).

As no significant relationships were found between participants' self-reported accuracy and actual judgement accuracy, H4 cannot be accepted.

Manipulated Characteristics of Profiles

To assess whether the individual manipulated characteristics of the fake profiles had an effect on participants' judgements (i.e., whether they judged a profile as real or fake), and the accuracy of said judgements, multiple general linear models (*glmer*) were conducted in R using the *'lme4'* package. Each of the seven factors (*Photo Type, Photo Number, Bio, Intro, Posts Content, Number of Comments, Number of Likes*) were entered into the model, with 'Prolific ID' as a random effect, and again in a different model with 'Prolific ID' with 'Profile Number' as a nested random effect. The addition of 'Profile Number' as a nested effect statistically significantly improved the fit of the model at *p* <.001 level, and this was the case for all models conducted, thus the models reported below include both 'Prolific ID' with 'Profile Number'.

Table 2 presents the results of the effects of each manipulated profile characteristic on Indian participant's judgements (Model 1) and accuracy of judgements (Model 2).

Table 2.
Results from 'glmer' Model's 1 & 2 where Indian participants' judgement and accuracy are regressed on the manipulated profile characteristics.

	Indian Participants									
	Model 1 – Judgement [b]					Model 2 – Accuracy [c]				
Predictors	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*	
Fixed effects										
Intercept	-2.16	0.23	-2.60	-1.71	<.001***	-0.83	0.26	-1.35	-0.32	<.001***
Photo Type [a]	2.91	0.29	2.34	3.47	<.001***	2.48	0.35	1.80	3.16	<.001***
Number of Photos [a]	0.20	0.26	-0.32	0.72	.453	-0.27	0.33	-0.92	0.39	.420
Bio [a]	-0.03	0.26	-0.55	0.49	.905	-0.38	0.34	-1.04	0.28	.264
Intro [a]	0.21	0.26	-0.30	0.73	.416	-0.22	0.33	-0.87	0.44	.518
Post Content [a]	0.56	0.26	0.04	1.07	.034*	0.12	0.33	-0.53	0.77	.719
Number of Comments [a]	0.58	0.27	0.06	1.10	.028*	0.24	0.34	-0.43	0.90	.486
Number of Likes [a]	0.24	0.26	-0.28	0.76	.370	0.03	0.34	-0.63	0.70	.917
Random effects										
Residual Variance (σ^2)	3.29					3.29				
τ_{00} PROLIFICID	0.91					0.21				
τ_{00} PROFILENUM	1.12					2.42				
Intraclass Correlation Coefficient	0.38					0.44				

Note. Number of Participants = 100, Number of Profiles = 148, Number of Observations = 1600. **p* = .05, ** *p* =.01, *** *p*<.001.
[a] Model 1: 0 = Judgement of Real, 1 = Judgement of Fake; Model 2: 0 = Non-Accurate Judgement, 1 = Accurate Judgement. [b] Conditional R^2 = .55. [c] Conditional R^2 = .52

Table 2 shows that *Photo Type* is a highly significant predictor of both participants' judgements and accuracy, specifically, if *Photo Type* has been manipulated in the profile being judged, participants are more likely to judge that profile as fake (B = 2.91) and that judgement of fake is more likely to be accurate (B =

2.48). Additionally, if *Posts Content* and *Number of Comments* had been manipulated in the profiles, Indian participants were more likely to judge that profile as fake (*B* = 0.56, *B* = 0.58 respectively). However, these factors were not also predictive of participant's accuracy, suggesting perhaps an overreliance on these factors when making their judgements – even if these factors had not been manipulated on the profiles participants are still using them as an area of the profile to inform their judgement, thus resulting in an inaccurate judgement.

Table 3 presents the results of the effects of each manipulated profile characteristic on Non-Indian participant's judgements (Model 3) and accuracy of judgements (Model 4).

Table 3.

Results from 'glmer' Model's 1 & 2 where Indian participants' judgement and accuracy are regressed on the manipulated profile characteristics.

	Non-Indian Participants									
	Model 3 – Judgement [b]					Model 4 – Accuracy [c]				
Predictors	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*	
Fixed effects										
Intercept	-2.08	0.17	-2.41	-1.75	<.001***	-0.94	0.25	-1.44	-0.44	<.001***
Photo Type [a]	2.45	0.23	2.00	2.90	<.001***	2.36	0.33	1.72	3.00	<.001***
Number of Photos [a]	0.31	0.22	-0.11	0.73	.152	-0.05	0.32	-0.68	0.57	.865
Bio [a]	0.08	0.22	-0.35	0.51	.709	-0.35	0.32	-0.98	0.28	.279
Intro [a]	0.41	0.22	-0.02	0.83	.061	0.10	0.32	-0.53	0.73	.753
Post Content [a]	0.38	0.22	-0.05	0.81	.081	0.06	0.32	-0.58	0.69	.864
Number of Comments [a]	0.33	0.22	-0.09	0.76	.124	-0.11	0.32	-0.74	0.52	.731
Number of Likes [a]	0.36	0.22	-0.06	0.79	.093	0.06	0.32	-0.57	0.69	.860
Random effects										
Residual Variance (σ^2)	3.29					3.29				
τ_{00} PROLIFICID	0.10					0.18				
τ_{00} PROFILENUM	0.64					2.31				
Intraclass Correlation Coefficient	0.18					0.37				

Note. Number of Participants = 100, Number of Profiles = 148, Number of Observations = 1600. **p* = .05, ** *p* = .01, *** *p*<.001.[a] Model 1: 0 = Judgement of Real, 1 = Judgement of Fake; Model 2: 0 = Non-Accurate Judgement, 1 = Accurate Judgement. [b] Conditional R^2 = .41. [c] Conditional R^2 = .21

Table 3 shows that *Photo Type* is the only predictor for both Non-Indian participants' judgements and accuracy. When *Photo Type* has been manipulated on the profiles, Non-Indian participants are more likely to judge the profile as fake (*B* = 2.45), and that judgement of fake is more likely to be accurate (*B* = 2.36).

To understand whether the culture of the profile had any effect on participants judgement and accuracy, and whether the factors relied upon differed between cultures, *Culture* was added in to each of the above models as a fixed effect. In regard to Indian participants, no statistically significant effects of *Culture* were found on either judgement (*B* = -0.08, *SE* = 0.26, 95% CI [-0.54, 0.37], *p* = .712) or accuracy (*B* = -0.08, *SE* = 0.30, 95% CI [-0.51, 0.68], *p* = .785). Even with the addition of *Culture,* the significant manipulated characteristics remained the same for model 1 (*Photo Type, Posts Content, Number of Comments)* and model 2 (*Photo Type).* Similarly, for Non-Indian participants, no statistically significant effects of *Culture* were found in the judgement model (*B = 0.24* , *SE = 0.19* , 95% CI [-0.14, 0.62], *p = .219*), or the accuracy model (*B* = 0.30, *SE* = 0.29, 95% CI [-0.27, 0.88], *p* = .298).

Overall, the models outlined above in Table 2 and Table 3, replicate findings from earlier studies in this research, in that they evidence that some of the manipulated characteristics are significant predictors of participants' judgements and accuracy of said judgements. As such, H5 and H6 can be accepted. Additionally, it was expected that *Photo Type* would be the strongest predictor of both judgements and accuracy of judgements (H7). This is the case across both cultures in the four models outlined above, thus H7 is accepted. Of importance here is the finding that *Culture* did not significantly influence participant's judgement or accuracy, for either Indian participants or Non-Indian participants, and the factors participants relied upon did not differ between the two cultures in the study. This suggests that participants used similar cues to judge the profiles regardless of their culture to the culture of the profile, meaning accurate judgements of fake Facebook profiles are not affected by cultural variables.

Social Media

Participants were asked a series of questions in relation to their use of social media to assess whether their usage had an effect on their judgement accuracy. It is expected that the time spent on social media per day will have an effect on the judgement accuracy of Indian participants (H8).

Platforms

To further understand how participants use social media, they were asked to select the social media platforms they use, and rank these from most to least used. Seven options were available: *Facebook, Twitter, Instagram, Snapchat, TikTok, YouTube, and Other*. The most popular platform selected by Indian participants was *YouTube* (N = 86), closely followed by *Facebook* (N = 76) . Of these 76, 11 ranked Facebook as their most used platform. Thirteen Indian participants selected *Other* and reported using the following platforms: Reddit, LinkedIn, Telegram, Discord, WhatsApp, and Twitch. In regard to the Non-Indian participants, *YouTube* was the most used platform (N = 81), followed by *Instagram* (N = 16) and *Facebook* (N = 72). Of the 72 who use Facebook, 19 ranked it as their most used platform. Ten Non-Indian participants selected *Other* and detailed using WhatsApp, BeReal, Reddit, Waveful, LinkedIn, Pinterest, Discord, and Wykop.pl.

Purposes

Participants were presented with 12 different purposes for social media use and were asked to select one or more of the specified reasons as to why they use social media (Appendix A). For Indian participants the most popular purpose selected was *Watching videos (TV/Films/YouTube etc.)* (N = 92), followed by *Socialising with friends/keeping in touch* (N = 85). Participants were also provided with the option of selecting *Other*, of which only 2 did, one detailed using social media for "Checking for local events or sales at store". Non-Indian participants' most popular purpose was *Socialising with friends/keeping in touch* (N = 87) which was very closely followed by *Watching videos (TV/Films/YouTube etc.)* (N = 86). Five participants selected *Other*, detailing that they use social media for: "Keeping up with and reading about hobbies and interests", "Networking", and "For learning purposes".

Daily Usage

Of the 100 Indian participants, 96 reported they were regular users of social media. When asked about how long they spend on social media per day, the most frequently reported number of hours spent was *1-2 hours* (N = 37) followed by *2-3 hours* (N = 25). Fourteen participants reported using social media for *Less than 1 hour* per day. Of the 100 Non-Indian participants, 29 report spending *1-2 hours* per day on social media, followed closely by *4+ hours* (N = 28). Only six participants reported using social media for *less than one hour*.

To investigate the effects of time spent on social media on participants' judgement accuracy, multiple regressions were conducted using fake and real profile accuracy scores for both Indian and Non-Indian participants. The predictor *Less than one hour* was used as the constant. The multiple regressions for hours spent on social media are presented below in Table 4.

Table 4.
Multiple regression for social media time predictors of real and fake judgement accuracy for each culture (Indian and Non-Indian).

	Indian Participants						Non-Indian Participants					
Predictors	Fake Profile Accuracy			Real Profile Accuracy			Fake Profile Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Hours spent on social media per day [a]	.052			.035			.079			.056		
Constant		4.21	0.55		3.00	0.22		5.50	0.57		3.00	0.28
1-2 Hours		0.35	0.64		0.35	0.26		-1.71**	0.63		0.48	0.31
2-3 Hours		0.55	0.69		0.48	0.28		-1.12	0.65		0.52	0.32
3-4 Hours		-0.91	0.79		0.46	0.32		-1.31	0.67		0.69*	0.33
4+ Hours		0.60	0.83		0.46	0.34		-1.21	0.63		0.32	0.31

Note. [a]*df* = 4,95. ***p < .01, *p < .05.*

Table 4 shows that the only significant coefficients are for Non-Indian participants. Spending *1-2 Hours* on social media per day significantly decreases judgement accuracy of fake profiles (*B* = -1.71) and spending *3-4* hours per day significantly increases judgement accuracy of real profiles. However, both regression models for Non-Indian participants are non-significant; Fake accuracy (*F*(4, 95) = 2.02, *p* = .097), Real accuracy *F*(4, 95) = 1.41, *p* = .238. None of the coefficients for Indian participants were significant, and both regression models were non-significant (Fake accuracy, *F*(4, 95) = 1.31, *p* = .271; Real accuracy, *F*(4, 95) = 0.86 , *p* = .492), as such, H8 cannot be accepted as time spent on social media per day does not have an effect on Indian participant's judgements of real or fake profiles.

Previous Experience in Creating a Fake Profile

After judging all 16 profiles, participants were asked to disclose whether they had any previous experience in creating a fake profile (Yes/No answer). If yes, participants were asked to outline their reasons for doing so. Of the 100 Indian participants, 11 answered *Yes* to having previous experience in creating a fake profile. Reasons given for doing so include *anonymity reasons* (" …to have an anonymous profile… and to have curated information feeds for my different profiles", "I didn't want to add any friends or family or post anything, just to look at memes all day"), *personal reasons* ("…when I was a teenager to chate with my classmates", "to promote a business"), *investigative purposes* ("to stalk someone and their friends since I got blocked on my main one"), and *malicious reasons* ("…play pranks on my friends"). Of the 100 Non-Indian participants, 20 answered *Yes* to previous experience creating a fake profile. Reasons given also include: *security/anonymity reasons* (" …I wanted to talk about a topic that was embarrassing for me", "I was very uncomfortable creating a digital profile of myself"), *personal reasons* ("to save some links/photos", "to see messages from the primary school"), *investigative purposes* ("to stalk an ex-lover", "to stalk an ex that we ended on bad terms [sic]"), and *malicious reasons* ("…to annoy my sister").

To investigate whether having previous experience of creating a fake profile can predict accuracy scores, multiple linear regressions were conducted using fake and real accuracy for each culture (Indian and Non-Indian). In regard to Indian participants, no significant coefficients were found in the fake accuracy model (*F*(1, 98) = 0.78, *p* = .378) or real accuracy model (*F(*1, 98) = 0.56, *p* = .456). The same results were found in regard to Non-Indian participants. No significant coefficients were found in the fake accuracy model (*F*(1, 98) = 0.40, *p* = .531) or real accuracy model (*F*(1, 98) = 0.51, *p* = .479). This study did not hypothesise an effect of previous experience creating a fake profile and judgement accuracy scores, however these variables were investigated and controlled for within the study. It is evident that this variable has no effect on judgement accuracy scores for either culture.

Personality

To understand whether participants' personality had an effect on their judgement accuracy, a multiple regression analysis was conducted using the personality traits from the TIPI and SS Scale as the predictors; Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness to New Experiences, and Social Sensitivity score, for total accuracy scores for each culture. As previous studies within this research have found no relationship between personality and profile judgement accuracy, a relationship was not hypothesised to be found, however the variables were still controlled for to analyse whether any variance in judgement accuracy scores can be attributed to personality variables.

No significant effect of personality variables were found on total accuracy scores for Indian participants (*F*(6, 93) = 2.86, *p* = .097), or for Non-Indian participants (*F*(6, 93) = 1.32, *p* = .151). None of the coefficients were statistically significant in either model. As such, it can be concluded that personality variables had no effect on participants' total judgement accuracy scores for either the Indian culture or the Non-Indian culture.

Post-Hoc Analysis

Heatmaps

As mentioned in previous studies and outlined in more depth in Study 1 (Chapter 2) each profile within the study had a heatmap layer which allowed for analysis in regard to the areas participants were clicking on the profile to inform their judgements. Each of the manipulated characteristics were outlined on the heatmap layer to capture the frequency of clicks in each area. The frequency of the clicks for each characteristic in each cultural profile type are displayed below in Table 5 for both Indian and Non-Indian participants.

Table 5. Table showing the frequency of clicks for each judgement type per manipulated characteristics for each cultural profile types across both data sites (Indian and Non-Indian)

	Indian Participants								Non-Indian Participants							
	Indian Profiles				Non-Indian Profiles				Indian Profiles				Non-Indian Profiles			
	0 Fakes	2 Fakes	4 Fakes	Real	0 Fakes	2 Fakes	4 Fakes	Real	0 Fakes	2 Fakes	4 Fakes	Real	0 Fakes	2 Fakes	4 Fakes	Real
Photo Type	226	130	132	164	153	140	146	139	197	143	138	144	151	131	159	148
Bio	0	4	20	2	26	27	21	0	0	3	26	2	45	31	20	0
Intro	63	50	31	87	47	47	27	54	73	48	37	89	49	39	35	66
Photo Number	31	26	21	25	30	24	18	34	42	37	39	51	48	39	52	51
Posts Content	126	108	117	139	121	115	102	139	130	108	119	137	118	110	107	138
Likes Number	14	20	13	18	11	13	17	10	31	25	25	37	25	26	26	23
Comments Number	11	17	14	14	18	11	13	6	23	23	15	14	33	14	15	2
Other	63	44	47	105	80	86	63	50	77	62	55	112	99	82	73	88
Total per profile type	534	399	395	554	486	463	407	432	573	449	454	586	568	472	487	516
Total per culture	1882				1788				2062				2043			

Table 5 shows that overall, Non-Indian participants had a larger frequency of overall clicks for both Indian profiles (N = 2062) and Non-Indian profiles (N = 2043). Interestingly, both cultures clicked more so on Indian profiles when making their judgements in comparison to Non-Indian profiles.

Photo Type was the characteristic with the highest frequency of clicks across all profile types (*0 Fakes, 2 Fakes, 4 Fakes, Real)* and both cultural types (*Indian profiles and Non-Indian profiles)* for both sets of participants (*Indian and Non-Indian).* This suggests that the visual stimuli on the profiles is relied upon most when determining the authenticity of a Facebook profile and is found cross-culturally. Further, *Photo Type* was clicked upon most, across all profile types and both cultures, when participants in both data sets (Indian and Non-Indian) were judging the profile as *Real.*

Again, as in previous studies, the characteristic of *Posts Content* and the region on *Other* had the second and third highest frequency of clicks respectively. This is true across both cultures and all profile types and was found to be consistent between Indian and Non-Indian participants; *Posts Content* was clicked on a total of 967 times by both sets of participants, and the region of *Other* received 538 clicks by Indian participants and 648 clicks by Non-Indian participants. Both of these regions, for both Indian and Non-Indian participants, received the most clicks when the profile was being judged as *Real.*

Once again, the region of *Other* consisted of inaccurate clicks, i.e. clicks on the profile that fell just outside the heatmap regions outlining the manipulated characteristics. Each of these clicks were manually checked by the researcher to ensure none were in area of the profile not manipulated that would suggest different areas are being relied upon to make a judgement. This was the case; all were inaccurate clicks.

However, due to the inability to overlap the heatmap regions within Qualtrics, the region of *Other* cannot be avoided.

Overall, the frequencies observed in these heatmaps mirror those found in earlier studies within this research, indicating that these are the primary areas relied upon to inform judgements. When considering these findings were also found with the introduction of the culture variable, the robustness of such findings is strengthened.

Discussion

Findings show that overall, a cultural difference was found in accuracy of social media profile judgements between Indian and Non-Indian participants. Indian participants achieved a higher mean accuracy score overall than Non-Indian participants and were more accurate at judging profiles from their own culture (Indian profiles) when those profiles contained *2 Fake* characteristics or were *Real*, and more accurate judging profiles from outside their culture (Non-Indian profiles) when those profiles contained *0 Fake* and *4 Fake* characteristics. Non-Indian participants were more accurate, across all profile types, at judging profiles from outside their culture (Indian profiles) than profiles from within their culture (Non-Indian profiles). A significant effect of profile type was found for both Indian and Non-Indian participants, however there were no significant mean differences between the pairwise comparisons of each profile type (*0 Fakes, 2 Fakes, 4 Fakes,* and *Real)* for each profile culture across both sets of participants (e.g., *0 Fakes Indian profiles vs. 0 Fakes Non-Indian profiles).* Such findings show that culture does have somewhat of an influence on judgement accuracy of Facebook profiles and suggest a level of cultural familiarity may play a role in distinguishing fake profiles from real.

Social Identity theory could provide a theoretical explanation as to why Indian participants were more accurate at judging Indian profiles. Tafjel (1978, p.63) coined the term social identity, defining it as "that part of the individual's self-concept which derives from knowledge of membership in a social group (or groups) together with the value or emotional significance attached to that membership". The overarching thread of the theory is that a persons' social identity can lead to a bias towards the in-group (members of their own group), whereby humans have a tendency to favour their own group over that of another.

When considering the role of culture in social identity, Hopkins and Reicher (2011) state it is impossible to talk of social identity without referring to culture, yet social psychology continues to fail to recognise the importance of culture. Of those who have researched the impact of culture on social identity have found that in-group bias in a universal concept (Fisher & Derham, 2022). Specifically, *collectivism* is associated with higher in-group bias, as collectivist individuals or groups are more likely to worry about identifying with the in-group (Hinkle & Brown, 1990). In other words, *collectivism* is strongly related to in-group favouritism (Yamagishi, Jin, & Miller, 1998). This could help explain why Indian participants (those from a *collectivist* culture) were more accurate at judging profiles from an Indian culture as they were biased with their in-group. Although, having higher accuracy in this context means they could accurately identify a profile as fake, and it could be argued that in-group favouritism could manifest as inaccurate judgements as the Indian participants could be biased to favour their own culture by believing the profiles to be real.

Expanding upon this, a finding of great interest in this study is that Non-Indian participants were most accurate at judging Indian profiles rather than profiles from their own culture (Non-Indian profiles). In fact, Non-Indian participants had the lowest accuracy scores recorded in the study, and these were for all types of fake Non-Indian profiles. These results suggest that perhaps the Non-Indian participants (*individualists)* were the ones who were biased towards their in-group, meaning they judged the majority of Non-Indian fake profiles inaccurately as real as a result of favouring the profiles from their own culture. Consequently, Indian profiles were judged more as fake resulting in higher accuracy scores of Indian profiles. However, due to the lack of literature regarding in-group biases between cultures in a social media context it is difficult to draw inferences in relation to the findings of this study. It does however point out the need for further research to understand how cultures behave online, specifically relating to how culture effects judgement and decision-making in regard to their culture or that of another.

With focus on the manipulated characteristics, *Photo Type* was the only predictive factor of both participants' judgement and judgement accuracy for Indian and Non-Indian participants. The factors of *Posts Content* and *Number of Comments* were also predictive of participant judgement for Indian participants; however, they were not predicative of Indian participants' judgement accuracy, suggesting that Indian participants over-rely on these characteristics of the profiles to inform their judgements, of which are not necessarily accurate.

Photo Type has consistently been found throughout this research to be the area of the profile participants rely on most to inform their authenticity judgements. This effect has now also been found cross-culturally; the *Photo Type* region had the highest frequency of clicks for all profile types, and both cultural profile types, by both Indian and Non-Indian participants, specifically when judging the profile as real. Several researchers have investigated how images, and more specifically profile pictures, are perceived or used to express the self across cultures.

Visual representations of the self through photographs can reveal certain aspects of our lives such as our values and what is important to us (Humphreys, 2018, p.60). Culturally, visual images in Western cultures are considered as a more natural form of self-representation than words and are therefore considered to be more truthful (Starret, 2003), suggesting that participants rely on the images (*Photo Type)* to provide the most accurate representation of the profile owner, and such presentation determines their level of trust of the image. Related works in consumer behaviour has shown that the addition of a profile picture of a consumer reviewer alongside the review text gained more trust in the review (Lu, Fan, & Zhou, 2016), and Airbnb guests infer trust in regard to the Airbnb host from the hosts profile photo (Ert, Fleischer, & Magen, 2016). Although not specifically related to profile photos on social media, these studies evidence that profile photos elicit a level of trust in decision-making processes, thus could help to explain the reliance (across both cultures) on *Photo Type* to make an authenticity judgement.

However, with specific reference to social media, Pelled, Zilberstein, Pick, Patkin, Tsironlikov and Tal-Or (2016) found that on Facebook profiles visual cues were less dominant than textual cues, whereas Edwards, Stoll, Faculak and Karman (2015) found that American LinkedIn users were judged as being more competent and socially attractive when they display a profile photo compared to those without a profile photo. Additionally, Van Der Heide, D'Angelo and Schumaker (2012) found that when the profile image contains an extroverted personality (portrayed by an individual socialising), the perceiver does not question their perception of the person in the image, i.e., they trust their first impression. However, when the cues from the image are seen as negative, more information is needed which is then obtained from the textual information. The mixed results in regard to social media profile images, and the levels of trust associated with them, suggest that further research is needed – how one represents oneself on a particular social media platform differs when on another platform (Davidson & Joinson, 2021), which might suggest that the reliance and perceived trustworthiness of the profile image differs across platforms.

In regard to the cultural differences of the content of the photographs, Zhao & Jiang (2011) found that when visual imagery is used, different cultures present themselves in different ways. Conservative cultures, otherwise known as *collectivist* or *tight*, posted more neutral photos of themselves, if any at all, and less conservative cultures, or *individualistic* or *loose*, posted less discrete imagery. In slight contrast to this, Dou (2011) found that individuals in high-context (*collectivistic)* cultures use more visual stimuli on their Facebook profiles than those in low-context (*individualistic)* cultures. However, Huang & Park (2013) found when studying Facebook photographs of East Asians and Americans that they showed similar presentations in regard to number of people in their profile pictures. Again, the literature provides mixed results in regard to cultural presentation in profile images and does not provide a clear answer to help explain why both cultures relied most on *Photo Type* to make their authenticity judgement of the profile. However, it does highlight the need for further research in this area, and perhaps this study could be considered as the first one to delve further into this topic from a profile judgement angle.

Perhaps the findings in regard to *Photo Type* are not at all surprising considering the recent shift on social media from text-focused to visual-oriented stimuli (Li & Xie, 2019). However, it is important, that when interpreting the results of this study, to be mindful that Facebook was developed in a western context meaning it's structure and assumptions are embedded in western prerogatives and values (Peters, Winschiers-Theophilus, & Mennecke, 2015), and the platform design features are associated with the culture in which the platforms were produced, and are designed to tailor to that culture (Zhao, Shchekoturov, & Shchekoturova, 2017). Further, Papacharassi (2009) reported that self-presentation and social interactions online are shaped by the design of the platform, or the architectural options offered by the platform.

When measuring the effects of time spent on social media per day on participants' judgement accuracy, minimal significant effects were found for Non-Indian participants only: spending *1-2 Hours* on social media per day significantly decreased judgement accuracy of fake profiles and spending *3-4* hours per day significantly increased judgement accuracy of real profiles. As mentioned previously, there have been several studies regarding Facebook addiction in India (Masthi, Cadabam & Sonaksi, 2015) hence the hypothesis that time spent on social media per day would have an effect on Indian participant's judgement accuracy (H8). This however could not be accepted based on the results. In fact, Indian participants most common time period was *1-2 hours* per day on social media. However, a universally accepted definition of social media addiction is yet to be finalised. Some researchers believe addiction to be spending more than nine hours per day on social media (Chegeni et al., 2021), whereas others believe it to be between two and three hours per day, and it is important to remember that frequent use of social media does not always indicate an addiction to social media (Griffiths, 2010).

Overall, these results do not support current literature on there being a Facebook addiction issue in India, particularly when considering most participants reported spending only 1-2 hours per day on social media. Although, this study had only 100 Indian participants and so is not wholly representative of the wider population and didn't directly measure for Facebook addiction. Future replications of this cultural study could replace the variable *time spent on social media per day* with the *Facebook*

Addiction Symptoms Scale (Andreassen, Torsheim, Brunborg, & Pallesen, 2012) to directly measure levels of addiction and how this might effect accurate identification and judgements of fake social media profiles.

Again, as in all previous studies conducted thus far within this research, the linear trend in judgement accuracy across the four different types of profile was again found and also found cross-culturally. Such finding suggests that the fake profiles are consistently judged in the same way regardless of whether they are from the participant's own culture or different culture. Profiles with *0 Fakes* are consistently judged the least accurately, and in this case, this was true for both Indian and Non-Indian profiles, suggesting that human judgement of fake profiles may not be accurate without any 'obvious cues' as to the 'fakeness' of the profile.

A methodological limitation of this study is in relation to heatmap regions and the frequency of clicks in said regions. The adoption of the new Facebook profile layout for the profiles used within this study meant that the characteristic *Bio* was not accurately captured. The Facebook layout used did not have a specific space for a *Bio* region to be outlined if the *Bio* wasn't present on the profile. This is due to Facebook incorporating *Bio* under the *Intro* section. This means that those fake profiles who do not have a *Bio* shown do not have a *Bio* heatmap region, therefore if any participant clicks in that area to indicate it is an area that informed their judgement (i.e., absence of a *Bio* denoted *fake* profile), then any clicks in that area would be encapsulated under *Intro*. However, as both *Bio* and *Intro* are not the areas of the profile relied upon heavily in this study, this issue does not have a major impact on the validity of the study itself, although it does mean that any inferences drawn from this, regarding the frequency of clicks in the *Bio* and *Intro* regions, are to be taken with caution.

A further, and rather important, limitation of this study is regarding the cultures used. When referring to the Non-Indian participants, these are participants that were not born in India and therefore are not considered as Indian. However, under the umbrella of Non-Indian participants are from locations such as Spain, Mexico, Portugal, South Africa, Poland, Brazil, Italy, Chile, Zimbabwe, Iraq, United States of America, and so on. The issue with this is that each of these countries could be considered a different culture type based on the *collectivist/individualist* and *tight/loose* paradigms. For example, United states of America are known as having an *individualist/loose* culture (Gelfand, 2018, p.28), whereas Iraq is considered as a *collectivist* culture (Hofstede, 1991), a culture similar to that of India. The combination of all of these culture types under the umbrella of Non-Indian means that any comparisons against the Indian participants, or *collectivist* versus *individualistic* is questioned. Additionally, of the 100 Indian participants, zero still currently live in India, rather they live in some of the following countries: Japan, USA, Germany, United Kingdom, Australia, New Zealand, Canada and so on. Some participants had only moved from India one month ago whereas others relocated 40 years ago. Of importance to note here is that most of the countries Indian participants currently reside in are *individualist* and *tight/loose* cultures, not *collectivist* like India. As such, any inferences taken from this research must be done so with the caveat that Non-Indian participants means exactly that; participants in this data set are not from India, and that Indian participants are those who were born there but no longer live there. No longer residing in the country of the culture being studied means participants may have become accustomed to the way of thinking of another culture and thus start believing the same beliefs as them. It is imperative for future researchers who may replicate this study, to choose participants from the specific cultures focused on within the study, i.e. only UK nationals in comparison to only Indian nationals.

Throughout this discussion recommendations have been made in regard to future researchers who may wish to replicate this study or expand upon it further. Firstly, more in-depth analysis of cultural in-group biases online – do cultures favour stimuli from their own culture when making authenticity judgements? Secondly, *Photo Type* is needed to understand the heavy reliance on this profile characteristic, for example are different cultures using different areas of the image itself (i.e. different content shown in the image) to inform their decision. Some of the literature suggests this might be the case in relation to judgements of emotion based on facial expression cues in images (Masuda, Wang, Ishii, & Ito, 2012; Matsumoto, Hwang, & Yamada, 2010) and perception of the scene in the image; some cultures perceive the scenes in a more holistic manner (Nisbett & Masuda, 2003). However, there is yet to be a study that examines the content of profile photos in the context of an authenticity judgement accuracy task.

Thirdly, measuring Facebook addiction directly rather than through *time spent on social media per day,* may provide further understanding about whether those who are exposed to Facebook more often are more accurate at identifying and judging fake and real profiles. Finally, methodological changes to the way in which culture was measured are important to be considered. Replicating this study using two distinct cultures only will further strengthen the findings from this study, and also help to further understand just how different cultures can be in authenticity judgements of Facebook profiles.

Facebook usage influences, and is influenced by, cultural practices (Peters et al., 2015), an effect that has been observed in this study; an effect of culture on participant judgement accuracy was found. These findings imply that cultural context can impact the accuracy of authenticity judgements of social media profiles. Such findings also evidence another level of complexity in understanding online behaviour, specifically how this presents on social media, and the strategies needed to equip users with the ability to accurately detect fake profiles.

Chapter 7: Study 6

Throughout this body of work thus far, data has shown that participants are most accurate at identifying real profiles, and less so fake profiles, specifically profiles with 0 Fakes of which participants consistently achieved the lowest accuracy score. Findings that also held true across cultures as demonstrated in Study 4 (Chapter 5). As such, this final study expects to find the same difference in accuracy between real and fake profiles (Hypothesis 1), and the same linear trend in judgement accuracy across all profile types whereby *0 Fakes* are least accurately judged, followed by *2 Fakes*, then *4 Fakes* and finally real profiles most accurately judged (Hypothesis 2). These hypotheses can provide a consistently grounded framework on which to carry forward this study beyond what has previously been shown and examine how training can be added as an intervention element to add a 'next step'.

Each study within this research has used different manipulations and conditions to measure participant judgement accuracy of fake and real Facebook profiles, including the introduction of a different type of fake profile (*0 Fakes)* in Study 2 (Chapter 3) and methodological manipulations in relation to this in Study 3 (Chapter 4), introduction of a time limit in Study 4 (Chapter 5), and analysis of performance across cultures using profiles from India (Study 5, Chapter 6). This final study introduces an intervention stage whereby participants will be randomly assigned to one of two groups, where only one of these will receive training on what to look for in a profile that denotes it is either real or fake. The purpose of such is to analyse whether participants can be trained to identify fake profiles accurately. If such a pattern were to be identified, then this could of course have much wider real-world consequences for preventative methods used by big social media companies.

The inclusion of an intervention within research is a widely used method for analysing and comparing participant performance pre- and post-intervention and has been in use since the 1940's (Solomon, 1949). General interventions as part of wider studies can now be found in diverse fields and can range in duration. For example, Sports psychology has examples of intervention studies designed to heighten performance in different fields (Driskell, Sclafoni & Driskell, 2014), or to build resilience through a pressure training intervention (Kegelaers, Wylleman, Bunigh & Oudejans, 2021). Health studies implement pre and post-test interventions for physical health improvements in a multitude of settings, from neonatal healthcare in the Democratic Republic of Congo (Berg, Mwambali, & Bogren, 2022) to the effects of anxiety in the workplace (Saunders, Driskell, Johnston, & Salas, 1996), and de-escalating patient aggression (Nau, Halfens, Needham & Dassen, 2010).

Within psychology there are also numerous examples of the use of interventions in studies. In specific relation to deception, many studies have implemented this methodology, of which Frank & Feeley (2003) discussed in their meta-analysis. The overarching finding within the literature analysed is that deception detection improved measurably when studies implemented training interventions. In a more recent meta-analysis of the same topic, it was further demonstrated that training interventions are consistently used in the study of deception detection, and their use can positively and significantly improve the accuracy of detection (Driskell, 2011).

Training interventions have been further developed and can be applied to the context of social media. *Nudges,* coined by Thaler and Sunstein (2009), are an intervention in decision-making that can steer the decisions made by acting on the decision-makers cognitive biases. An example of such is the placement of healthy foods in the queue of a cafeteria to increase their prominence and accessibility, thus increasing the probability a customer will select them (Congiu & Moscati, 2021). Recent work on *nudges* and the ever-growing issues with misinformation and fake news on social media has resulted in the development of *accuracy nudges*, which are nudges designed to prime people to think about the accuracy of the information they see online (Pennycook, Epstein, Mosleh, Arechar, Eckles & Rand, 2021). Although this concept is relatively early in its conception, studies have evidenced that accuracy nudges do have an effect on decision-making in a social media context (Pennycook & Rand, 2022), specifically decreasing the spread of sharing false information. There is also evidence in its infancy that indicates *accuracy nudges* are effective across cultures (Arechar et al., 2022). However, Allard and Clavien (2023) have found that accuracy nudges in regard to the credibility of scientific communication were not effective in preventing people being influenced by their own biases. Overall, the work in this area is of promise and suggests that individuals can be prompted, or rather 'trained', to question their decision-making and judgements of the accuracy of stimuli on social media.

A further technique used in the attempt to protect social media users by preventing the spread of misinformation and fake news online is *inoculation theory*. Inoculation theory is defined as "a theory postulating that resistance to persuasion, can be created by exposing people to weak persuasive attacks that are easily refuted" (APA Dictionary of Psychology, n.d.). Using a medical analogy of inoculations giving the body a taste of a virus in order to strengthen it, the theory relies on a pre-treatment of exposure to an argument (and the means of making that argument) in order to develop resistance against the message and the way of conveying it. With the idea that prevention is better than a cure, the theory relies on 'pre-bunking' peoples' misconceptions rather than having to ‘de-bunk’ them once those beliefs have already formed (Roozenbeek, van der Linden, Goldberg, Rathje, & Lewandowsky, 2022).

The theory has had multiple applications and is described as the ‘grandparent theory of resistance to attitude change’ (Eagly & Chaiken, 1993, p.561). Developments in the original theory which align it more in the direction of this study are the way in which it has now being applied to online disinformation contexts. This is a relatively recent area of study but has proved to be quite dynamic. In a recent review of psychological inoculation against misinformation, researchers Traberg, Roozenbeek, & van der Linden (2022) reported that the theory has successfully been applied to inoculation against conspiracy theories, and climate change misinformation (van der Linden et al., 2017).

An example of the way that inoculation theory research has now moved into the social media sphere is the creation and use of the ‘Bad News’ game as an intervention against online misinformation (Roozenbeek & van der Linden, 2019). The Bad News game encourages players to create fake news stories and build up a following to maintain their credibility, the idea being that by using the techniques of scammers and those pushing misinformation, the players would have a better idea of when they were being exposed to these same tactics when using social media. The overall consensus in regard to the Bad News game is that it is successful in improving participants’ ability to identify misinformation and is an effect that has also been found cross-culturally (Roozenbeek , van der Linden, & Nygren, 2020). Playing the game has been shown to increase peoples’ awareness of the techniques they were exposed to and also to have increased their judgment confidence (Basol, Roozenbeek, & van der Linden, 2020). Both of these factors are what this study can seek to replicate with the introduction of a training intervention. The discussed research above demonstrates a wide breadth of evidence as to the success and applicability of pre-test post-tests designs, the developing success of *accuracy nudges* in a social media context, and the success of the Bad News game in inoculating users against misinformation online. As such, this study will include a pre-test posts-test training intervention to investigate whether participants can be trained to identify, or be 'inoculated' against, fake profiles.

There are certain elements of inoculation theory which this study can transfer over for use in the training intervention. The current misinformation interventions have used underlying repeated tropes to be targeted in the inoculation – rhetoric, false dichotomies, scapegoating etc. (Roozenbeek et al., 2022), whereas this study can use the pre-existing manipulated characteristics of *Photo Type, Bio, Intro, Photo Number, Posts Content, Number of Comments*, and *Number of Likes,* to produce the same effect. Additionally, the effect of *pre-bunking,* where individuals are taught to identify the target stimuli (fake profiles) before exposure, contributes a methodology to follow, to a degree, for the intervention in this study. As a consequence, rather than explaining afterwards why a particular judgement was wrong, this study will instead seek to expose subjects to methods of fake profile creation and then expose them to fake profiles and measure their judgement accuracy. However, this study will not be following a direct inoculation theory approach, but rather a pre-test post-test design, of which the theory contributes. As participants will be exposed to a series of fake profiles before the intervention it will not strictly be a *pre*-bunking. However, the reasoning behind this is to get a baseline accuracy score that their post training accuracy score can be compared to, therefore meaning the effectiveness of the training is measured.

Based on this literature it is expected that there will be an effect of participant condition (training or no-training) on judgement accuracy, and that the judgement accuracy scores of the participants in the training condition will improve when judging the second set of profiles and thus be higher than that of the participants in the control (no-training) condition (Hypothesis 3).

Additionally, it is predicted that there will be a relationship between participants' self-reported accuracy of their judgements and their actual accuracy (Hypothesis 4). Based on the mixed results in previous studies regarding this relationship, the direction of such cannot be stated. However, with the inclusion of a training intervention it is important to investigate any effects of the training on self-reported accuracy.

In relation to the profiles and specifically the manipulated characteristics, a consistent trend has been found across all studies thus far in that the manipulation of 'Photo-Type' on the profiles is a predictor of both participants' judgement of the profile and their accuracy of said judgement. Thus, this study expects to find these relationships again between the manipulated characteristics and participants' judgements (Hypothesis 5), and the manipulated characteristics and participants' judgement accuracy (Hypothesis 6). Specifically, it is expected, based on all studies conducted in this research thus far, that 'Photo-Type' will be the strongest predictor of both participant's judgements and accuracy of said judgements (Hypothesis 7).

Data from the previous studies in this research have consistently found minimal or non-existent results regarding the relationship between individual differences variables and social media variables on judgement accuracy. As such, the researcher cannot hypothesise any relationships within this study, although the variables will be controlled for. The results of these analyses will be presented after the main analysis.

Method

Participants

An A-Priori power analysis of a repeated measures within-between subjects ANOVA was conducted using G* Power (Faul et al., 2007) prior to data collection to determine the appropriate sample size for this study. The analysis indicated that a sample size of 16 participants would be sufficient to detect a medium effect size of f = 0.25, with an alpha level of α = 0.05 and a power of 1−β = 0.80. However, 300 participants were recruited (150 per intervention group). As outlined in Study 1 (Chapter 2), Study 3 (Chapter 4), Study 4 (Chapter 5), and Study 5 (Chapter 6) the reasoning behind this decision was to enhance the reliability and generalisability of the findings by reducing the errors associated with the estimates, improve the representativeness of the sample, and reduce the risk of Type 1 and Type 2 errors. Additionally, as the design of this study included a within-subjects factor the researcher felt it was important to increase the sample size to 300 participants in comparison to the previous studies within this research (120 - 200 participants) to account for any additional sources of variability in the data.

Three hundred and two participants were recruited on Prolific. Three entries were not included in the final analysis. One participant did not give full consent. One participant experienced technical difficulties where all profile screenshots did not display so judgements could not be given, thus meaning entry was incomplete. One participant entry was rejected due to completion time – the participant completed the study in five minutes, a time under the minimum allowed time by prolific of three standard deviations lower than the average time (M = 36:04 minutes, SD = 18:03 minutes). The total number of participants included in analysis is 299, 146 in the Training Group and 153 in the Control Group.

Participants' ages ranged from 18-62 years, with a mean age of 28.04 years (SD = 0.49) and a mode of 23 years. 132 participants (44.15%) identified as Male, 156 (52.17%) identified as Female, 1 (.33%) identified as Transgender, 1 (.33%) identified as Gender Fluid, 5 (1.67%) identified as Non-Binary, and 4 (1.34%) participants selected 'Prefer not to say'. Ethnicities reported by 298 participants were; Asian or Asian British (N = 6, 2.01%), Black, African, Black British, or Caribbean (N = 78, 26.17%), Mixed or Multiple Ethnic Groups (N = 17, 5.70%), White, including any White backgrounds (N = 178, 59.73%), and Another Ethnic Group (N = 14, 4.70%). A total of 5 (1.68%) participants selected the option of 'Prefer not to say'.

Participants were also asked to provide some location data, including the country they were born in, the country they currently live in, and how long in years they have lived in their current country. Of the 298 participants who entered this information, locations spanned across five continents: Africa, Asia, Europe, North America, and South America. The most popular birth locations were South Africa (N = 77, 25.84%), Poland (N = 58, 19.46%), and Portugal (N = 47, 15.77%). The most popular locations participants currently reside in were South Africa (N = 83, 27.85%), Poland (N = 60, 20.13%), and Portugal (N = 49, 16.44%). 273 (91.61%) participants reported that they were still residing in the same country they were born in, and 25 (8.39%) of participants had moved location.

The six participants who provided their real Facebook profiles for use within the study have an age range of 30-64 years (M = 44), ethnically identify as White British, with three (50%) identifying as Female and three (50%) as Male.

Design

A 4 (Real profiles, 0 Fakes profiles, 2 Fakes profiles, '4 fake' profiles) x 2 (Accurate judgement vs. Inaccurate judgement) x 2 (Training vs. Non-Training) experimental Turing test design will be used to investigate the study hypotheses. The Dependent Variable (DV) is the accuracy of the judgements, and the Independent Variables (IV) are the Facebook profiles and the Training condition. Both variables are between subjects' measures following a repeated measures design.

Measures, Materials, Equipment

Measures

As per both previous studies, the same four self-report questionnaires were administered to the participants online using Qualtrics software; a social media questionnaire specifically created for these research studies, the Ten Item Personality Inventory (TIPI) (Gosling, Rentfrow & Swann, 2003), the Social Sensitivity (SS) scale (Riggio, 1986), and a follow up questionnaire containing two short questions that was also created specifically for the purposes of this research to measure participants' self-reported accuracy of their judgements and previous experience in creating a fake profile (See Appendices A-D for details on each).

Materials

The same Facebook profile screenshots were used in the study as were in study 4; 68 fake profiles containing different combinations of the seven manipulated profile characteristics ('Photo - Type', 'Photo – Number', Bio, Intro, 'Posts – Content', 'Comments – Number', and 'Likes – Number'), and six real profiles obtained from persons known to the researcher. Within the total 68 fake profiles, 12 profiles had '0 fake characteristics', 21 profiles had '2 fake characteristics', and 35 profiles had '4 fake characteristics'. The profiles with 0 Fakes technically had no manipulations of the fake characteristics, as these profiles were created specifically to fool participants into judging the profile as real, i.e., the profiles were created to look as real and authentic as possible. To allow for a level of variability and random assignment to participants, 12 profiles with 0 Fakes were created.

Training materials were created specifically for this research using photo-altering software (Adobe Photoshop) and Canva. The training consists of two different images of a profile screenshot, one real and one fake, with boxes around the edge detailing the specific things to look on the profiles that denote the profile to be either real or fake, accompanied with arrows pointing to the specific area of the profile the description refers to (see Appendix AB). As the training images could not use profiles that will be used within the study, two new profiles were created. The fake profile included *all* of the manipulated characteristics to create an 'ultimate fake'. Using the 4 Fakes characteristic framework (Appendix P), each of the seven manipulated characteristics highlighted as fake were used in this training fake profile, so it could be considered as a '7 fakes' profile. The real profile was the researchers' own profile, with names and identifying information removed as per the real profiles used in the study. The training was created in this way (as static images) to replicate how participants will be viewing the profiles in the study.

Procedure

The procedure near identically replicates that of Studies 2, and 3. All participants were recruited via Prolific.co and redirected to the study on Qualtrics once informed consent was obtained. During the study, all participants were first required to complete three of the self-report measures, the social media questionnaire, TIPI, and SS Scale. Following this, participants were provided with a set of instructions in relation to the profile phase of the study, whereby they were informed that they will see 12 Facebook profile screenshots in a random order and asked to make a judgement as to the authenticity of the profile. Participants were also asked to identify the areas of the profile they used when making their authenticity judgement by clicking on the specific areas of the profile screenshot. Following the completion of the profile phase, all participants were asked to complete the follow-up questionnaire and enter some brief demographic details. As per both previous studies, participants were fully debriefed and provided with both further information on the research and contact details of the researcher should they wish to withdraw or ask any further questions.

Procedurally where this study differs from that of Studies 2 and 3 is participants were randomly assigned into one of two groups, training and no-training, after giving their informed consent. Participants, regardless of the group they were in, followed the procedure outline above; completed three self-report measures, viewed 12 random profiles, and completed the follow-up questionnaire. Depending on the group participants were assigned to, one of two things would then happen. Participants in the training group were informed that they were to see two example profiles, one real and one fake, and to take as much time viewing each image as they wished. However, they were restricted from moving on from each image for the first 30 seconds to ensure that they benefitted from the training. Following the training, participants were then informed that they would see a *different* set of 12 profiles and would be asked to judge the authenticity of said profiles. Once all profiles had been viewed and judged, participants were again given the follow-up questionnaire, however the second version had been slightly altered to ask participants if *after* viewing the training profiles they were confident in their judgement abilities and how accurate they believe their judgements to be *after* viewing the example profiles (training). Participants in the non-training group will follow the same basic procedure, and once they have judged the first set of 12 profiles, they will be informed they will then see *another* set of 12 *different* profiles, followed by the follow-up questionnaire again.

Ethics

This research was fully approved by the ethics committee at Lancaster university on 14th May 2021, under an amendment to the same ethics submission as Studies 1 and 2. All participant data was stored on a secure hard drive, in line with GDPR guidelines, and only accessible to the researchers.

Results

Before analysis, the data was inspected for normality and outliers via visual inspections of histograms and normal Q-Q plots, which showed normal distributions and linear patterns respectively. There were 10 outliers identified in participant accuracy scores that were greater than ±3 standard deviations (SD) of the mean, with 7 being more than -3 SD and 3 more than +3 SD . These outliers were assessed as being genuine values and not data entry or measurement errors, rather they were either particularly high or low accuracy scores, and as such they remain in the analyses. Data from 299 participants was analysed. Of these, 146 were in the Training Group and 153 in the No Training (Control) Group.

Participants saw a total of 24 profiles, 12 pre-intervention and 12 post-intervention. In each set of 12, participants were presented with nine fake profiles, three of each type (0, 2, and 4 fake characteristics), and 3 real profiles. To measure participant performance, they were given 18 different accuracy scores dependent on the type of profile and the stage of the study (pre or post intervention). Each participant had six different overall accuracy scores; *Overall Accuracy* scored out of 24, *Overall Fakes Accuracy* scored out of 18, and *Overall Real Accuracy*, *Overall 0 Fakes Accuracy*, *Overall 2 Fakes Accuracy*, and *Overall 4 Fakes Accuracy*, all scored out of 6. The same process was repeated for the profiles judged pre-intervention and post-intervention, giving each participant a further 12 accuracy scores; *Total Pre/Post Accuracy* both scored out of 12, *Total Pre/Post Fakes Accuracy* both scored out of 9, and *Total Pre/Post Real Accuracy*, *Total Pre/Post 0 Fakes Accuracy, Total Pre/Post 2 Fakes Accuracy*, and *Total Pre/Post 4 Fakes Accuracy*, all scored out of 3.

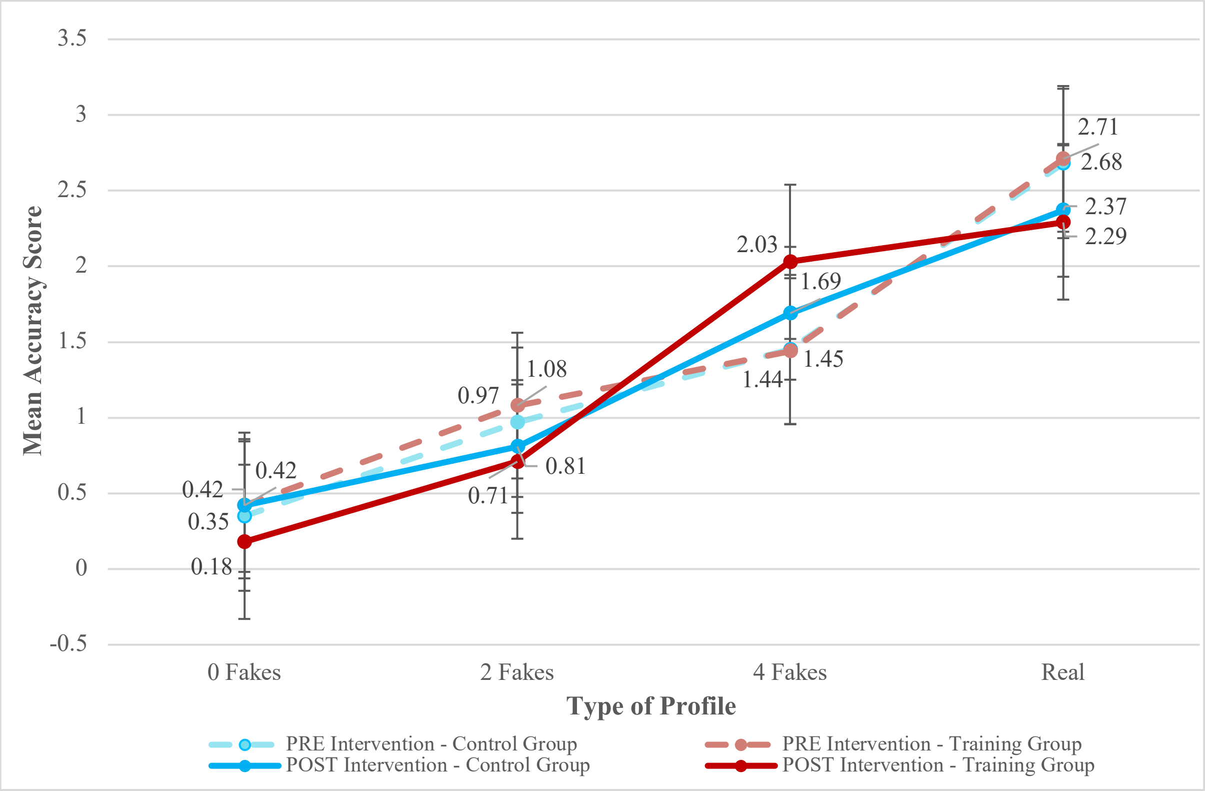
Profile Accuracy

Participants' mean judgement accuracy scores for all fake profile types, and real profiles, were compared. Results show that overall participants in both conditions, pre and post intervention (*Overall Accuracy*), performed better at accurately judging fake profiles (*M* = 5.78, *SD* = 2.47) than real profiles (*M* = 5.03, *SD* = 0.99). The data set was split into the two intervention groups and the same effect was found for total accuracy scores in each group, with participants performing better at accurately judging fake profiles (Training group: *M* = 5.86, *SD* = 2.26; No Training group: *M* = 5.70, *SD* = 2.67, *d* = .06), than real profiles (Training group: *M* = 5.00, *SD* = .95; No Training group: *M* = 5.05, *SD* = 1.03, *d* = .05).

However, when the mean scores were analysed for each profile type and each intervention group, participant accuracy was highest when judging real profiles. Figure 1 shows a linear pattern of mean judgement accuracy across all types of profiles, in both conditions (pre and post intervention), increasing as the number of fake profile characteristics increases. Real profile accuracy scores are the highest for both conditions.

Figure 1

Mean judgement accuracy scores for each type of profile in each of the intervention conditions (N = 299).



Training Group

As shown in Figure 1, post intervention, participants' scores improved for the profiles with *4 Fakes* only, with scores pre intervention of *M* = 1.44 (*SD* = .863), and post intervention of *M* = 2.03 (*SD* = .866) – the post intervention score being the highest for the *4 Fakes* profile type across all conditions. Regarding the fake profiles with zero fake and two fake characteristics, and the real profiles, participants' mean judgement accuracy scores were higher *pre*-intervention for each. Specifically, the pre-intervention scores for these profiles were the highest scores out of both conditions, whereas the post-intervention scores were the *lowest* of both conditions.

Additionally, participants in the Training group only were asked if the training helped them when judging the profiles post-intervention. Of the 146 participants in the Training group, the majority (N = 67, 22.4%) reported that the training ‘Helped a lot’, and 55 (18.4%) reported it ‘Helped a little’. Only 11 (3.7%) participants selected ‘Didn’t help much’, with only 1 (0.3%) selecting ‘Didn’t help at all’.

No Training Group

Participants’ judgement accuracy for the second set of profiles (post-intervention half of the study), improved for the *0 Fakes* and *4 Fakes* profile types only. Accuracy increased markedly more so for *4 Fakes* profiles with a .24 increase in mean scores (Pre-intervention: M = 1.45, SD = .85; Post-intervention: M = 1.69, SD = .91) when compared to a .07 increase for *0 Fakes* profiles (Pre-intervention: M = .35, SD = .57; Post-intervention: M = .42, SD = .63).

Overall analysis – both groups

Analyses of overall accuracy scores for both intervention groups revealed that zero participants achieved an overall maximum judgement accuracy score of 24. The highest score achieved was 18 (N = 2, 0.67%). A linear trend was found when looking at maximum overall accuracy scores of each profile type; zero participants correctly judged all six 0 Fakes profiles, 1 (0.33%) correctly judged all six 2 Fakes profiles, 14 (4.67%) correctly judged all six 4 Fakes, and 108 (36.12%) correctly judged all six Real profiles.

To analyse whether participants’ accuracy scores were better than that of chance, multiple t-tests were conducted using participants’ overall accuracy score, and both the pre and post intervention total accuracy scores. Participants’ overall accuracy, scored out of 24, was M = 10.81, SD = 2.29, a score statistically significantly lower than the chance level of 12; $t(298) = -9.00$, $p < .001$, CI [.40, .64], with a mean difference of -1.19. Similar results were found for both pre- and post-intervention total accuracy scores. Participants' total accuracy pre-intervention, scored out of 12, was *M* = 5.54, *SD* = 1.53, a score statistically significantly lower than the chance level of 6; *t*(298) = -5.17, *p* < .001, CI [.18, .42], with a mean difference of -.46. Post-intervention total accuracy, also scored out of 12, was *M* = 5.26, *SD* = 1.41, a score statistically significantly lower than the chance level of 6; *t*(298) = -9.03, *p* < .001, CI [.40, .64], with a mean difference of -.74. Each score evidence that participants did not perform better than chance either as a whole or in either intervention group.

To further investigate the mean differences in accuracy scores, a two-way repeated measures ANOVA (Analysis of Variance) was conducted. The within subjects' factors were Group and Profile Type. Group consists of two levels: *Training group* and *No Training* group, and Profile Type consists of four levels: *0 Fakes, 2 Fakes, 4 Fakes,* and *Real*. The between subjects' factor was Time – pre- or post-intervention.

When testing that the data fit the assumptions of the ANOVA, results of Mauchly's test of sphericity showed a violation, $X^2(5) = 61.28$, *p* <.001. To correct this, the Greenhouse-Geisser correction was used in following ANOVA analyses. There was a statistically significant two-way interaction between accuracy scores pre- and post-intervention and type of profile judged, *F*(2.64, 783.32) = 40.25, *p*<.001, partial n^2 = 1.00. There was also a statistically significant three-way interaction between accuracy scores pre and post intervention, the type of profile judged, and the intervention group (Training or No Training), *F*(2.64, 783.32) = 7.15, *p*<.001, partial n^2 = .97. However, there was not a statistically significant two-way interaction between the type of profile judged, and the intervention group, *F*(2.66, 789.43) = 2.66, *p* = .054, partial n^2 = .61, or between accuracy scores pre- and post-intervention and the intervention group, *F*(1.00, 297.00) = 1.49 *p* = .22, partial n^2 = .005. Additionally, a non-significant effect was found between intervention group (Training or No Training) and judgement accuracy scores, $F(1,297) = 0.176$, $p = .675$, partial $\eta^2 = .001$. Pairwise comparisons were also conducted between accuracy scores for each profile type pre- and post-intervention. Such comparisons are displayed in Table 1.

Table 1.
Pairwise comparisons of mean accuracy scores for each profile type, pre and post intervention

Measures	*M*	*SE*	*95% CI*
0 Fakes PRE vs. 0 Fakes POST	0.08	0.04	[-0.01, 0.16]
2 Fakes PRE vs. 2 Fakes POST	0.26***	0.06	[0.15, 0.38]
4 Fakes PRE vs. 4 Fakes POST	-0.42***	0.07	[0.29, 0.55]
Real PRE vs. Real POST	0.36***	0.04	[0.28, 0.44]

Note. ****p* < .001.

Table 1 shows mean judgement accuracy scores were higher pre-intervention for profiles with zero and two fake characteristics, and real profiles, however this difference was only statistically significant at the p <.001 level for two fakes profiles and real profiles. A statistically significant difference in accuracy scores for zero fakes profiles from pre-intervention to post-intervention was not found. For the profiles with four fake characteristics, the mean judgement accuracy scores were higher post-intervention, a statistically significant difference to scores pre-intervention. These results suggest that the intervention was only somewhat effective at increasing accuracy scores.

For a more in-depth analysis into participants' decision-making process when judging a profile as either real or fake, and their response bias, Signal Detection Theory (SDT) was used. Overall accuracy scores for fake and real profiles, and total pre and post intervention accuracy scores, were transformed into hit rate and false alarm scores. Hit rate was calculated by dividing the number of hits (number of accurate judgements) by the number of signal trials (possible correct judgements), and the false alarm rate was calculated by the number of false alarms (inaccurate judgements) divided by the number of noise trials (the total number of signal trials incorrectly identified as noise trials). Both types of profile, fake and real, had a hit rate and a false alarm rate, whereby from these a d-prime (*d'*) value and criterion (*c*) score were calculated - *d'* is a sensitivity measure used to indicate participants' abilities at distinguishing between fake profiles (signals) and real profiles (noise), and *c* is a measure of response bias, specifically whether participants had a stronger tendency to say yes or no (real or fake).

Overall, participants were able to distinguish fake profiles (signals) from real profiles (noise), *d'* = 0.92, 95% CI [5.49, 6.06], meaning they were able to identify the fake profiles as fake. Additionally, participants showed a bias to judging a profile as fake (responding 'yes'), with a liberal *c* score of -1.40. Pre-intervention, participants' were again able to identify the fake profiles as fake, *d'* = 0.61, 95% CI [2.68, 3.03], and had a response bias to judging a profile as fake with a *c* score of -0.60, whereas post-intervention participants were less able to identify the fake profiles as fake, *d'* = 0.30, 95% CI [2.76, 3.09], and showed a stronger response bias to judging a profile as fake, *c* = -0.80. These findings further support why participants in the Training Group had the lowest accuracy score for real profiles post-intervention.

To further analyse the effectiveness of the intervention on judgement accuracy, Pearson's correlations between the time spent on each intervention (the training example profiles) and mean accuracy scores post-intervention were conducted. Total accuracy scores, for both fake and real profiles, post-intervention were not significantly correlated with time spent looking/studying the intervention material; Fake accuracy , M = 48.04s, SD = 33.15, *r*(144) = -.13, *p* = .131, Real accuracy, M = 39.83s, SD = 13.39, *r*(144) = -.05, *p* = .515. Further, linear regressions were conducted to analyse if the time spent on studying each of the training profiles predicted judgement accuracy for participants in the training group. However, no significant relationship was found between total post-intervention fake accuracy and time spent studying the fake profile training, *F*(1,145) = .08, *p* = .774. Similarly, no significant relationship was found between total post-intervention real accuracy scores and time spent studying the real profile training, *F*(1,297) = .43, *p* = .515. The results from both the correlation and regression models suggest that any effectiveness of the intervention on judgement accuracy scores was not related to the time participants spent looking at the intervention materials/example profiles.

In summary, there is a statistically significant difference in the accuracy of profile judgments between real profiles and fake profiles. Additionally, profiles with a greater number of fake characteristics are more reliably identified as fake compared to those with fewer characteristics. As a result, hypotheses H1 and H2 can be accepted. However, results have shown that judgement accuracy of participants in the training group improved only for one type of profile following intervention, and they performed worse than control group post-intervention for all other profile types, showing that the intervention was only somewhat effective in increasing judgement accuracy scores. Additionally, the time spent studying the intervention material (example profiles) had no effect on judgement accuracy scores post-intervention. As such, H3 can only be partially accepted.

Self-reported Accuracy

Participants were asked to disclose how confident they feel in relation to their judgements of the profiles by rating how accurate they think their judgements were on a scale from 1 (Unconfident) – 7 (Confident), with 'Neutral' in the middle (4). Participants were asked this question twice; all participants were asked this question after the first set of profiles (pre-intervention), and then again once post-intervention after they had judged the second set of profiles. As a result, there are three different self-reported accuracy scores; pre-intervention (all participants, N = 299), post-intervention Training group (N = 146), and post-intervention Non-Training group (N = 153).

Pre-intervention, 30 participants (10.0%) felt confident their judgements were accurate, with the most frequently selected choice on the scale being 'Slightly confident' (N = 93, 31.1%). Post intervention, in the No-Training group 16 participants (10.5%) reported feeling 'Confident', compared to 22 (15.1%) in the Training group. The most frequently selected choice by participants in the No Training group was 'Slightly Confident' (N = 47, 30.7%), whereas the most frequently selected choice in the Training group post-intervention was 'Moderately Confident' (N = 58, 39.7%). Overall, pre-intervention 194 participants (64.9%) selected a level of confidence in their judgement accuracy ('Confident', 'Moderately Confident', or 'Slightly Confident'). In the Training group 119 participants (81.5%) in the selected a level of confidence compared to only 98 (64.1%) in the No Training group. Participants who received the training self-reported that they feel more confident overall in their judgement accuracy than those in the No Training group.

To understand whether particular levels of self-reported accuracy have a relationship with actual judgement accuracy scores pre and post intervention, multiple regressions were conducted. Table 2 displays the results of these regression models with 'Confident' used as the constant.

Table 2.
Multiple regression of self-reported accuracy and overall judgement accuracy measured both pre- and post-intervention.

	Pre-Intervention [a]			Post-Intervention: Training [b]			Post-Intervention: No Training [c]		
Predictors	Overall Accuracy			Overall Accuracy			Overall Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Self-Reported Accuracy	.027			.033			.011		
Constant		5.27	0.28		5.23	0.28		5.19	0.39
Unconfident		0.96	0.58		-0.73	0.96		0.56	0.86
Moderately Unconfident		0.76*	0.37		-0.31	0.47		0.13	0.49
Slightly Unconfident		0.07	0.46		0.77	0.96		0.19	0.67
Neutral		0.03	0.37		-0.41	0.48		-0.24	0.53
Slightly Confident		0.30	0.32		0.26	0.35		0.22	0.45
Moderately Confident		0.20	0.33		-0.04	0.33		0.13	0.47

Note. [a] N = 299, *df* = 6, 298, [b] N = 146, *df* = 6, 145, [c] N = 153, *df* = 6, 152. * *p* = .05

Table 2 shows that there is only one significant coefficient in the three models; ‘Moderately Unconfident’ is a significant predictor of overall judgement accuracy pre-intervention. Participants who reported feeling ‘Moderately Confident’ pre-intervention had higher mean overall accuracy scores by 0.76. However, the regression model was not statistically significant, *F*(6, 298) = 1.36, *p* = .233. Similarly, the regression model’s post-intervention were not statistically significant for either the Training group (*F*(6, 145) = 0.79, *p* = .579) or the No Training group (*F*(6, 152) = 0.27, *p* = .950). Overall, participants’ self-reported accuracy is not a good predictor of actual judgement accuracy, whether that be pre- or post-intervention. As such, H4 cannot be accepted.

Manipulated Characteristics of Profiles

To assess whether the individual manipulated characteristics of the fake profiles had an effect on participants’ judgements (i.e., whether they judged a profile as real or fake), and the accuracy of said judgements, multiple general linear models (‘*glmer’*) were conducted in R using the *‘lme4’* package. Each of the seven factors (Photo Type, *Photo Number, Bio, Intro, Posts Content, Number of Comments, Number of Likes*) were entered into the model, with ‘Prolific ID’ as a random effect, and again in a different model with ‘Prolific ID’ with ‘Profile Number’ as a nested random effect. The addition of ‘Profile Number’ as a nested effect statistically significantly improved the fit of the model at p <.001 level, and this was the case for all models conducted, thus the models reported below include both ‘Prolific ID’ with ‘Profile Number’.

Model’s 1 and 2, reported in Table 6, measured participants overall judgement and overall accuracy scores against the manipulated characteristics predictors.

Table 6.

Results from 'glmer' Model's 1 & 2 where judgement and accuracy are regressed on the manipulated profile characteristics.

Predictors	Model 1 – Judgement [b]					Model 2 – Accuracy [c]				
	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*	
Fixed effects										
Intercept	2.36	0.19	1.99	2.73	<.001***	-1.19	0.29	-1.75	-0.63	<.001***
Photo Type [a]	-2.17	0.22	-2.61	-1.73	<.001***	1.75	0.34	1.09	**2.42**	<.001***
Number of Photos [a]	-0.67	0.22	-1.11	-0.24	.002**	0.36	0.34	-0.31	1.02	.297
Bio [a]	-0.19	0.22	-0.63	0.24	.383	-0.08	0.34	-0.75	0.58	.804
Intro [a]	-0.29	0.22	-0.73	0.14	.187	-0.04	0.34	-0.71	0.63	.911
Post Content [a]	-0.48	0.22	-0.91	-0.04	.031*	0.13	0.34	-0.54	0.79	.712
Number of Comments [a]	-0.32	0.22	-0.76	0.11	.146	0.01	0.34	-0.66	0.69	.968
Number of Likes [a]	-0.42	0.22	-0.86	0.01	.056	0.13	0.34	-0.54	0.80	.698
Random effects										
Residual Variance (σ^2)	3.29					3.29				
τ_{00} PROLIFICID	0.55					0.15				
τ_{00} PROFILENUM	0.69					1.77				
Intraclass Correlation Coefficient	.27					.37				

Note. Number of Participants = 299, Number of Profiles = 74, Number of Observations = 7176. **p* = .05, ** *p* = .01, *** *p*<.001.

[a] Model 1: 0 = Judgement of Fake, 1 = Judgement of Real; Model 2: 0 = Non-Accurate Judgement, 1 = Accurate Judgement. [b] Conditional R^2 = .46. [c] Conditional R^2 = .44

Table 6 shows that *Photo Type* is a highly significant predictor of both participants' judgements and accuracy, specifically, if *Photo Type* has been manipulated in the profile being judged, participants are more likely to judge that profile as fake (*B* = -2.17) and that judgement of fake is more likely to be accurate (*B* = 1.75). Additionally, if 'Number of Photos' has been manipulated, participants are more likely to judge the profile as fake (*B* = -0.67), however this is not a predictor of participants' accuracy, suggesting that participants may over rely on the *Number of Photos* to make their judgement, i.e., even if the *Number of Photos* has not been manipulated, participants are using this as an area of the profile to make their judgement, thus resulting in an inaccurate judgement. The same effect was found for *Posts Content* (*B* = -0.48).

To assess whether a relationship between the manipulated characteristics and participant judgement and accuracy differed between pre- and post-intervention, the models were ran for a second time using all participants pre-intervention (after viewing the first set of 12 profiles), and each intervention group post-intervention (after viewing the second set of 12 profiles). Table 7 reports these models.

Table 7.

Results from 'glmer' Model's 3 & 4 where judgement and accuracy are regressed on the manipulated profile characteristics both pre-intervention for all participants and post-intervention for participants in the Training and No Training groups.

	Model 3 – Judgement																	
	Pre-Intervention [b]					Post-Intervention Training Group [c]					Post-Intervention No-Training Group [d]							
Predictors	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*			
			LL	*UL*				*LL*	*UL*				*LL*	*UL*				
Fixed effects																		
Intercept	2.26	0.24	1.80	2.72	<.001***	2.99	0.34	2.32	3.67	<.001***	2.12	0.24	1.64	2.59	<.001***			
Photo Type [a]	-2.15	0.28	-2.70	-1.60	<.001***	-2.73	0.44	-3.60	**-1.86**	<.001***	-2.22	0.33	-2.85	-1.58	<.001***			
Number of Photos [a]	-0.27	0.31	-0.88	0.33	.373	-1.06	0.39	-1.82	-0.29	.007**	-0.79	0.29	-1.35	-0.23	.006			
Bio [a]	-0.25	0.30	-0.84	0.33	.395	-0.11	0.43	-0.95	0.72	.793	0.47	0.31	-0.15	1.08	.137			
Intro [a]	-0.24	0.29	-0.81	0.33	.403	-0.16	0.41	-0.96	0.64	.697	-0.14	0.30	-0.73	0.44	.632			
Post Content [a]	-0.27	0.28	-0.82	0.29	.348	-0.98	0.42	-1.80	-0.17	.018*	-0.75	0.30	-1.35	-0.15	.014			
Number of Comments [a]	-0.18	0.29	-0.74	0.38	.519	-1.15	0.46	-2.06	-0.24	.013*	-0.45	0.33	-1.10	0.21	.182			
Number of Likes [a]	-0.36	0.28	-0.90	0.19	.196	-1.06	0.43	-1.91	-0.21	.014*	-0.59	0.32	-1.21	0.04	.065			
Random effects																		
Residual Variance (σ^2)	3.29					3.29					3.29							
τ_{00} PROLIFICID	0.51					0.59					0.51							
τ_{00} PROFILENUM	0.53					0.98					0.47							
Intraclass Correlation Coefficient	.24					.32					.23							

	Model 4 – Accuracy														
	Pre-Intervention [f]					Post-Intervention Training Group [g]					Post-Intervention No-Training Group [h]				
Predictors	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*	Estimate	*SE*	95% CI		*p*
			LL	*UL*				*LL*	*UL*				*LL*	*UL*	
Fixed effects															
Intercept	-0.94	0.40	-1.72	-0.16	.018*	-1.92	0.49	-2.87	-0.97	<.001***	-1.14	0.37	-1.87	-0.42	.002**
Photo Type [e]	1.67	0.47	0.75	2.59	<.001***	2.32	0.63	1.08	**3.56**	<.001***	1.89	0.49	0.92	2.85	<.001***
Number of Photos [e]	0.10	0.52	-0.92	1.11	.854	0.67	0.56	-0.44	1.77	.236	0.44	0.44	-0.42	1.31	.313
Bio [e]	-0.14	0.50	-1.11	0.83	.778	-0.10	0.61	-1.30	1.10	.867	-0.64	0.48	-1.58	0.29	.179
Intro [e]	-0.11	0.49	-1.07	0.84	.815	-0.11	0.59	-1.26	1.05	.853	-0.16	0.46	-1.06	0.74	.732
Post Content [e]	-0.12	0.48	-1.05	0.82	.806	0.61	0.60	-0.57	1.78	.312	0.44	0.47	-0.48	1.35	.351
Number of Comments [e]	-0.24	0.48	-1.18	0.70	.622	0.94	0.66	-0.35	2.23	.152	0.20	0.51	-0.80	1.20	.690
Number of Likes [e]	0.02	0.47	-0.90	0.93	.969	0.77	0.62	-0.45	1.99	.214	0.36	0.49	-0.60	1.31	.464
Random effects															
Residual Variance (σ^2)	3.29					3.29					3.29				
τ_{00} PROLIFICID	0.20					0.01					0.10				
τ_{00} PROFILENUM	1.69					2.31					1.37				
Intraclass Correlation Coefficient	.36					.41					.31				

Note. *p =.05, ** p =.01, *** p<.001.

[a] **0 = Judgement of Fake, 1 = Judgement of Real.** [b] Number of Participants = 299, Number of Profiles = 38, Number of Observations = 3588, Conditional R^2 = .41. [c] Number of Participants = 146, Number of Profiles = 36, Number of Observations = 1752, Conditional R^2 = .59. [d] Number of Participants = 153, Number of Profiles = 36, Number of Observations = 1836, Conditional R^2 = .42

[e] **0 = Non-Accurate Judgement, 1 = Accurate Judgement.** [f] Number of Participants = 299, Number of Profiles = 38, Number of Observations = 3588, Conditional R^2 = .42. [g] Number of Participants = 146, Number of Profiles = 36, Number of Observations = 1752, Conditional R^2 = .55. [h] Number of Participants = 153, Number of Profiles = 36, Number of Observations = 1836, Conditional R^2 = .39

Table 7 shows that consistently in both models, ‘Photo-Type’ was a highly significant predictor of participant’s judgement and judgement accuracy across both groups pre- and post-intervention. When ‘Photo-Type’ has been manipulated in a profile, i.e. profiles display photos showing Celebrities, cartoon characters, landscapes, or artwork, the probability that participants will judge that profile as fake is increased (*B* = -2.15, -2.73, or -2.22 respective to each group), and the probability that these judgements will be accurate is increased (*B* = 1.67, 2.32, or 1.89 respective to each group). Of particular note here is the estimate for the training group is the highest of the three tested conditions for both judgement and accuracy, suggesting that the intervention had a significant effect, specifically the training profiles assisted participants in making more accurate *fake* judgements when *Photo-Type* was a manipulated factor. This effect was also found pre-intervention and for the No-Training group, however this was to a lesser extent as the estimates for each were lower, suggesting that intervention may not be necessary in relation to *Photo-Type.* However, the only other significant results are found post-intervention for the Training group; *‘Number of Photos’* (*B* = -1.06) , *‘Posts - Content’* (*B* = -0.98), *Number of Comments* (*B* = 1.15), and *Number of Likes* (*B* = -1.06) were all significant predictors of participants’ judgement, increasing the probability that participants will judge the profiles that contain these manipulated characteristics as fake. However, these characteristics did not significantly predict judgement accuracy, suggesting that perhaps the intervention was necessary to highlight the areas of the profiles that can be manipulated, but participants who received this intervention over-relied on these areas to make their judgements thus effecting their judgement accuracy.

Overall, these models provide further support for H3 in that there is an effect of participant condition on judgement accuracy, and partial support for H5 and H6 as some of the manipulated characteristics were significant predictors of both participants' judgement and judgement accuracy. However, H7 can be accepted – *Photo Type* was the strongest predictor of participants' judgements and accuracy of judgements, regardless of group or condition.

Again, as in Studies 3, 4, and 5 (Chapters 4-6), this study did not hypothesise any relationships between individual differences variables (personality traits as measured by the TIPI and social sensitivity) or social media variables and judgement accuracy, due to the very minimal, or non-existent, relationships found in all previous studies in this research. However, these variables were still controlled for and the results of which are presented below.

Personality

To understand whether participants' personality had an effect on their judgement accuracy, a multiple regression analysis was conducted using the personality traits from the TIPI and SS Scale as the predictors; Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness to New Experiences, and Social Sensitivity score. As previous studies within this research have found no relationship between personality and profile judgement accuracy, a relationship was not hypothesised to be found, however the variables were still controlled for.

Prior to analysis, linearity was assessed by visual inspection of the scatterplots of each personality variable. The data first had to be 'jittered' to allow for this assessment due to the variables consisting of discrete data that overplotted on one another. Adding the noise with jittering showed a linear relationship between each personality variable and accuracy scores, and further showed homoscedasticity. Additionally, each of the variables had a Durbin-Watson score close to 2, which showed independence of residuals. As all assumptions were met, the multiple regression was deemed a suitable method of analysis. Table 3 shows the results of the multiple regression between personality variables and judgement accuracy scores.

Table 3.
Multiple regression for personality predictors of overall judgement accuracy for each type of profile (pre and post intervention scores combined).

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
TIPI	.014			.019			.014			.039		
Extraversion		-0.01	0.04		-0.03	0.05		-0.08	0.05		0.05	0.04
Agreeableness		-0.07	0.05		-0.13	0.08		-0.04	0.08		0.06	0.05
Conscientiousness		0.02	0.04		0.02	0.06		-<0.01	0.06		-0.10*	0.05
Emotional Stability		0.06	0.05		-0.01	0.07		0.10	0.07		-0.03	0.05
Openness to New Experiences		0.03	0.05		0.12	0.07		0.05	0.07		-0.01	0.05
SS Scale		0.01	0.01		<0.01	0.01		<0.01	0.01		0.01	0.01

Note. df = 6, 292. **p < .05.*

As is evident from Table 3, personality variables are not good predictors of participants' overall judgement accuracy (pre and post intervention) for any of the profile types, as there is only one significant coefficient. The regression model for overall real profile accuracy shows that an increase in conscientiousness is associated with a decrease in accuracy score (*B* = -0.10), meaning the more conscientious a participant is the less accurate they are in judging real profiles as real. However, this significant coefficient is not from a statistically significant model - the multiple regression models for each type of profile were not statistically significant; 0 Fakes, *F*(6, 292) = .70, *p* = .654; 2 Fakes, *F*(6, 292) = .96, *p* = .453; 4 Fakes, *F*(6, 292) = .72, *p* = .638; Real; *F*(6, 292) = 1.95, *p* = .073.

From this, it can be concluded that there is no significant relationship between the personality variables and participants' overall judgement accuracy for any of the profile types.

Social Media

Participants were asked a series of questions in relation to their use of social media to assess whether their usage had an effect on their judgement accuracy.

Platforms

Participants were asked to select the social media platforms they use, and rank these from most to least used. Seven options were available: *Facebook, Twitter, Instagram, Snapchat, TikTok, YouTube, and Other*. Of the participants who selected at least one platform, *YouTube* was selected by the most participants (N = 258, 86.29%), followed by *Instagram* (N = 233, 77.93%) and *Facebook* (N = 212, 70.90%), meaning 212 participants are active users of Facebook and familiar with the workings of the platform. Of these 212, 49 (23.11%) ranked Facebook as their most used platform. The option of *Other* was selected by 46 participants (15.38%) who detailed they use the following platforms: Discord, LinkedIn, WhatsApp, Reddit, Tumblr, 4Chan, Pinterest, Mastadon, VK, and Be Real.

Purposes

Participants were presented with 12 different purposes for social media use and were asked to select one or more of the specified reasons as to why they use social media (Appendix A). The most popular purpose selected was *Watching videos (TV/Films/YouTube etc.)* (N = 261, 87.29%), closely followed by *Socialising with friends/keeping in touch* (N = 246, 82.27%). Participants were also provided with the option of selecting *Other*, of which 16 (5.35%) did so. Those 16 participants detailed the following reasons: “Follow brands and organisations”, “Engage in fandom spaces”, “Promoting my art/writing”, “Writing stories”, “Create and share memes”, “Date”, “Learn new things”.

Daily usage

Of the 299 participants, 293 (98%) reported they were regular users of social media. When asked about how long they spend on social media per day, the most frequently reported number of hours spent was *4+ hours* (N = 84, 28.1%), followed by *1-2 hours* (N = 70, 23.4%). Only 21 participants (7%) reported using social media for *Less than 1 hour* per day.

To investigate the effects of time spent on social media on participants' judgement accuracy, multiple regressions were conducted using overall accuracy scores (for all participants in both intervention groups) for each profile type. The predictor *Less than one hour* was used as the constant. The multiple regressions for hours spent on social media are presented below in Table 4.

Table 4.
Multiple regression for social media time predictors of overall judgement accuracy for each type of profile (pre and post intervention scores combined).

Predictors	0 Fakes Accuracy			2 Fakes Accuracy			4 Fakes Accuracy			Real Profile Accuracy		
	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*	R^2	*B*	*SE*
Hours spent on social media per day [a]	.008			.034*			.003			.006		
Constant		0.52	0.19		1.62	0.28		3.43	0.29		5.19	0.22
1-2 Hours		0.12	0.22		0.11	0.32		-0.13	0.33		-0.25	0.25
2-3 Hours		0.16	0.22		0.10	0.32		-0.19	0.33		-0.10	0.25
3-4 Hours		0.13	0.23		-0.15	0.33		-0.20	0.34		-0.10	0.25
4+ Hours		0.27	0.22		0.52	0.32		-0.05	0.32		-0.23	0.24

Note. [a]*df* = 4,298. $^*p < .05$.

Table 4 above shows that none of the coefficients of any of the predictors are statistically significant, meaning the specific amount of time each participant spends on social media per day does not predict their judgement accuracy scores. However, the 2 Fakes model was statistically significant at the *p* = <.05 level (*F*(4, 298) = 2.62, *p* =

.035), meaning the model overall explains 3.4% of the variance in judgement accuracy scores for profiles with 2 fake characteristics, but this is not specific to one particular predictor (time spent on social media). Each of the models for the remaining profile types were non-significant; 0 Fakes (*F*(4, 298) = 0.58, *p* = .677), 4 Fakes (*F*(4, 298) = 0.21, *p* = .934), and Real (*F*(4,298) = 0.47, *p* = .76). To assess whether there is a relationship between accuracy, regardless of type of profile, and time spent on social media, a further multiple regression was completed using overall accuracy as the dependent variable, which found a non-significant effect of hours spent on social media on judgement accuracy scores for all types of profiles; *F*(4,298) = 1.44, *p* = .22.

To understand if the intervention group and time spent on social media had an effect on accuracy scores, several correlations were conducted with overall accuracy scores, and overall fake and real accuracy scores, for each group (Training or No Training). As the dependent variable is continuous and the independent variable is ordinal, a Spearman's rank correlation test was used. For participants in the training group, no significant correlations were observed between overall accuracy scores (r_s (144) = .08, *p* = .363), overall fake accuracy scores (r_s (144) = .06, *p* = .498), or overall real accuracy scores (r_s (144) = -.02, *p* = .788). Similarly, in the no training group, no significant correlations were found between time spent on social media and overall accuracy scores (r_s (151) = .10, *p* = .215), overall fake accuracy scores (r_s (151) = .11, *p* = .192), or overall real accuracy scores (r_s (151) = -.04, *p* = .672).

Previous Experience in Creating a Fake Profile

After judging all 24 profiles, participants were asked to declare whether they had previous experience in creating a fake social media profile (not specific to Facebook profiles) by answering a Yes/No question. A total of 49 participants (16.4%) reported that they had previous experience creating such profiles, 22 (14.4%) in the 'No

Training' group (N = 153) and 27 (18.5%) in the 'Training' group (N = 146). Participants who selected 'Yes' were asked to provide reasoning behind the creation of the profile. Reasons given include *security/anonymity reasons* ("Wanted to sell something anonymously. Wanted to not upload personal data", "I did not want to reveal my personal information on Facebook"); *investigative purposes* ("Wanted to stalk someone", "….I created a fake account so that I can stalk my boyfriend and also in school we used to catfish other schoolmate", "I wanted to see if my boyfriend at the time would cheat on me"); *personal reasons* ("I simply needed and account where I could be myself without the possibility of my acquaintances knowing. Basically no judgements", "Low self-esteem and paranoia. By hiding behind a veil I could communicate more naturally with those I was curious about"); or *gaming purposes/for fun* ("…so I could play games in a secondary account…", "…just to have fun writing something to my friends", "…just for contests…"). Only 3 participants (1%) reported that they created the profile for *malicious reasons* ("…to troll people at Facebook groups", "…because I wanted to give someone some information and I didn't want them to know I was the source", "just to troll some people").

To investigate whether having previous experience of creating a fake profile can predict accuracy scores, a multiple linear regressions were conducted using overall accuracy score, and overall real and fake scores, split across each intervention group, with 'No' used as the constant. These are displayed in Table 5 below.

Table 5.
Multiple regression of previous experience creating a fake profile and overall judgement accuracy for each type of profile split between intervention groups.

	Training[a]						No Training[b]					
Predictors	Overall Accuracy[c]		Overall Fake Accuracy[d]		Overall Real Accuracy[e]		Overall Accuracy[f]		Overall Fake Accuracy[g]		Overall Real Accuracy[h]	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Experience creating fake profile												
Constant	10.92	0.19	5.29	0.21	5.00	0.09	10.64	0.22	5.57	0.23	5.08	0.09
Yes	-0.29	0.49	-0.33	0.48	0.05	0.20	0.77	0.57	0.94	0.61	-0.17	0.24

Note. [a] *df* =1,145, [b] *df* = 1,152, [c] R^2 = .003, [d] R^2 = .003, [e] R^2 = <.001, [f] R^2 = .012, [g] R^2 = .015, [h] R^2 = .003

As is displayed in Table 5, there are no statistically significant coefficients of the predictor (previous experience in creating a fake profile) and overall judgement accuracy scores for either intervention group. Additionally, each of the models were not statistically significant for the Training group (Overall accuracy, *F*(1, 145) = .41, *p* = .524; Overall Fake Accuracy, *F*(1, 145) = .47, *p* = .493; or Overall Real accuracy, *F*(1, 145) = .05, *p* = .824), or the Non-Training group (Overall accuracy, *F*(1, 152) = 1.83, *p* = .178; Overall Fake Accuracy, *F*(1, 152) = 2.34, *p* = .128; or Overall Real accuracy, *F*(1, 152) = .50, *p* = .481)

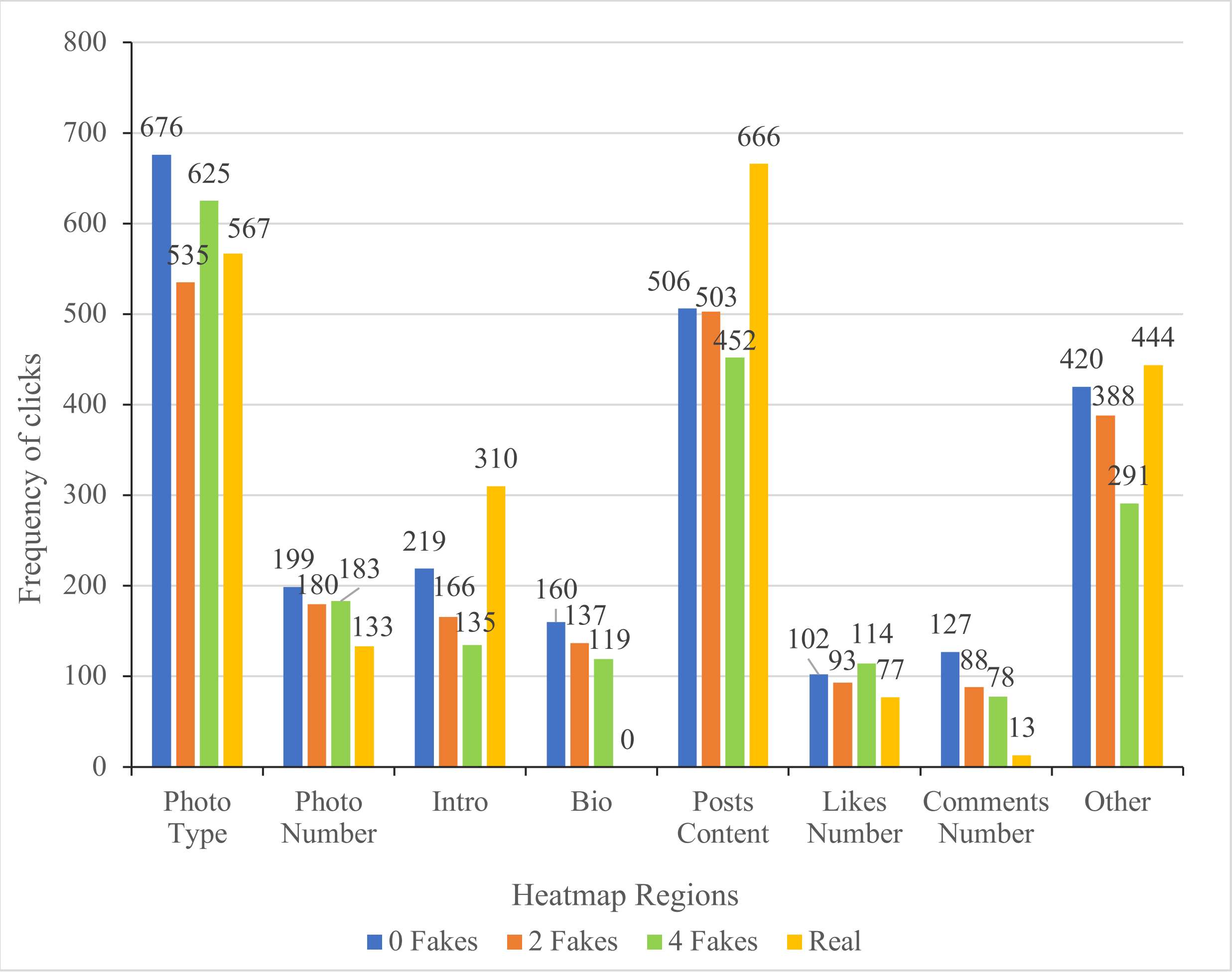
It has been evidenced using both descriptive and inferential testing that overall, there are no relationships between social media predictors, hours spent on social media, or previous experience creating a fake profile, and judgement accuracy scores.

Post-Hoc Analysis

Heatmaps

Each of the profiles used within this research utilised heatmap layers to analyse the specific areas that participants focused on when making judgments about each profile, by recording the location of any clicks made on the profile. To capture these clicked areas, heatmap regions, that were visible only to the researcher during analysis, were added around the manipulated characteristics of each profile (Appendix K). The regions were intentionally made as wide as possible, whilst also avoiding overlap with other regions, to ensure any imprecise clicks that may be on the edge of a region were encapsulated and accounted for. Each click appeared on the profile as a red dot, and participants were informed they were allowed a maximum of ten clicks per profile, a maximum set by Qualtrics software. The collated results of all clicks on each profile are denoted by colours; blue areas represent fewer clicks, or the 'cooler' end of the heatmap, and red represent the highest number of clicks at the 'hotter' end of the scale. The frequency of these clicks across all profiles, for each of the manipulated characteristics, are shown in figure 2 for profiles judged pre-intervention, and figure 3 for profiles judged post-intervention.

Figure 2.
Graph showing the frequency of clicks per manipulated characteristics for all profile types judged pre-intervention.



As evidenced in Figure 2, participants pre-intervention clicked mostly on the regions of *Photo Type, Posts Content*, and *Other* across all profile types when making their judgements of the authenticity of the profiles. The region of *Other* relates to any other area of the profile that is not covered by one of the heatmap regions outlined by the researcher. After manual checking of each of the clicks in these areas it is apparent that they are inaccurate clicks in areas where there is no content, such as the grey space or borders within the profile, rather than clicks on any other specific areas not covered by the specified regions.

Figure 3.

Graph showing the frequency of clicks per manipulated characteristics for all profile types judged post-intervention.

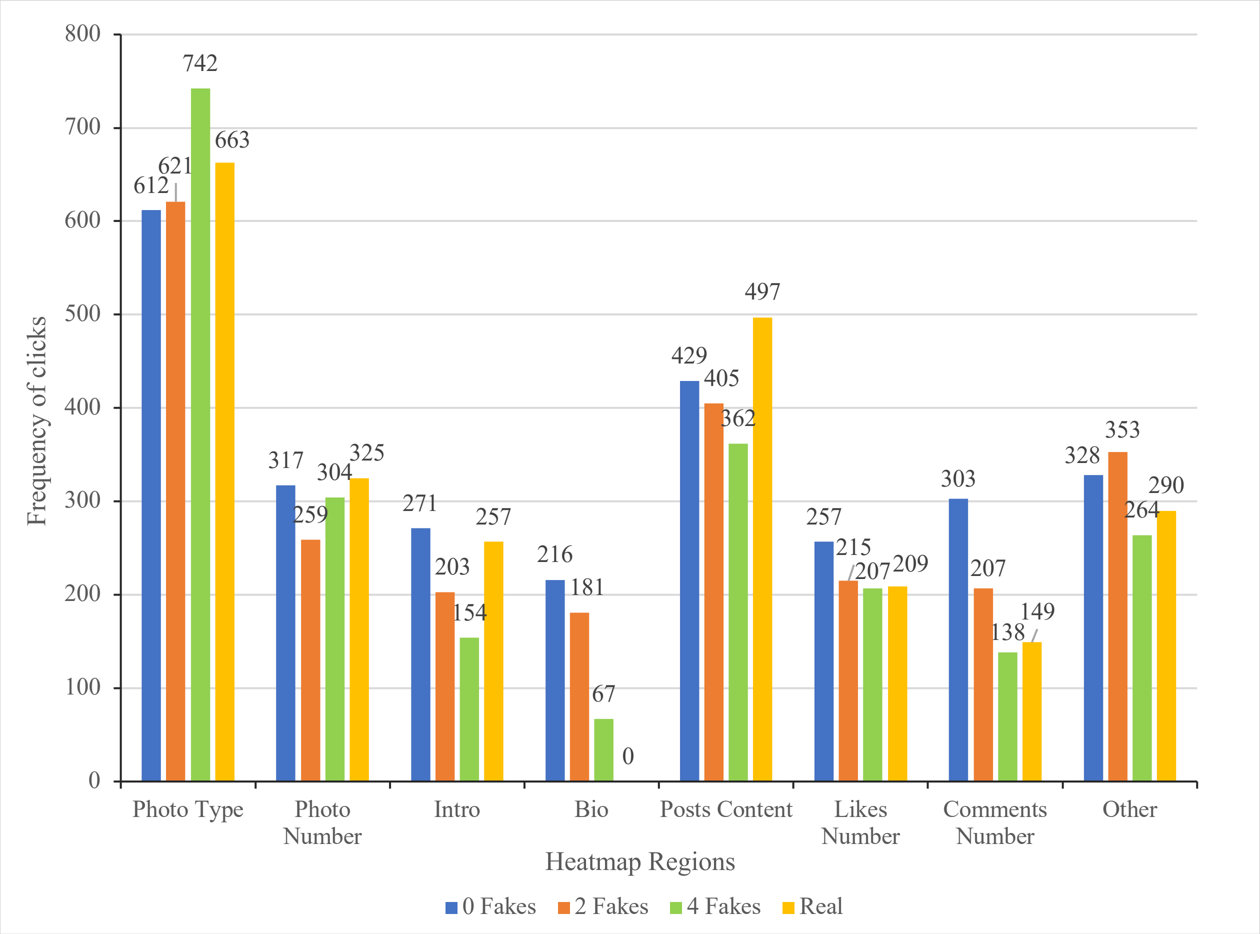
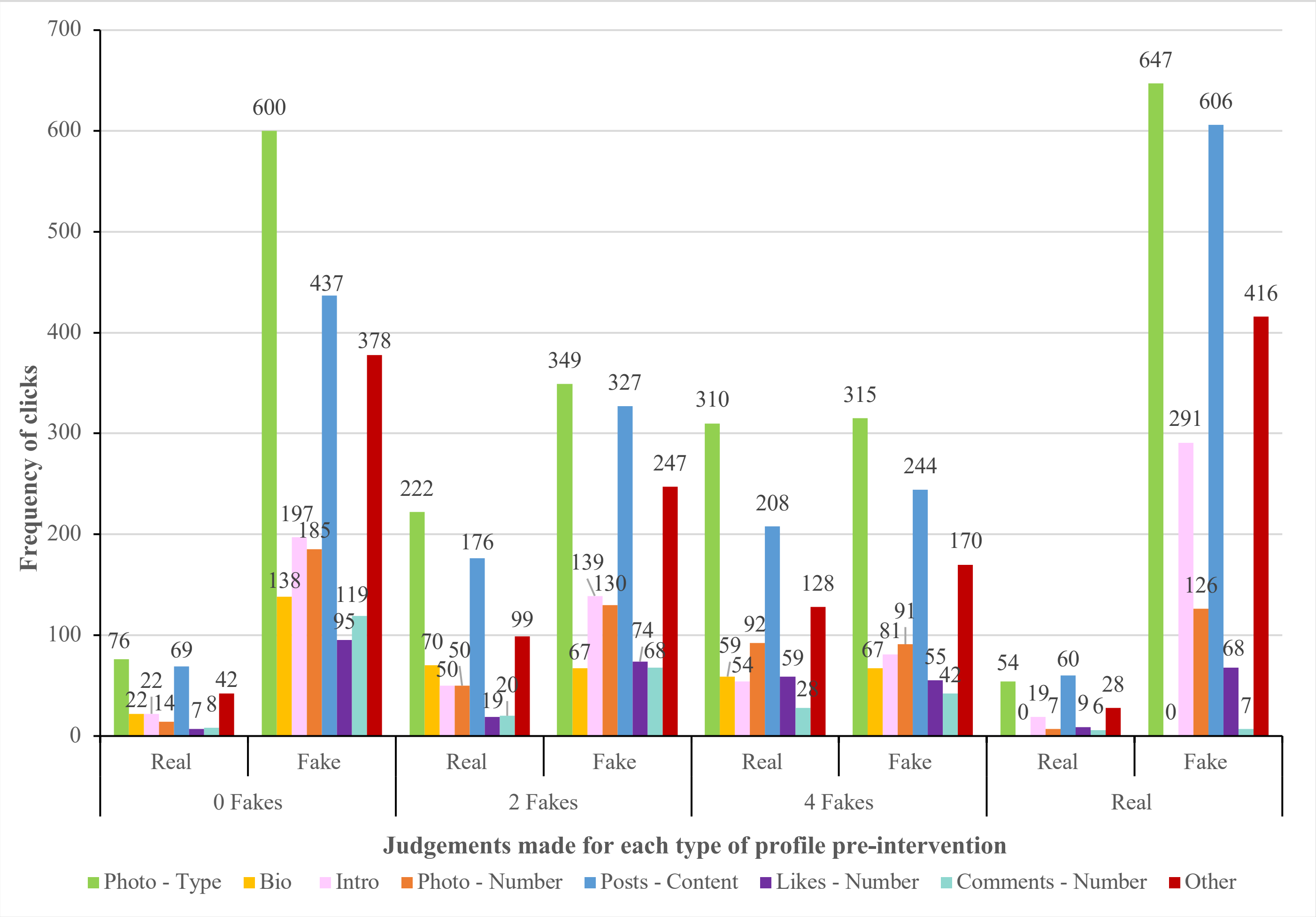


Figure 3 shows a similar pattern, in that participants post-intervention click most in the ‘Photo -Type’ and ‘Posts-Content’ regions when making their judgements of the profiles. Again, the Other clicks were quite frequent, however, post-intervention they were not as frequent as those pre-intervention. What is most evident here is that there is a higher frequency of clicks in general across all regions post-intervention when compared to pre-intervention, specifically a 12.62% increase in clicks.

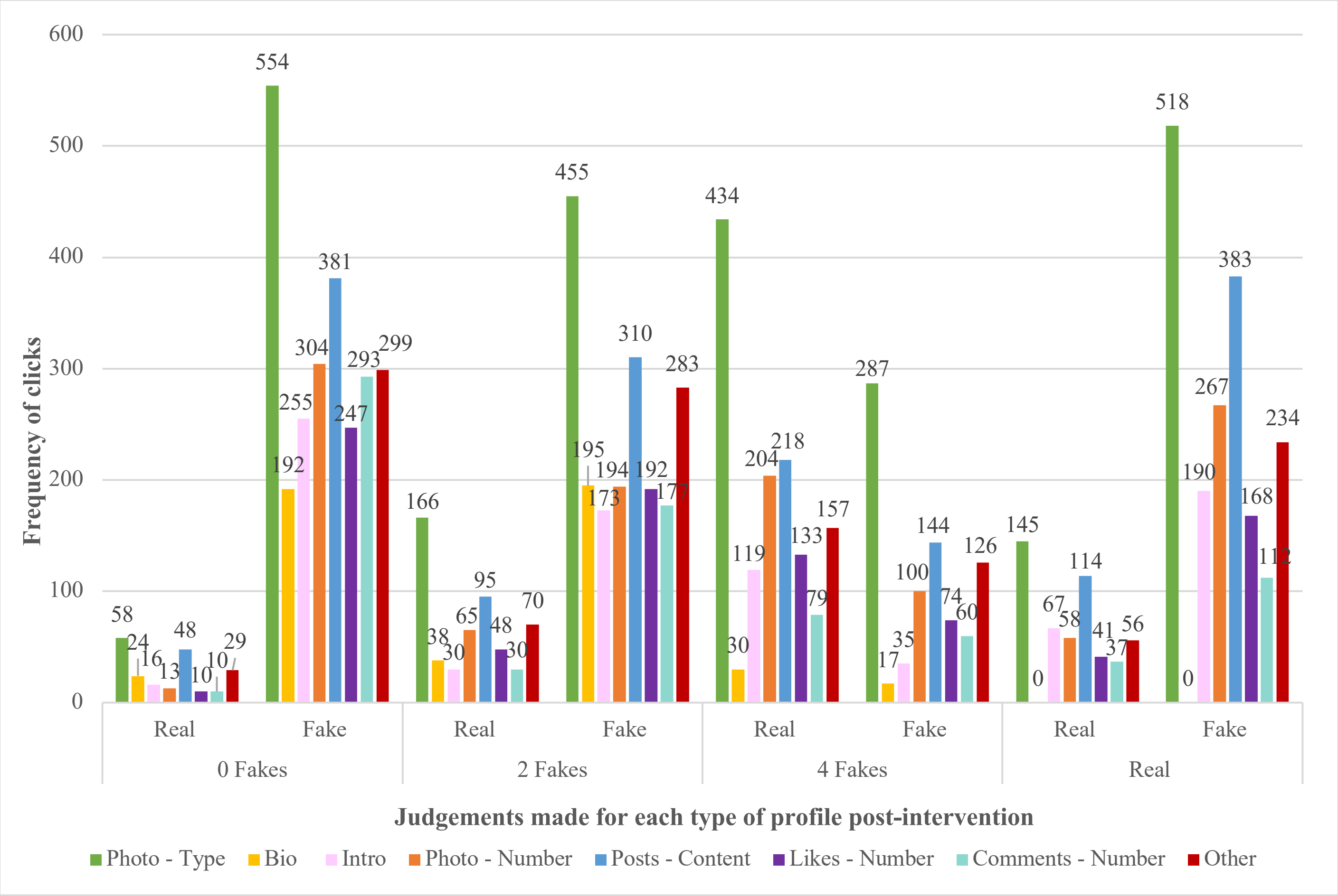
To further understand the areas participants used most when making their profile authenticity judgements additional analyses were carried out on the heatmap frequencies to examine if participants were clicking on different areas to inform real/fake judgements. The analyses were across all profile types both pre- and post-intervention. Figure 4 displays these results.

Figure 4.
Graph showing the frequency of clicks for each judgement type per manipulated characteristics for all profile types judged pre-intervention.



From figure 4 it can be said that on the whole pre-intervention, participants clicked more frequently on the profile when they were judging the profile as fake, a finding that is consistent across all profile types but most evident for *0 Fakes* profiles and *Real* profiles. However, a different pattern emerged post-intervention, as shown in figure 5.

Figure 5.
Graph showing the frequency of clicks for each judgement type per manipulated characteristics for all profile types judged post-intervention.

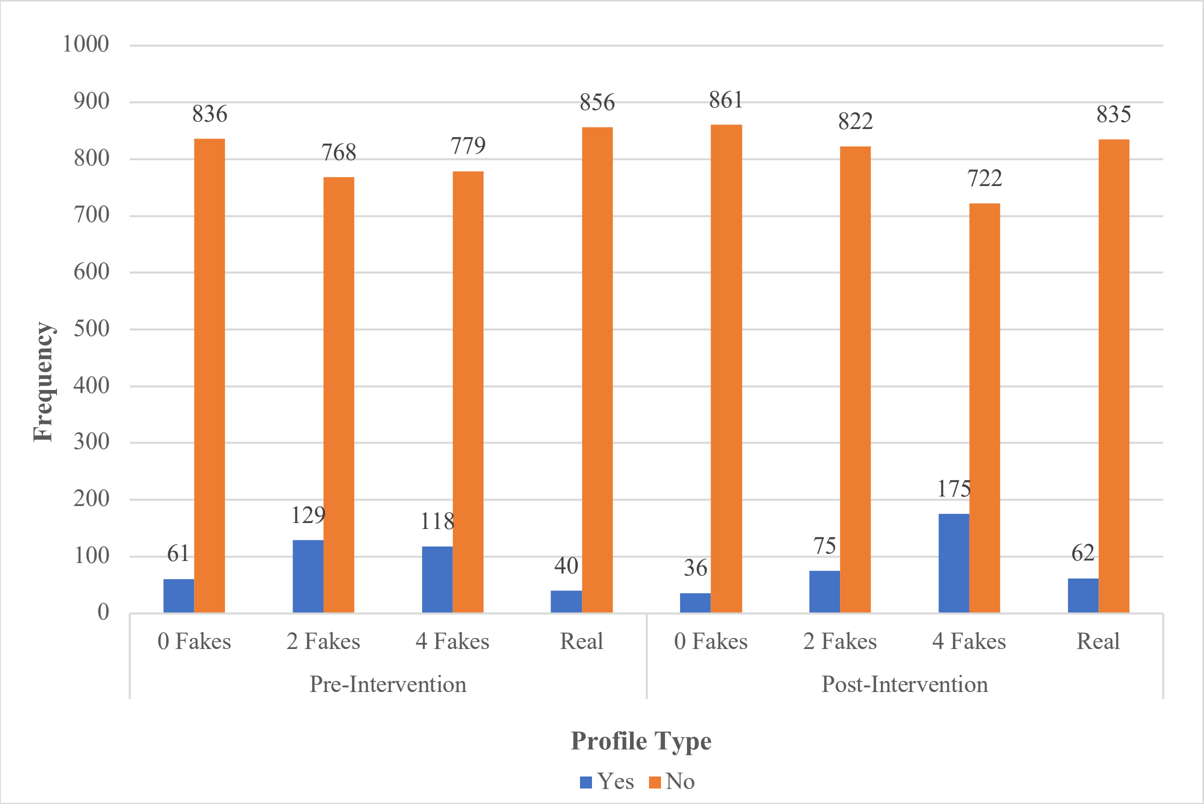


As can be seen in Figure 5, a different pattern in click frequency has emerged between pre- and post-intervention; post-intervention participants have a higher frequency of clicks when judging the profile as fake in relation to *0 Fakes, 2 Fakes*, and *Real* profiles, however this is the opposite for *4 Fakes* whereby their frequency of clicks is highest when judging the *4 Fakes* profiles as Real. However, consistently across all profile types both pre- and post-intervention, the characteristics with the highest frequency of clicks are *Photo Type, Posts Content,* and *Other* respectively, for both judgements of Real and Fake. Importantly, this suggests a reliance on these three characteristics whenever a profile is being judged, and it is the specific content of these characteristics that help participants make a decision and thus guide the direction of the judgement.

Deceptive Purposes of Profiles

After judging each profile, participants were asked to answer yes or no as to whether they believe the profile was created with deceptive intentions/malicious intent. Figure 6 shows that across all profile types, both pre- and post-intervention most participants perceived the profiles as non-deceptive.

Figure 6.
Graph showing the frequency of yes/no answers to the question 'Do you think this profile was created for deceptive purposes?' for all profile types both pre- and post-intervention.



Interestingly, pre-intervention participants reported that the profiles with 2 fake characteristics were the most deceptive, however post-intervention the profiles with 4 fakes were most deceptive. When asked why they think the profiles were deceptive, participants reported a plethora of reasons such as; "Catfish", "Trying to impersonate a celebrity", "I think that all people who make fake profiles have bad intentions", "There is no name and place", "To mislead people", "Only 3 photos, weird posts and no reaction to them", "To spread misinformation". However, there is a consistent trend here that overall participants did not believe the profiles, of any type, were made with deceptive intentions.

Discussion

The main purpose of this study was to investigate whether it is possible to train people to accurately identify fake Facebook profiles. Findings show that participants in the Training group had lower accuracy scores for three of the four types of profile post-intervention than those who did not receive the training; participants only performed better when judging profiles with 4 Fakes. Such result could be due to the fact that the example profile used to train participants included all possible manipulations of the fake characteristics, meaning effectively participants were shown a profile with *7 Fakes*. This raises the question of whether exposure to all of the possibilities regarding areas on the profile to be faked helped participants significantly when at least four of those were manipulated in the *4 Fakes* profiles. Considering participants in the training group post-intervention had the lowest mean accuracy scores of all participants for profiles with *0 Fakes* and *2 Fakes*, this suggests that more than two of the fake characteristics outlined in the training need to be manipulated on the profiles to increase the likelihood participants will recognise they have been manipulated, and thus judge the profile as fake. However, as previously mentioned in earlier studies, the *0 Fakes* profiles were created with the purpose of replicating a *Real* profile, thus directly manipulating all of the characteristics to be *Real*. As such, it would be expected that participants who received the training would achieve the lowest accuracy scores for *0 Fakes* profiles as they cannot identify any of the areas of the profile outlined within the training. This does not however provide an explanation as to why these same participants achieved the lowest accuracy score for profiles with *2 Fakes.*

Training overall did have an effect on participants' ability to accurately detect fake profiles, albeit a small effect (*d* = .06). This finding was also found in relation to deception detection in a meta-analysis of 30 studies; training improved overall ability to detect deception with a small-medium effect size (Hauch, Sporer, Michael, & Meissner, 2016). However, the pre-test post-test methodological design used in this study, albeit a popular and well used design, has been criticised by many, namely for the threat it generates to the internal validity of the study (Kim & Willson, 2010). The main argument behind this threat is that the use of a pre-intervention test can influence the intervention which in turn cause issues when measuring the effect of the intervention, an argument better known as *Pretest Sensitisation* coined by Campbell & Stanley (1963, p. 24). Researchers have explained such pretest effects as increasing arousal or attention to the post-intervention tests (Sime & Boyce, 1969), or direct participants' attention towards the specific aims of the research (Kim & Willson, 2010). This may be the case when participants are made aware of the pre-test post-test design, however within this study participants were not informed of this specific detail prior to commencement of the study, but rather that they will judge a number of profiles and were debriefed around the design and randomisation at the end of the study. Doing so, along with presenting participants with a different set of random profiles post-intervention, controlled for any such effect of pretest sensitisation, or other threats to internal validity.

In regard to the format of the training, the researcher used certain recognisable aspects of multimedia learning (Mayer, 2009) to choose a simplified but informative display to assist the participants. The training material designed used two screenshots of example fake and real profiles and was designed in such a way as to replicate the materials used within the study. The fake profile consisted of all possible fake characteristic manipulations outlined on one screenshot, as providing the three different types of fake profile could increase the probability that participants would be aware of the design of the study, meaning a host of biases could come into play, such as order effects, researcher effects, or pretest sensitivity as mentioned previously. However, several researchers have advocated for the use of video training for inoculation or pre-test post-test studies, due to improved performance when participants are trained using video when compared to others who were trained using static images including text or just text alone (Hanley, Herron, & Cole, 1995; Al-Seghayer, 2001). This being said, the researcher decided against using video instructions and used static screenshots instead. The purpose of such was to replicate the design of those used within the study and accurately represent the content of the profiles. Due to the method of which the profiles were created, showing a live Facebook feed would not have been possible, not only for ethical reasons but also the fact that a dynamic, moving profile does not present all areas of the profile that have been manipulated across one screen; to see each area, specifically the posts, would involve scrolling. Additionally, using screenshots avoided potential participant confusion around the profiles and thus negated any effects that could diminish the impact of the training.

Several researchers however have reported evidence as to the benefits of using words and images rather than videos for training purposes. A meta-analysis conducted by Hauch, Sporer, Michael, & Meissner (2016) has shown that using written instructions is the most effective method to train or inoculate participants, or written instructions alongside a video; video instructions alone are not effective at successfully inoculating participants. Additionally, Mayer (2009, p.3) reported in his multimedia theory of learning that an explanation is better understood when presented with pictures and words rather than words alone. Such combination positively impacts the cognitive processing of the training material by reducing the cognitive load placed upon the learner and making it easier for the learner to process the information (Mayer & Moreno, 2003; Brunken, Plass, & Leutner, 2004). Thus, the decision to use screenshots of the profiles that combined both images and words rather than an instructional video is warranted and appropriate for the overall design of this research. However, some researchers have reported that longer interventions are more effective, specifically those using verbal techniques of delivery. Hauch, Sporer, Michael, & Meissner, (2016) detailed in their meta-analysis investigating whether training improves deception detection, that training using verbal content cues had the largest training effect on accuracy of deception detection, recommending that any training interventions implemented in such research should focus on verbal content training.

Another area of note in regard to the training materials is that they required participants to focus on both a 'lie and a 'truth'. The fake profile trained participants to look for the fake characteristics, or 'lies', and the real profile trained participants to look for the real characteristics, or 'truths'. As a result of this, judgement accuracy may have been affected post-intervention as it has been found that when participants are trained to focus their attention on identifying truths their truth accuracy increased, and when trained in cues to deception no effect on their truth accuracy was found (Hauch, Sporer, Michael, & Meissner, 2016). This suggests that the type of training given can influence the accuracy of detection, and raises the question of whether it is easier to spot a truth than a lie? To investigate this further, future studies could employ a one-sided approach here by either only providing cues to deception or cues to truth across both types of profile, and then measure the accuracy of profile identification of either real profiles or fake profiles.

Participants were not given a maximum time limit for viewing the training, however they were restricted, unbeknownst to them, from moving on for 30 seconds by ensuring the arrow button did not appear until after 30 seconds had bypassed. Additionally, participants were not able to go back to the fake profile training screenshot once they had clicked off that page. This was not a conscious decision of the researcher, but rather a software driven restriction as Qualtrics did not allow backwards movements on just one question in the study. The addition of the backwards button at this stage would have introduced a backwards option for every question within the study which is not within this study design. Time spent on training was analysed and no effect was found between the time spent viewing the training materials and judgement accuracy for either real or fake profiles post-intervention, suggesting that length of exposure to the training material is not a factor in accuracy of judgements post-training. Future studies could introduce a time-limit here to measure any effect this may have on judgement accuracy.

With a focus on judgement accuracy overall, participants were more accurate at identifying and judging fake profiles than they were real profiles; a linear trend in judgement accuracy similar to that of previous studies within this research. When looking at specific mean accuracy scores per profile of participants in each intervention group, both pre- and post- intervention, participants were most accurate at judging real profiles, and less accurate at judging profiles with 4 fake characteristics, followed by those with 2 Fakes and finally least accurate at judging those with 0 Fakes.

In regard to the areas participants relied upon most to inform their judgements, *Profile Type and Posts Content* were relied upon most as demonstrated by the frequency of clicks on the heatmaps. The region of *Other* received the third highest frequency of clicks. These are the same areas as those in previous studies within this research, however within this study these areas were also found under different conditions, namely pre- and post-intervention. This suggests that the intervention did not have an impact on the areas participants rely upon to when making their judgement. However, the frequency of clicks post-intervention across all characteristics was higher than those pre-intervention. Again, as mentioned previously, this could perhaps be due to the intervention making participants aware of all the possible areas that could be manipulated, leading to more clicks overall on the profiles. This could also be testing effects in that the judgments of the first set of profiles pre-intervention, and therefore exposure to the profiles pre-intervention, influenced the outcome, or judgement accuracy, post-intervention. However, researchers have found that inoculation studies are likely to be influenced by item effects whereby any effects observed may be due to the items within the study (the profiles), rather than testing effects where exposure pre-intervention may have an influence on the outcome post-intervention (Roozenbeek, Maertens, McClanahan, & van der Linden, 2021). This research is specific to game-based inoculation studies on misinformation, however the theory behind this is still applicable to this study and may aid in explaining why the frequency of clicks post-intervention are significantly higher than those pre-intervention, although further exploration into this in future studies would further aid in explaining this result.

When looking at the frequency of heatmap clicks per type of judgement, it was evident that participants mostly had a higher frequency of clicks when judging all types of profiles as fake, with the exception of judgements of *4 Fakes* profiles post-intervention. Participants had a higher frequency of clicks for all characteristics when judging these profiles as Real. An interesting finding when coupled with the fact these are the only profiles that participants in the Training group post-intervention judged accurately more so than the same type of profile pre-intervention. This suggests that perhaps participants click more on profiles when unsure of their judgement, when accurately judging the profile participants clicked on all characteristics a lot less.

Several suggestions for future research and replications of this study have been discussed above, however one more significant point of note is in regard to possible future adaptions to the training intervention. It has been evidenced by many that task feedback, a method where participants are offered feedback on their performance, is effective in increasing participant accuracy of detecting deception (Kessler & Ashton, 1981; Kluger & DeNisi, 1996; Elaad, 2003). Implementation of this, prior to training, could assist in increasing accuracy of the training group further. Additionally, with the continued success of the Bad News game in inoculating participants against misinformation, it might be prudent to consider implementing some of the techniques used in those studies to test whether participants' judgement accuracy increases as a result. Doing so could assist in further understanding, and measuring more accurately, whether creating a fake profile helps in identifying fake profiles in the future. For example, participants could be required to create a series of fake profiles, of which they are then exposed to and asked to judge the authenticity of.

Although many suggestions of methodological alterations have been made for future researchers who may want to replicate this work, these do not undermine or distract from the findings of this study that the training delivered did have an effect on participants' judgement accuracy of fake Facebook profiles.

Chapter 8: Super Recognisers

Throughout this research, when considering participant's judgement accuracy, each study has identified any participants who achieved maximum judgement accuracy. This was investigated to understand whether any participants have an extremely high or innate ability to accurately identify fake profiles and thus could be considered as *Super Recognisers.*

The definition of the term *Super Recogniser (*SR) is currently lacking consensus amongst SR researchers (Ramon, 2021). Broadly speaking, persons considered SRs are those who have a significantly high ability to recognise faces - SRs sit on the extreme high point of the scale of facial recognition ability. Opposite to this are the individuals with a condition named developmental prosopagnosia, also known as 'face blindness', who sit at the extreme low points of the scale. People with this condition are exceptionally poor at recognising faces yet have a distinct lack of cognitive deficits or brain trauma that may explain why they have such a condition.

The term *Super Recogniser*, first coined in a pioneering paper by Russell, Duchaine, and Nakayama (2009), was based upon research on four individuals who had self-disclosed that they had a strong ability to recognise faces after seeing media coverage regarding the researchers' previous work on developmental prosopagnosics. These individuals reported that they could recognise faces they had seen in crowds, faces of strangers they had not seen for years, or even faces whose appearance had changed drastically, such as changing from a child to an adult, and as a result felt they had a natural ability to accurately recognise faces. To test their ability, each of the four individuals were asked to complete two different facial recognition tests. *The Before They Were Famous* (BTWF) Test was created specifically for use in their research, and simply involved showing participants images of famous people as a child, i.e., before they were famous. Doing so meant that a correct identification of the person required an ability to recognise faces that have changed distinctly over time. The second test, the Cambridge Face Memory Test (CFMT) (Duchaine & Nakayama, 2006a), involved participants learning a set of six different unknown male faces from a variety of different views and angles and under different conditions, including different lighting and images with added ‘noise’. Russell, Duchaine, and Nakayama (2009) extended this test further to the CMFT long form by adding an extra round whereby participants were also exposed to ‘distractor images’ that differed greatly to the original images shown. Across both tests, all four individuals performed at ceiling, gaining maximum scores, and outperformed the control subjects greatly. And so, from this, the term *Super Recogniser* was introduced.

Despite Russell et al. (2009) seminal paper being published 12 years ago, the research field of SRs is still a relatively new phenomenon, and interest is still growing (Ramon, 2021), not only in the research field but also in the real-world application domain. Due to the recency of this work into SRs, there is a distinct lack of empirical evidence regarding the different factors that underwrite the extreme abilities of SRs (Nador, Zoia, Pachai, & Ramon (2021). Additionally, several researchers have indicated that there are several limits to ‘Super Recognition’, particularly in that there may be several biases that SRs are subject to. For example, Bates et al., (2019) investigated the effects of the *own-race bias* or *other-ethnicity effect* on SRs, a well-known bias whereby it has been found that recognition memory is much higher when presented with faces from the recognisers’ own ethnicity than those from a different ethnicity. The researchers found that SRs are not immune to the effects of the other ethnicity effect and are in fact subject to this effect in both face matching and face memory tasks. Similarly, Rhodes and Anastasi (2012) found supporting evidence for another bias that effects SRs – the *own-age bias* – whereby recognition memory for faces of persons within one's own age group are more accurately recognised than faces from a different age group. Additionally, Bate, Bennetts, Murray and Portch (2020) found the own-age bias to be present among SRs when face matching children's and adults faces, with adult participant SRs performing better when matching adult faces. Studies of this nature raise the question of whether being a super-recogniser is as useful to public bodies as is suggested. For example, recent use of SRs has mainly been in police and criminal investigations, however the use and acceptance of this within these domains is questionable when the biases are taken into consideration – being susceptible to the *own-race* and *own-age* biases can become problematic particularly in reference to identifying an offender of a crime who could then be both charged and sentenced with said crime. However, as stated, the research on SRs is still new to the field, and so the research on the biases that SRs are susceptible to is therefore also new. Thus, drawing conclusions regarding the problems that relying on SRs can cause is perhaps slightly premature, yet is an issue that should be carried through in thought when researching SRs and the applicability of their skills in real-life sectors.

Even with the limited empirical evidence thus far, the phenomenon of SRs is one which is not only very interesting and ever expanding, but it has also garnered attention from practitioners outside of Neuropsychology, namely within police forces for use within criminal investigations (Dunn, Towler, Kemp & White, 2023). However, evidence for SRs being present within aspects of life different to that of recognising faces, is an area of SR research that has very limited literature. To aid in the expansion of the SR literature and offer a fresh approach the literature, the data captured within this thesis was further analysed to measure the possible presence of SRs of fake social media profiles. To identify possible SRs in the data, the judgement accuracy scores of each participant within each study were analysed. Each participant had a total accuracy score for all profiles judged, with a maximum score of 12, and an individual accuracy score for the judgements of real profiles and the judgements of fake profiles. As slightly different experimental designs were used in each study, the number of real and fake profiles (and type of profile for the fake profiles) judged by each participant differed across studies. The criteria for an SR within this body of work was an individual who scored either 100% with a maximum judgement accuracy score, or 91% with a judgement accuracy score of one less than the maximum.

Only two of the six studies identified SRs – Study 1 identified five and study 2 identified 1. The participants in each of these studies achieved a score of 11 (91% accuracy) – no participants achieved 100%. Methodologically, Study 1 implemented a within participants repeated measures design whereby participants viewed a total of 12 screenshots of Facebook profiles; six real profiles, and six fake profiles, of which three had '2 Fake characteristics' manipulated by the researcher, and three had '4 Fake characteristics'. All participants viewed the same six real profiles and a random selection of fake profiles. The five participants that were identified as SRs all had a total accuracy score of 11 out of 12, and all inaccurately judged a *different* fake profile, however all the fake profiles inaccurately judged were those with two fake characteristics. Focusing on the demographics, these five participants' ages ranged from 20 to 37 (M = 27), two identified as Male and three as Female, all identified as 'White (Includes British, English, Scottish, Welsh, Northern Irish, Irish, Irish Traveller or Gypsy and any other white backgrounds)', and all disclosed they live in European countries and have done so for their whole lives. Additionally, each participant completed two different personality questionnaires: Ten-Item Personality Inventory [TIPI] (Gosling, Rentfrow, & Swann, 2003), and the Social Sensitivity Scale [SS Scale]

(Riggio, 1986). Looking at their scores, each participant had different scores, and there were no identifying factors that were consistent across all five identified potential SRs that could be considered a 'trait of SRs'.

Study 2 differed methodologically from Study 1 in that it used a between participants design with three different conditions. Participants were randomly assigned to a condition whereby the saw six fake profiles with either *0 Fakes, 2 Fakes*, or *4 Fakes*, and all participants saw the same six real profiles. The only SR present in Study 2 also scored a total of 11 out of 12, however different to that of the five SRs from Study 1, the profile that was judged inaccurately had four fake characteristics. Though, this was the only possibility as the participant only viewed fake profiles with four fake characteristics. As there was only one participant in this study identified as an SR it seems that there may always be an element of chance when judging authenticity of online profiles.

Be it as it may, the data discussed is not significant to be of note in terms of identifying SRs which is evident from looking at the data distribution graphs of the total accuracy scores. Due to the low number of SRs, there is not a significant uptick in the normal distribution curve to categorically state that these studies have identified SRs, and their recognising ability is not due to chance. However, it is to be noted that these studies have highlighted that some people seem to have an advanced ability to accurately identify, or recognise, fake Facebook profiles from real Facebook profiles. Although, this may be due to several extraneous variables such as exposure to fake profiles or individual differences, there is a possibility it could be an innate ability – a relationship that should be of interest to investigate further in future.

The inclusion of SR analysis in future research on fake profiles is necessary to understand whether firstly the SR definition can include recognition of fake profiles, and secondly whether there are people who have a natural ability to recognise such profiles. If this were to be found, particularly if on a larger scale, the implications of this would be impactful in several ways. Identification of such SRs could not only contribute to helping the social media platforms fight the battle against fake profile by helping said platforms understand the areas of the profiles they use to recognise it to be fake, but also contribute further to this literature in terms of garnering further understanding of the specific areas of the profiles used when accurately identifying the profiles as fake .

Chapter 9: General Discussion

This main aim of this research was to study the judgement accuracy of fake social media profiles from a human perspective. As the human element of fake profile detection has been largely overlooked so far, the intention was to not only investigate how humans make authenticity judgements and whether they were accurate doing so, but also to begin laying the foundations for future research to build upon. Technological solutions have been well studied, and attested to in this research, however as deceiver and detectors lean ever more heavily into technological advancements, it is difficult to see how a solution can lie in the technological field, where the inevitable arms race leads to parity at best. This collection of studies can be considered as a foundation stone upon which further research can be built to inform the fields of psychology, particularly judgement/decision-making and deception detection, and computer science, to truly understand human judgement of deception online.

Overall, this research fits into a general pattern of previous work on deception generally which places human deception detection accuracy as frequently no better than chance. No participant managed to accurately identify all the fake profiles in any of the individual studies, however a consistent linear pattern in judgement accuracy was found across all studies within this research: participants were more accurate at correctly identifying real profiles as real than they were at correctly identifying fake profiles as fake. Between the types of fake profiles, a consistent linear trend was also found - participants were most accurate at judging fake profiles with four manipulated characteristics (*4 Fakes)*, and least accurate at judging profiles with zero manipulated characteristics (*0 Fakes).* This linear trend remained true under each experimental manipulation, whether that be participant conditions (Study 2, Chapter 3), time pressures (Study 4, Chapter 5), cross-cultures (Study 5, Chapter 6), or training interventions (Study 6, Chapter 7), demonstrating the robustness of this accuracy trend.

Participants were shown to have a bias towards judging the profiles as fake. Each study employed the use of *Signal Detection Theory* (SDT) (Green & Swets, 1966) to further understand participant judgements, particularly how well participants can make decisions under uncertainty. With the exception of Study 1 (Chapter 2), participants were able to distinguish between the signals (fake profiles) and the noise (real profiles). All studies, including Study 1, showed that participants overall had a bias towards judging the profiles as fake but participants in Study 1 were unable to distinguish between the signals and the noise, meaning they were unable to distinguish the fake profiles from the real profiles. As suggested in Study 1, this might be a result of participant misunderstanding, task confusion, or other extraneous variables. An important factor to consider here is Study 1 did not contain profiles with *0 Fakes*. This implies that the inclusion of the *0 Fakes* in all other studies might have enabled participants to differentiate more effectively between profiles with *2 Fakes* and *4 Fakes,* thus improving their ability to distinguish between the signals (fake profiles) and noise (real profiles) in subsequent studies.

Overall, the presence of such bias towards judging the profiles as fake is a surprising finding considering it has been shown numerous times that humans have a bias, or naturally default, to the truth. Levine (2014, 2020) developed the *Truth-Default Theory* (TDT) which posits that humans tend to operate on a default presumption that another person is honest (2020, p.94), a presumption that makes sense considering the majority of human communication is honest. However, having such a default means that humans are vulnerable to deceit. Levine (2020) further outlines that the *truth-default* is a failure to consider the possibility of deceit and is used as a fall-back cognitive state after failing to retrieve sufficient evidence to affirm the presence of deception (p. 94). The theory overall is comprised of multiple modules, or mini theories, which can be understood both as a standalone theory or as part of the TDT as a whole (p. 96). The modules of relevance to this study are *The Veracity Effect* and *Sender Honest Demeanour.*

The Veracity Effect refers to the honesty (veracity) of the communication and how such honesty predicts whether the message communicated will be judged correctly, stating that honest messages produce higher accuracy than lies (p.203). In fact, in a summary of his collective works on veracity and deception detection, Levine (2020, p.202) reported that truth accuracy is always more than 50%, usually significantly higher than 50%, whereas lie accuracy is always below 50%. Additionally, the stronger the *truth-bias/default,* the stronger the *veracity effect* – the *truth-bias* actually causes the *veracity effect*.

Sender Honest Demeanour refers to the believability of a non-verbal communication signal or cue, independent of actual honesty of the cue and has been found to explain 98% of the variance in deception detection accuracy (Levine, Serota, Shulman, Clare, Park, Shaw, Shim & Lee, 2011). Effectively, the way in which the sender of a message presents themselves (their demeanour) can influence the way in which the message is received and consequently judged. This could be applied to this study in that the way in which a profile is presented, through such content as the photos or posts, and how it can influence the judgements of the profile, i.e., if the profile, or profile user, appears honest then the profile is more likely to be judged as real, even when it is fake.

In regard to this study, these modules of TDT can provide somewhat of a theoretical explanation as to why participants judge real profiles more accurately than they do fake profiles – they have a *truth-bias* and believe the profiles to be *Real* based on the *sender demeanour cues.* However, as mentioned earlier, the overall *Signal Detection* Theory results are the opposite to that of TDT – participants are biased to judge profiles as fake. Thus, TDT does not help to explain why participants showed such bias. However, recent works in the social media domain can provide some evidence in support of this finding.

TDT has recently been applied to deception in an online context. Researchers Luo, Hancock and Markowitz (2020) relied upon the theoretical framework of TDT, specifically the *veracity effect* and *sender demeanour cues,* to investigate whether human participants could detect fake and real news on social media. Their findings were inconsistent with the expectations of TDT. In their first study, participants had a *deception-bias* in that they were inclined to judge the fake news as fake, and such bias led to more accurate fake judgements than real judgements. However, the researchers concluded this might be due to general levels of suspicions regarding news presented on social media, and conducted a further study to consider other variables that may have an effect on a person's *truth-bias.* Said further study investigated whether a high number of Facebook likes on fake news articles increased the credibility of the message. Findings showed that a high number of Facebook likes increased credibility of the articles which reduced the earlier found *deception-bias*, and also increased the detection of real news articles. The results of these two studies are supportive in that they apply TDT to online deception detection, specifically the detection of fake and real stimuli (news articles) and provide evidence to support the participant bias to judge profiles as fake found in this research. The nod to people largely believing news on social media to be inaccurate suggests that perhaps the same beliefs might be held in regard to social media profiles, however based on TDT this is likely to not be the case as people are more likely to be biased to believe them to be true or real. In fact, this is evidenced in the findings of this research whereby participants, by an overwhelming majority, consistently reported that the profiles were not made with deceptive or malicious intent. This suggests that in regard to social media, perhaps human judgement of the authenticity of profiles will always be flawed with the prevalence of such biases.

What might seem like two contradictory findings presented above – participants are more accurate at judging real profiles correctly and also have a bias towards judging profiles as fake – can be reconciled with the notion that a judgement of real might not necessarily be a positive affirmation of the veracity of the profile, but rather a default position taken when there are no discernible cues to the profile being fake. As participants are primed to identify fake cues, by the nature of the study and the specific instructions given to participants regarding judgements of the profiles, they are specifically looking for fakery, and lack of such fakery results in a default real judgement. In line with this supposition, across all studies profiles containing zero fake manipulated characteristics (*0 Fakes)* have been shown to be the hardest to accurately detect as fake, particularly in Study 3b where more time was required when judging *0 Fakes* profiles. As these profiles were purposefully created to be the most realistic of the fake profiles, such results suggest that participants found it harder to identify any obvious signs of fakery, and as such defaulted to judging the profile as real. Thus, the inclusion of *0 Fakes* in each study may have had an influence over the SDT scores and biases.

However, to reliably conclude this, future researchers will need to study this in much more detail with a more direct focus on the *veracity effect* and *sender honest demeanour cues* to identify whether this still holds true, and perhaps further refinement to the instructions given to participants may minimise the presence of the bias towards judging profiles as fake.

Another consistent finding across all studies in this research is in relation to the manipulated profile characteristics, specifically the type of photograph used in the profile or cover picture (*Photo Type)*. The *Photo Type* characteristic was a significant predictor of participants' judgements (whether they judge the profile as real or fake) and the accuracy of their judgements; specifically when *Photo Type* had been manipulated on the profile participants were more likely to judge that profile as fake, and the judgement of fake was more likely to be accurate. In fact, with the exception of Study 1 (Chapter 2), *Photo Type* was the only manipulated characteristic that was a significant predictor of both type of judgement and judgement accuracy in all other studies. In Study 2 (Chapter 3) and the Non-Indian participant condition in Study 4 (Chapter 5) *Photo Type* was the *only* significant predictor, meaning no effects were found for any of the six remaining manipulated characteristics.

The strong effects of *Photo Type* on participants' judgements and judgement accuracy were reflected in the heatmap click data. Across all studies and all experimental conditions, *Photo Type* was the region of the profile that consistently had the highest frequency of clicks to denote that participants used this area of the profile to inform their judgement. The manipulated characteristics used in each study (post Study 1) were different to those in Study 1, yet the same effect was found. This demonstrates the strength of these findings – even when manipulated under different conditions, with different characteristics (e.g., *Age, Number of Friends etc.)*, *Photo Type* was still clicked on the most.

There are several theoretical frameworks within the literature that could explain why such an effect was found, of which some have been touched upon throughout the course of this thesis. Firstly, in relation to the position of the photo on the social media platform itself, researchers have found through use of eye-tracking methodology, that internet users tend to follow set patterns when looking at a webpage. As first introduced in relation to this research in Study 2 (Chapter 3), researchers have found across a multitude of studies that users tend to look at a webpage in an *F-shaped*, *L-shaped,* or *Z-shaped* pattern (Nielson, 2006; Scott & Hand, 2016), meaning anything in the top left-hand corner tends to be the first thing users look at. Any one of these patterns, in regard to the Facebook profiles, would capture the profile photo as one of the first things looked at when judging the profile, as the profile picture is situated in the top left-hand corner. Facebook did alter their layout in the time between Study 1 and the commencement of Study 2, whereby the profile picture was moved into the middle of the top banner of the profile. In this context that would mean that the *L-shaped* pattern would become redundant in terms of it capturing the profile picture as the first thing they see. However, the characteristic of *Photo Type* did not relate only to the *profile picture* on the profile, but also included the *cover photo*. The *cover photo* is an image that runs the whole width of the top of the profile upon which the *profile picture* sits. This means that any effect of *Photo Type* is not confounded by any layout changes as the background upon which the *profile picture* sits is encapsulated under the same manipulated characteristic. With that being said, it can be suggested that *Profile Type* is the area of the profile relied upon most and has a strong influence over a persons' judgement of the profile, as it is located in an area of the profile that is often looked at first.

Secondly, in regard to the decision-making and judgement processes, an explanation could be found from the thin-slicing literature, as mentioned previously in Study 4 (Chapter 5) whereby the term *thin slice* is used to describe short snippets of social behaviour (of less than 5 minutes) that perceivers draw inferences from (Carney, Colvin, & Hall, 2007). It has been shown that accurate judgements can be made in under 10s (Murphy et al., 2003), and even under 100ms (Willis & Todorov, 2006).

In reference to images and social media, thin slicing has been found to yield accurate judgements of personality from condensed profiles, or rather profiles with limited information (Stecher & Counts, 2008), and personality judgement accuracy for Facebook profiles was highest when judgements were based solely on the profile picture (Ivcevic & Ambady, 2012). These findings suggest that perhaps the photos, either the *profile picture* or *cover picture*, are all participants need to form a judgement or that strong evidence is required from other aspects of the profile to overturn the initial conclusion provided by the *cover picture* or *profile picture.*

In parallel with this, research on the cognitive processes involved in decision making have shown that the brain typically makes rapid, or even automatic, decisions. The brain is said to utilise two systems to both form impressions and make decisions: System 1 and System 2 (Kahneman, 2011). In summary, System 1 is fast and operates automatically with very little effort, and System 2 is slower, more effortful and considered as the lazy system (Kahneman, 2011, pg. 64), which we are predisposed to avoid using (Dennis & Minas, 2018). System 1 is found to be the system that influence human perceptions and judgements (Dennis & Minas, 2018), and when processing information using System 1 only the information immediately available is used (de Castro Bellini-Leite, 2013), meaning that the saliency of available information drives the decision-making process.

The saliency of photographs has been investigated by several researchers, with a recent focus on the ever-growing issue with AI manipulated photographs and whether manipulated or altered photos can be accurately identified. Findings have shown that people have a limited ability to accurately detect manipulated photographs of real-world scenes (Nightingale, Wade & Watson, 2017), and are only slightly better than chance at correctly determining the authenticity of a manipulated photograph (Nightingale, Wade, Farid & Watson, 2019). With this, and the literature regarding use of System 1 decision making processes, it is of concern that people cannot accurately identify manipulated, or fake, photographs when they are most likely to use the most salient information available to inform their decision-making – photographs or visual imagery.

In specific relation to this study, participants relied heavily on *Photo Type* when making a judgement as to the authenticity of a Facebook profile. Such reliance suggests that participants are predominantly relying on System 1 process to form their decision which prioritises immediate and salient cues. Specifically, participants gravitate toward the most salient available information on the profile – the photographs. However, it is essential to note that the images themselves were not directly manipulated. Instead, they featured a celebrity or a piece of artwork. Despite this, participants still considered such images as potentially fake when the images did not corroborate or align with the remaining context of the profile. With the research mentioned above, and the findings of the research, it is of importance to further understand such reliance on *Photo Type* and whether direct manipulation of the images shown on the profile has an effect on judgement accuracy of said profile.

In essence, *Photo Type* has been a consistent thread throughout this research and has had the strongest effects on participant accuracy across all study manipulations, an effect that could be as a result of any or a combination of all of the explanations presented above. As such, further research is needed to further understand specifically why this is the case and investigate whether it is the location of the image on the profile that elicits such results, or whether it is the specific content of the images that influence judgements. Additionally, understanding the interplay between System 1 decision making process, saliency, and image context can provide valuable insights into how people evaluate authenticity in the online domain.

Consistently throughout each study, there were no significant effects found between the personality variables (TIPI scores and Social Sensitivity scores) and judgement accuracy. It is widely known within psychological literature that discovering any effects of personality variables are notoriously difficult, and any effects that are found are typically small, as reported in a recent meta-analysis of personality research (Bühler, 2023). However, given that personality is dynamic and can influence a multitude of experimental outcomes, it is imperative to continue to research the constructs of personality.

The validity and methods of the psychometric tests used to test for personality have been widely critiqued. For example, Kumar et al. (2023) report that psychometric tests are criticised mainly due to the need for participants to have the ability to introspect and have a good understanding of their true character to accurately answer the tests. Similarly, there are several biases at play when measuring personality, such as social-desirability bias whereby participants answer in a way that would make them look ‘better’ to the researcher, or subject bias whereby participants or subjects act or respond in a way in which they think is expected of them, thus not capturing the true personality of the participants. As Cipresso and Riva (2016, Chapter 18, p.240) note, these biases are not easy to control for due to the fact that personality, by its nature, is such a personal thing. It is recommended that in any future replications or expansion upon this research that personality variables still remain, however perhaps an alteration to the measurement, or a stronger focus on such variable, is needed to elicit any effects.

Throughout each study a series of mixed results were found in relation to the social media variables – *time spent on social media per day* and *previous experience creating a fake profile.* In some studies, there were no effects found of either of these variables on participant judgement accuracy. In those where an effect was found, the effects were minimal, i.e., some regression models reported one significant coefficient, but the overall model was non-significant. Whilst these results could be considered underwhelming, some researchers have found significant results in relation to social media experience and sensitivity to fake profiles created by bots, whereby participants who reported greater experience in social media were less sensitive to being duped by a bot (Kenny et al., 2022).

Thus, whilst inferences in regard to social media experience (time spent per day and fake profile creation) and judgement accuracy of fake profiles cannot be made from this research, the inclusion of such variables in future research is warranted based on the significant findings of others presented above. Such findings suggest there is a relationship between social media and fake profiles, and to elicit such findings methodological changes to the studies in this research are needed.

Limitations

Whilst this research has shown consistent and robust findings in relation to judgement accuracy of fake social media profiles, there are a few limitations to the research that shall be considered below.

One of the main limitations of this research is in relation to participants' judgements of the profiles and the characteristics of the profiles they used when making said judgements. As discussed, *Photo-Type* was relied upon most when making judgements and is the only manipulated characteristic that had a significant effect on participants' judgements and judgement accuracy. However, it is not known if participants relied upon *Photo-Type only* when making their decision, or whether they used other available information on the profile, alongside the photo, when making their judgements, i.e., did participants look solely at the photo to make their decision, or did they look at the remaining areas of the profile and then conclude that the type of photo doesn't match the other available information, thus denoting the profile as fake? The researcher tried to control for this by the inclusion of heatmaps on each of the profiles, which allowed participants to click the areas they used when making their judgements. However, the heatmap functionality itself does not come without methodological concerns.

The heatmap software extension in Qualtrics has several limitations of which may have had an adverse effect on the results in this research, many of which have been mentioned throughout this thesis. Firstly, when outlining the aspects of the profile to be measured on the heatmap, the regions were rigid in that they could only be in the shape of a square or a rectangle. This was not necessarily detrimental in regard to the majority of the profile, however the *profile picture* on the profile is round in shape, so the region had to be made wider than the circular shape to ensure the whole image was encapsulated. Again, this shouldn't have caused any adverse effects as the profile picture sits atop the *cover photo* which was also outlined as the region *Photo Type,* meaning any inaccurate clicks due to the oversized region surrounding the profile picture would still be encapsulated under *Photo Type*. This also raises an issue in regard to *profile picture* and *cover picture* both being recorded under *Photo Type*, as it is not known which of the two received the most clicks without manual checking of every click by every participant in all studies – a task that is incredibly unfeasible. Ideally it would be prudent in future to split these two regions apart and have them named as either *profile photo* or *cover photo.* However, these regions would not be able to be accurately outlined due to the limited capabilities of the Qualtrics heatmap software. Perhaps replication using a different heatmap software would be beneficial.

Further, the region with the third highest frequency of clicks across all studies was *Other.* This was not a specified region by the researcher, but rather a way in which Qualtrics classifies any clicks that fall outside of the prescribed regions. The edges of each region seemed to be quite sensitive to clicks, as any clicks that touched the edge of the region border were automatically seen as inaccurate and classed as *Other*. The researcher became aware of this in the early stages of this research, and edited the instructions given to participants accordingly to try to counteract this effect. From Study 2 onwards, participants were instructed to be as accurate as possible when clicking and to try and click directly on the stimuli they used. However, this did not seem to have the acquired effect as *Other* was consistently the region with the third highest number of clicks. What this means in context is that the heatmap data is not reporting an accurate a picture as possible of the areas of the profile's participants used when making their judgements – further clicks for the manipulated characteristics could be captured incorrectly under the umbrella of *Other,* thus negatively skewing the reported numbers. In future, perhaps the instructions given to participants need to be tailored further to try to counteract as many inaccurate clicks as possible.

Additionally, Qualtrics only allows a maximum of ten clicks per heatmap, meaning participants' ability to click on any of the areas they used to make their judgement is limited. If they exceed the prescribed amount, the first click made will now change and become their tenth click. Participants were informed of this; however, it is of note to remember when considering the results of the heatmap clicks. Additionally, the order in which participants click on the heatmap are unknown, meaning we cannot infer the areas they clicked on first to make their judgements, but rather the total frequency of clicks in relation to each manipulated characteristic.

Despite the limitations of the heatmaps, the results garnered provide the literature with a dynamic insight into areas of social media profiles used to inform judgements of authenticity – something of which has yet to have been shown elsewhere. Additionally, they show a consistent pattern across all studies – participants rely most on the photos (*Photo Type)* and the posts (*Posts Content)* when making a decision as to the authenticity of the profiles.

A limitation which runs across all of the studies is the creation of the fake profiles themselves. Whilst each fake profile created was fake in that it was created by the researcher and didn't represent a real person or entity, it was created for a research study, they were not genuinely deceptive profiles taken from Facebook trying to fool the audience for malicious purposes. Given the wide range of reasons for which a deceptive profile may be set up, it is worth considering that the fake profiles created by the researcher would not necessarily reflect all these design types; a profile designed to elicit money from a social media user may well differ from a profile looking to infiltrate a network through being accepted as a friend by multiple users. These differences could not be accounted for without either accessing truly fake profiles from Facebook (which brings its own problems of establishing what is truly fake) or by focusing solely on a particular kind of fake profile – e.g. a trolling profile – which would entail narrowing the research to too great an extent, therefore reducing its applicability.

Furthermore, users who took part in the tests throughout the study were clearly aware that they were involved in research on fake profiles. Whilst some provided feedback to express an interest in, and enjoyment of, the study, this foreknowledge would have given the suspicion of a fake profile a more prominent place in a participants mind than would ordinarily be the case if they were to be using a social media platform recreationally in their own time. Without the researcher being able to know what characteristics a normal user may choose to use when making an accuracy judgement of a fake profile in their normal lives, it is not possible to say with complete certainty that those characteristics chosen by the researcher were the only, or most suitable ones to employ during the study. Whilst the heatmap analysis and changes made after Study 1 took into account these factors in order to mitigate against them, it is still a possibility that there are some other characteristic users rely upon which was not included in this study and which further research may uncover.

Implications

A clear theme running throughout each set of results in this study is the need for further research. The large disparity in research geared towards software and algorithm-based solutions to fake profiles and online deception generally as opposed to human based work clearly highlights the need for this study to have been done, but also for future studies to pursue a similar direction. If we still want social media to be designed and built for humans to use, then ultimately human judgement will always have to be a factor in determining how effective and safe that use is. So far, there has not been enough recognition of this.

For the researcher of this study, it would be of great interest to see this work taken forward for a further study combining all of the aspects of this thesis into one new experiment. This could potentially involve further work along the same lines whereby participants are provided with a training intervention before any exposure to fake profiles and then asked to make fast, time pressured judgements on cross-cultural profiles. This would be an interesting approach to see how the different experimental methods interact with each other. The researcher fully endorses this avenue of exploration to others wishing to use it.

It is too easy to assume that technological problems must beget technological solutions. A truly proficient algorithm for one social media site may not be the most effective on another, which will then create the need for each site to run separate algorithm systems when a human judgement perspective could provide cover across the whole social media landscape. As a rule, technological solutions will also quickly be outdated and also counter-attacked by deceptive user's own updated software. Facebook now, as of late 2023, allows users to create multiple profiles linked to a single account. As with most technological updates, this provides opportunity for improvement and susceptibility to deceptive practices. Software based algorithms which might have looked at multiple profile-based accounts as a cue to fakery would now be instantly rendered obsolete. Given the pace of change in the online world, there is no guarantee that any company will survive and thrive years into the future, regardless of how accurate their detection algorithms may be. Human based judgement, which will outlive the online sites themselves, therefore seems a preferable solution. In addition, a hybrid model of software and human judgement, which could weed out the majority of easy to spot fakes before they reached the stage of human interaction with them, may seem of great benefit and would mitigate the problem. This could also leave the same end result in terms of fake profiles that are difficult to spot (as *0 Fakes* were in this study) being present on a platform that people believe is then safer to use.

What this study does contribute is the basis for which a framework can be built that outlines of the areas of a Facebook profile that constitute a fake profile. With some confidence, the areas added to this framework are the photos displayed on the profile (*Photo Type)* and the posts (*Posts Content),* as this research has consistently shown that people look at these two areas when judging the authenticity of a social media profile. Further development of this framework with subsequent further research can help to inform human users of social media of the *red-flags* to look for when assessing the validity of a Facebook profile – a tool which could prove to be rather useful in combatting the fake profile epidemic.

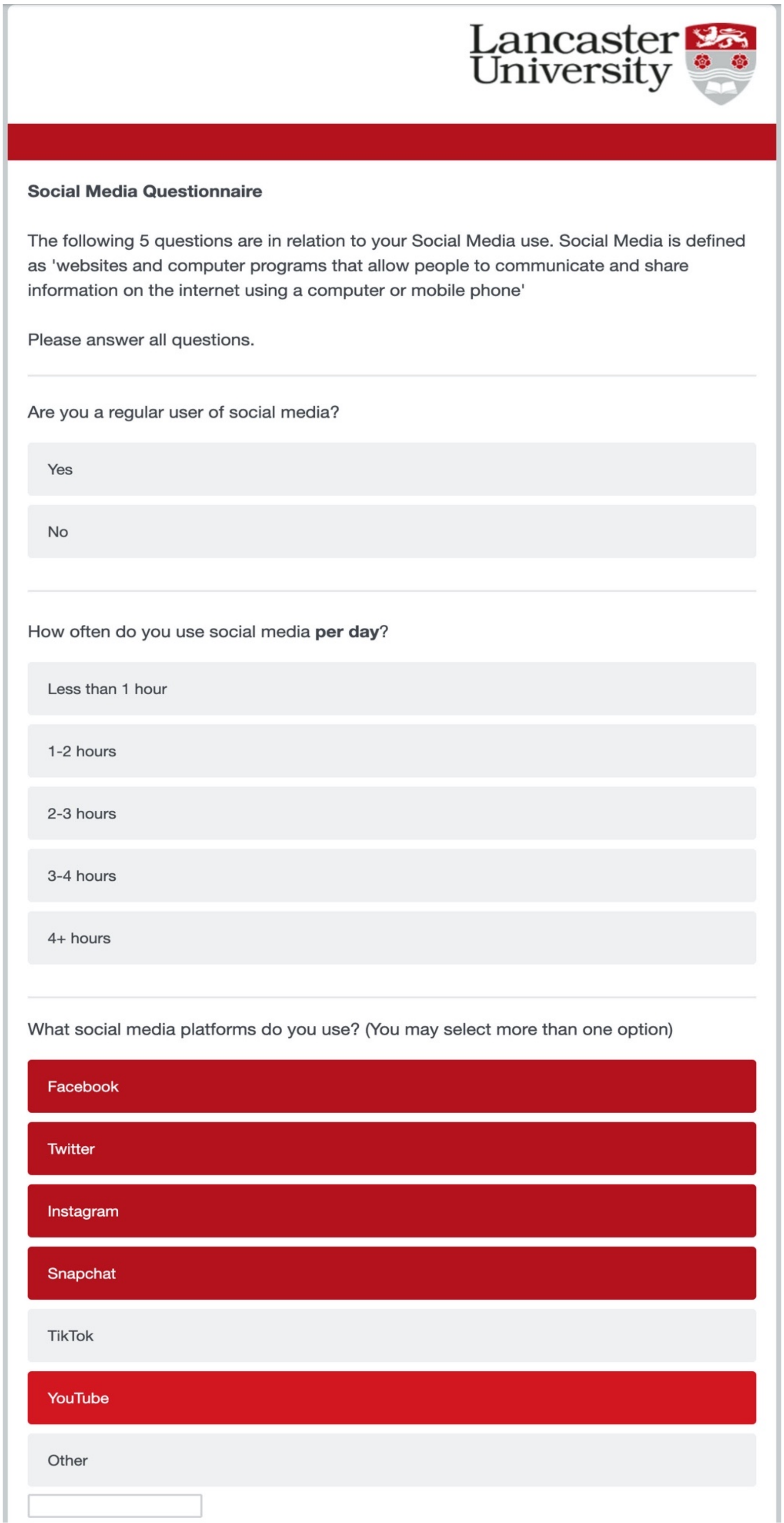
Conclusion

This research began with an attempt to help fill the gap that exists in the study of human judgement accuracy of fake social media profiles. It is hoped that this will eventually allow for individuals to make better and safer decisions in their online use and reduce the impact of those who generate profiles for their own malicious practices. Ultimately, this may be an aspiration which is impractical. In many respects the online world and social media are both analogous and opposing to our evolutionarily defined traits. Throughout human history we have had to rely on face-to-face interactions, with their abundance of visual and non-visual stimuli to help us make a judgement on someone's personality and motives. Now the online world forces us to make judgements of people when we lack all of these. We have no idea if the photo we see is the true face of the person we're interacting with. It is static and devoid of visual clues. We rely on text which may have been edited many times before sending or even generated by AI driven bots. We are almost driven to judgement in a way which is wholly unnatural. When people can still struggle to judge people and their motives in face-to-face interactions after thousands of years of practice, is it feasible at best to expect them to be able to perform more effectively on a brand-new medium.

In 1818 Mary Shelley's work *Frankenstein* could pose the question "when falsehood can look so like the truth, who can assure themselves of certain happiness?" (p.96). In the digital realm of the 21st century our online falsehoods mean we're still asking ourselves the same question over two hundred years later. It may be unclear if humans will ever possess the judgement accuracy needed to truly tell the difference between falsehood and truth in the form of real and fake profiles, but if this research has anything to do with it, it will not be for want of trying.

Appendices

Appendix A – Social Media Questionnaire given to participants (Qualtrics)



Lancaster University

Social Media Questionnaire

The following 5 questions are in relation to your Social Media use. Social Media is defined as 'websites and computer programs that allow people to communicate and share information on the internet using a computer or mobile phone'

Please answer all questions.

Are you a regular user of social media?

Yes

No

How often do you use social media **per day**?

Less than 1 hour

1-2 hours

2-3 hours

3-4 hours

4+ hours

What social media platforms do you use? (You may select more than one option)

[✓ Facebook]

[✓ Twitter]

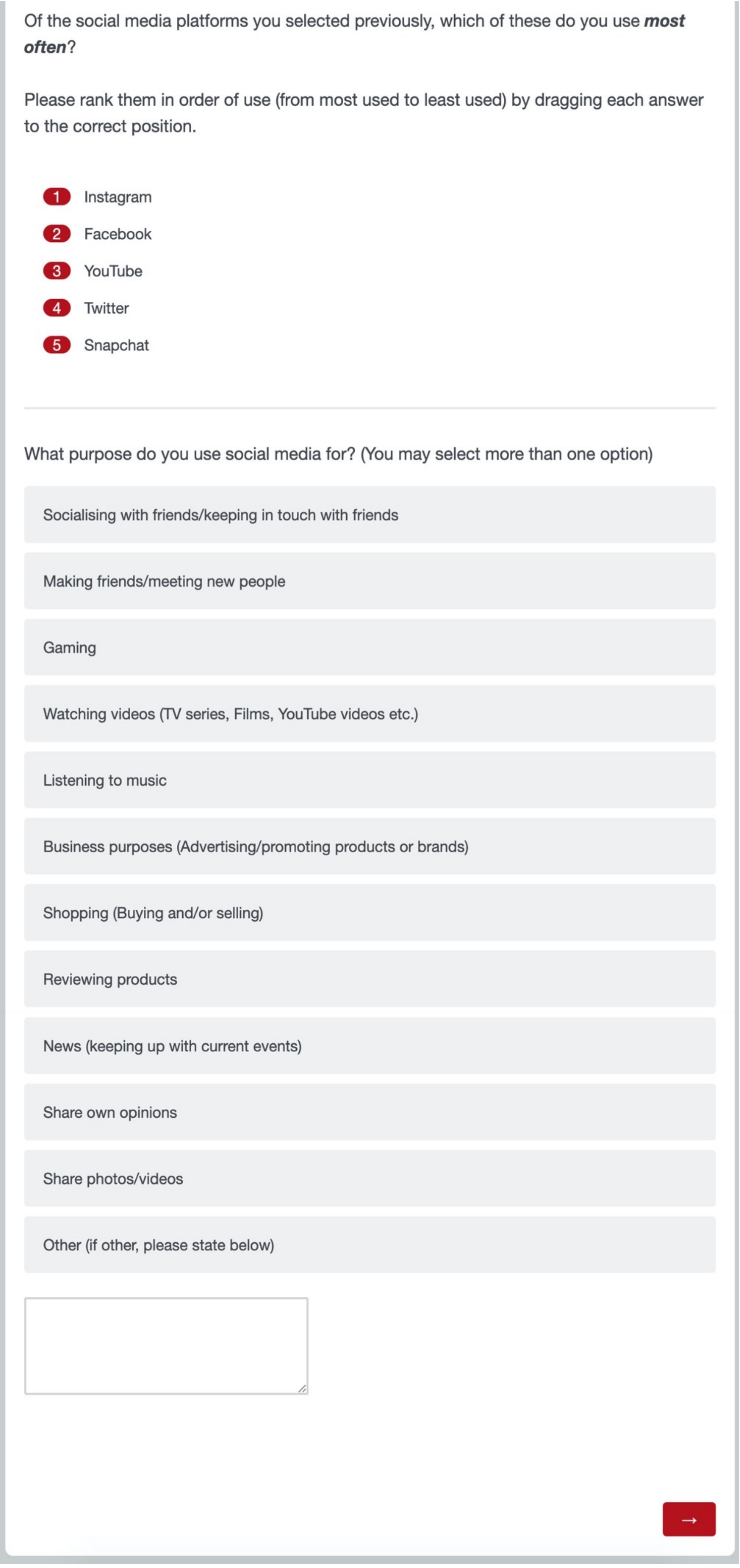
[✓ Instagram]

[✓ Snapchat]

TikTok

[✓ YouTube]

Other



Of the social media platforms you selected previously, which of these do you use ***most often***?

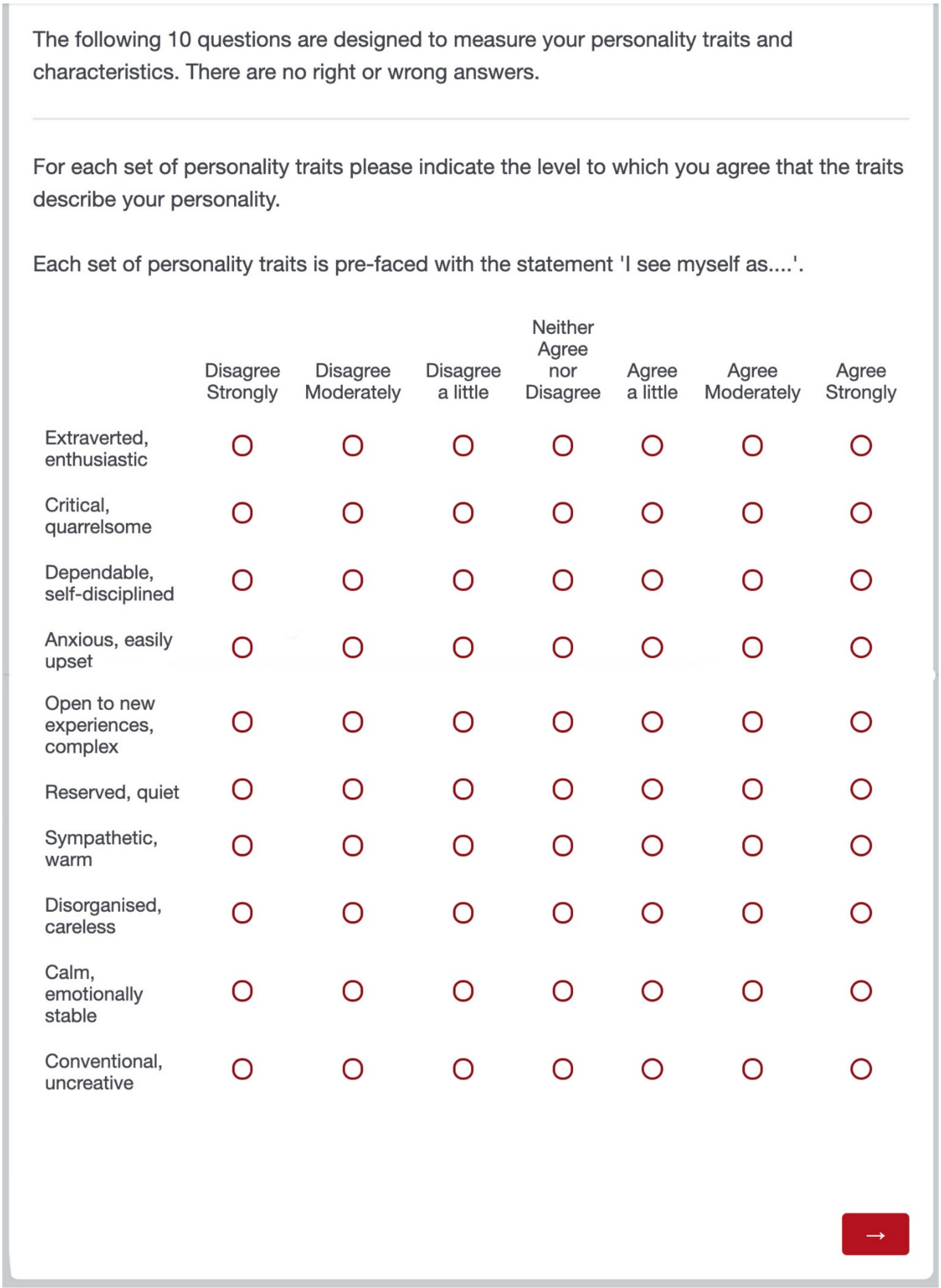
Please rank them in order of use (from most used to least used) by dragging each answer to the correct position.

1. Instagram
2. Facebook
3. YouTube
4. Twitter
5. Snapchat

What purpose do you use social media for? (You may select more than one option)

- Socialising with friends/keeping in touch with friends
- Making friends/meeting new people
- Gaming
- Watching videos (TV series, Films, YouTube videos etc.)
- Listening to music
- Business purposes (Advertising/promoting products or brands)
- Shopping (Buying and/or selling)
- Reviewing products
- News (keeping up with current events)
- Share own opinions
- Share photos/videos
- Other (if other, please state below)

Appendix B – Personality questionnaire - TIPI (Qualtrics)



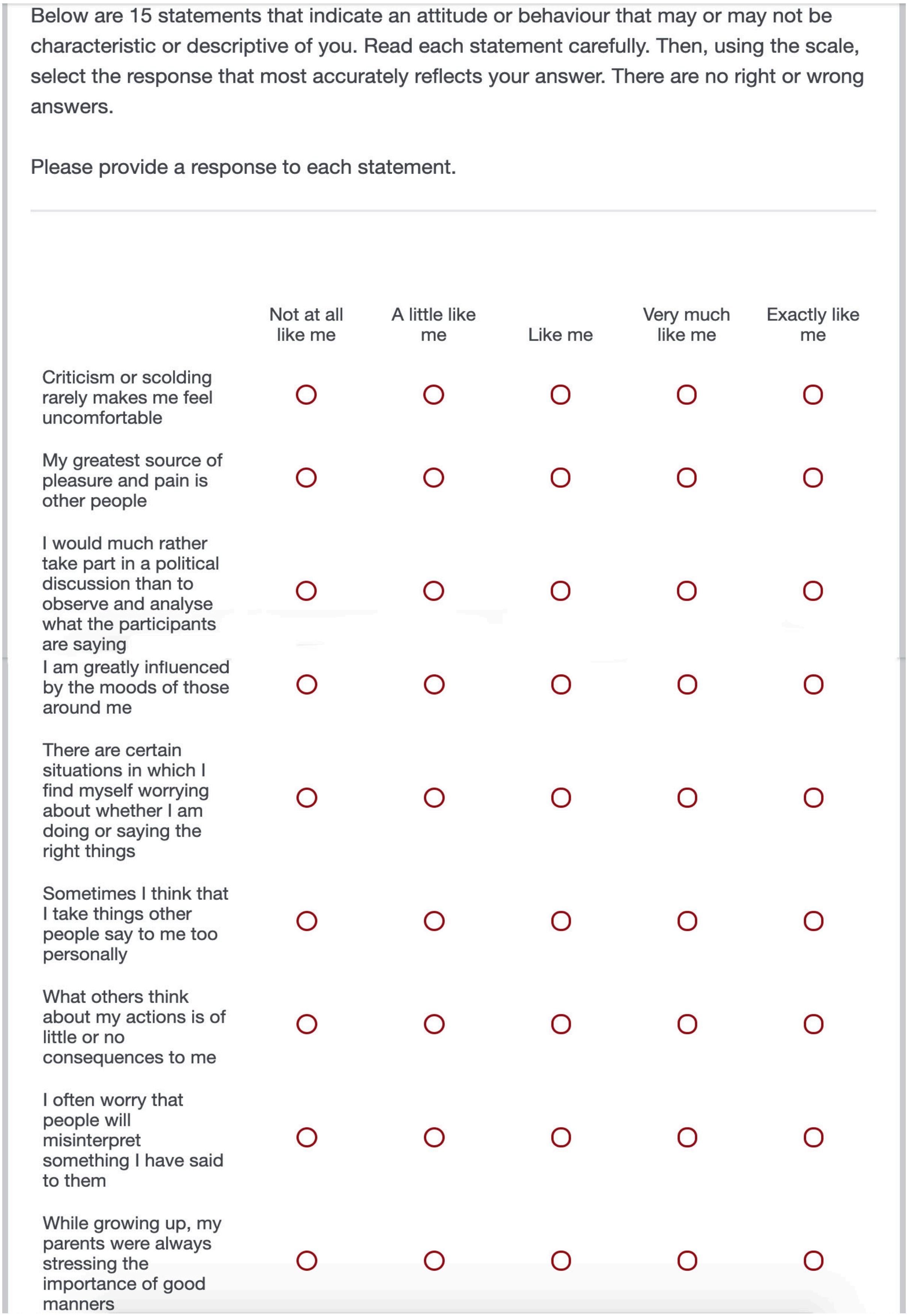
The following 10 questions are designed to measure your personality traits and characteristics. There are no right or wrong answers.

For each set of personality traits please indicate the level to which you agree that the traits describe your personality.

Each set of personality traits is pre-faced with the statement 'I see myself as....'.

	Disagree Strongly	Disagree Moderately	Disagree a little	Neither Agree nor Disagree	Agree a little	Agree Moderately	Agree Strongly
Extraverted, enthusiastic	○	○	○	○	○	○	○
Critical, quarrelsome	○	○	○	○	○	○	○
Dependable, self-disciplined	○	○	○	○	○	○	○
Anxious, easily upset	○	○	○	○	○	○	○
Open to new experiences, complex	○	○	○	○	○	○	○
Reserved, quiet	○	○	○	○	○	○	○
Sympathetic, warm	○	○	○	○	○	○	○
Disorganised, careless	○	○	○	○	○	○	○
Calm, emotionally stable	○	○	○	○	○	○	○
Conventional, uncreative	○	○	○	○	○	○	○

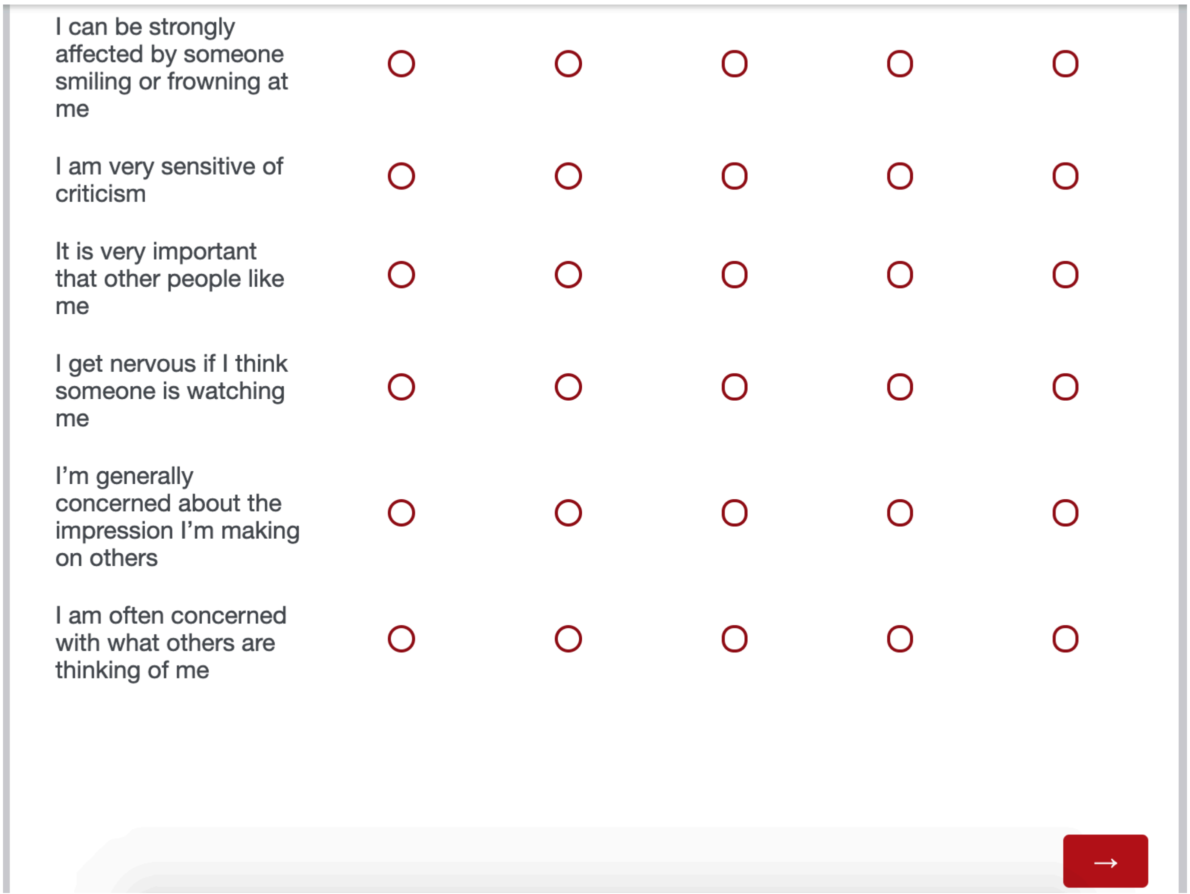
Appendix C – Personality Questionnaire – Social Sensitivity Scale (Qualtrics)



Below are 15 statements that indicate an attitude or behaviour that may or may not be characteristic or descriptive of you. Read each statement carefully. Then, using the scale, select the response that most accurately reflects your answer. There are no right or wrong answers.

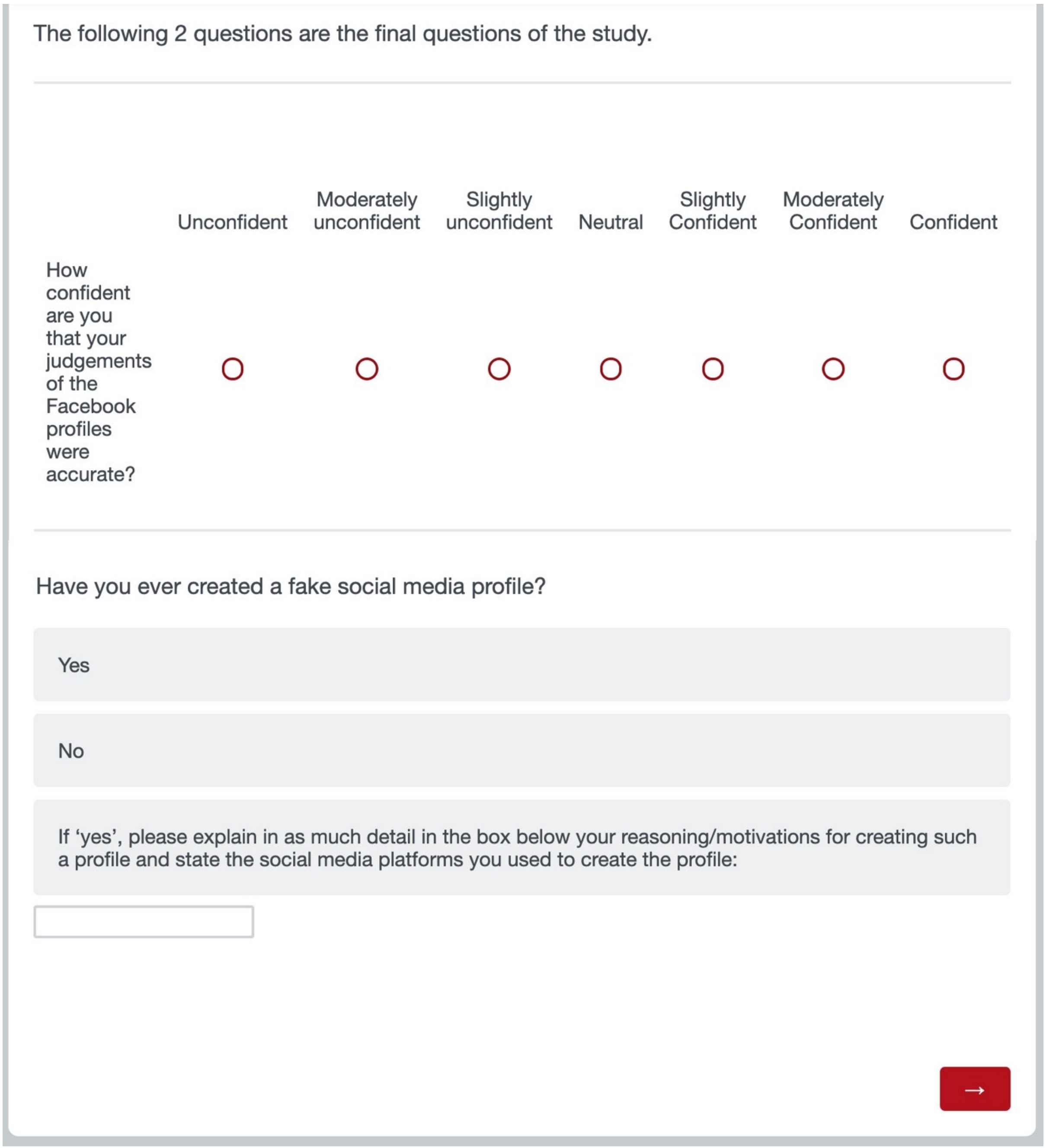
Please provide a response to each statement.

	Not at all like me	A little like me	Like me	Very much like me	Exactly like me
Criticism or scolding rarely makes me feel uncomfortable	○	○	○	○	○
My greatest source of pleasure and pain is other people	○	○	○	○	○
I would much rather take part in a political discussion than to observe and analyse what the participants are saying	○	○	○	○	○
I am greatly influenced by the moods of those around me	○	○	○	○	○
There are certain situations in which I find myself worrying about whether I am doing or saying the right things	○	○	○	○	○
Sometimes I think that I take things other people say to me too personally	○	○	○	○	○
What others think about my actions is of little or no consequences to me	○	○	○	○	○
I often worry that people will misinterpret something I have said to them	○	○	○	○	○
While growing up, my parents were always stressing the importance of good manners	○	○	○	○	○



I can be strongly affected by someone smiling or frowning at me	○	○	○	○	○
I am very sensitive of criticism	○	○	○	○	○
It is very important that other people like me	○	○	○	○	○
I get nervous if I think someone is watching me	○	○	○	○	○
I'm generally concerned about the impression I'm making on others	○	○	○	○	○
I am often concerned with what others are thinking of me	○	○	○	○	○

Appendix D – Follow-up Questionnaire given to participants at the end of the study (Qualtrics)



The following 2 questions are the final questions of the study.

	Unconfident	Moderately unconfident	Slightly unconfident	Neutral	Slightly Confident	Moderately Confident	Confident
How confident are you that your judgements of the Facebook profiles were accurate?	○	○	○	○	○	○	○

Have you ever created a fake social media profile?

Yes

No

If 'yes', please explain in as much detail in the box below your reasoning/motivations for creating such a profile and state the social media platforms you used to create the profile:

Appendix E – Framework for manipulated characteristics in each profile in STUDY 1. Characteristics highlighted in red denote fake and green denote real.

Profiles with 2 fake characteristics:

	CHARACTERISTICS						
PROFILE NAME (Will be omitted on profile)	**AGE**	**PHOTOS** (Number)	**PHOTOS** (Type)	**FRIENDS** (Number)	**POSTS** (Number & regularity)	**GROUPS** (Number)	**SOCIAL MEDIA** (Links)
LORENZO MARTINEZ	46	6 Non-Regular	Selfies Groups Pets	191	102 Regular	4	Instagram Twitter
ALICE ROUX	39	11 Regular	Landscape Celebrities	197	87 Regular	3	Twitter
IAN MARTIN	52	162 Regular	Selfies Groups	17	361 Regular	2	Instagram Twitter
JAMES BENRIDGE	35	72 Regular	Selfies Groups	125	3 Non-regular	2	Twitter
SASKIA SWANSON	89	39 Regular	Groups	117	68 Regular	28	None
FRED DIMBLE	18	3 Non-regular	Celebrities	103	81 Regular	4	None
SARAH LOCELSO	22	6 Non-regular	Selfies Groups	36	56 Regular	3	Snapchat
MARTINE SHELDON	27	11 Non-Regular	Selfies	136	21 Non-Regular	2	Twitter

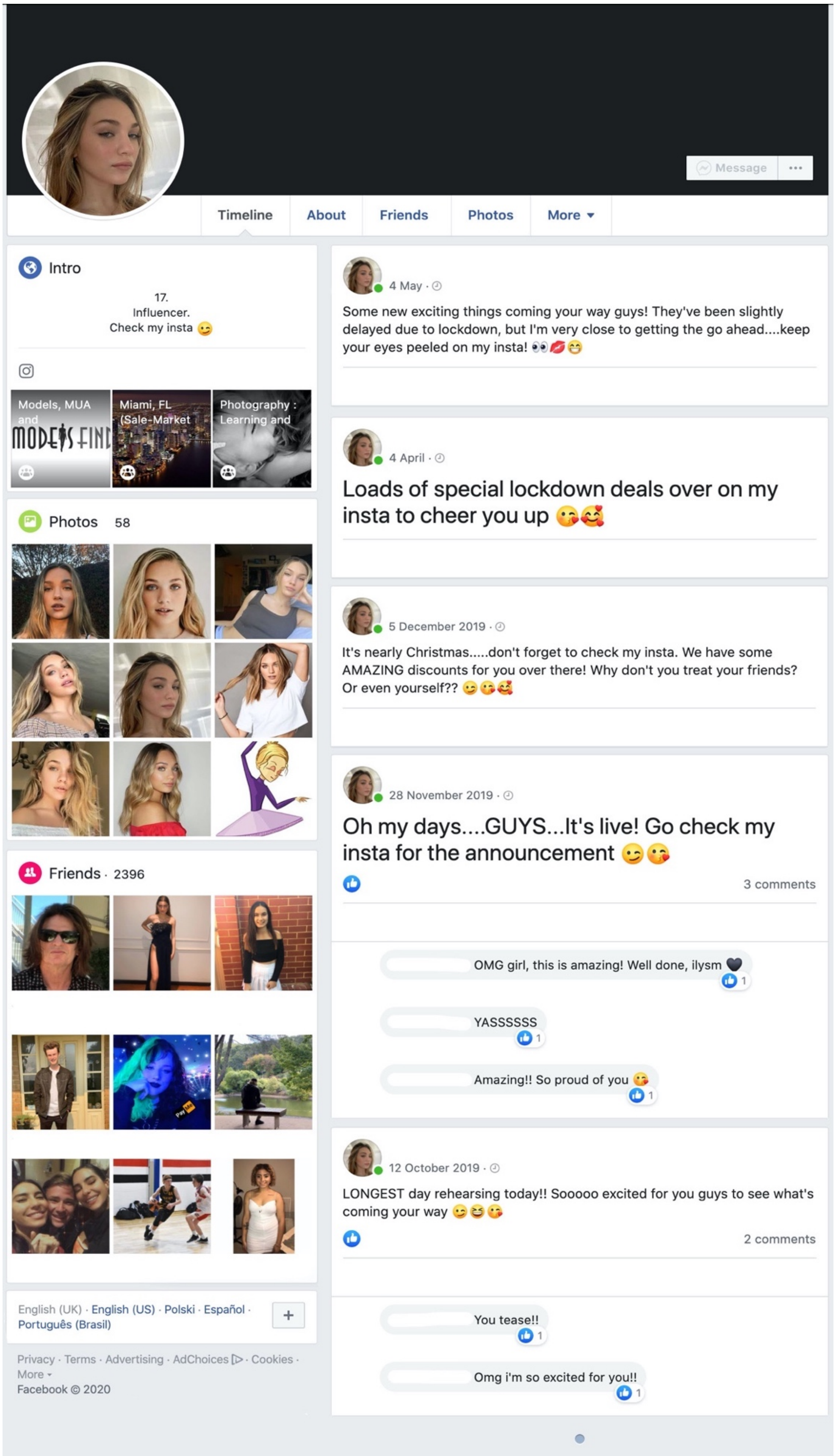
TOM WILSON	19	17 Non-Regular	Selfies	179	63 Regular	21	Instagram
SUNITA RANSPUT	26	237 Regular	Celebrities Landscapes Cartoon characters	61	374 Regular	2	Instagram Twitter
MARCUS IRONTORE	69	73 Regular	Landscapes	111	2 Non-Regular	4	None
SHEILA WHEELWRIGHT	75	51 Regular	Landscapes Artwork	107	64 Regular	0	Instagram
JOE COURTNAY	66	92 Regular	Selfies Groups Pets	942	9 Non-regular	1	None
HAYLEY POPOU	19	416 Regular	Selfies Groups	947	Hundreds Regular	26	Instagram Snapchat YouTube
ESME DAVIES	23	297 Regular	Selfies Groups	139	6 Non-regular	57	

Profiles with 4 fake characteristics:

	CHARACTERISTICS						
PROFILE NAME (Will be omitted on profile)	**AGE**	**PHOTOS** (Number)	**PHOTOS** (Type)	**FRIENDS** (Number)	**POSTS** (Number & regularity)	**GROUPS** (Number)	**SOCIAL MEDIA** (Links)
RANSIL SINGH	14	7 Non-regular	Cartoon characters Celebrities	1,362	96 Regular	4	Instagram Twitter Snapchat
PENNY TAYLOR	85	2 Non-regular	Landscape	135	4 Non-regular	2	Instagram
JENNIFER PARTON	16	12 Non-regular	Cartoon characters Celebrities	128	236 Regular	47	Instagram Twitter YouTube
MOHAMMED KHAN	87	3 Non-regular	Selfies Groups	3	6 Non-Regular	4	None
PATRICIA MACKENDALL	88	9 Non-regular	Groups	7	51 Regular	67	Twitter
DAVID GORN	91	4 Non-regular	Groups	107	5 Non-Regular	51	None
ABIGAIL DAWN	17	58 Regular	Cartoon characters Celebrities	2,396	9 Non-Regular	3	Instagram Twitter
STUART BRIGHTON	82	39 Regular	Artwork	15	64 Regular	0	None

PAULINE SMETHWICK	85	71 Regular	Landscapes Artwork	168	3 Non-Regular	29	Twitter
TERRY SMELTON	15	347 Regular	Selfies	874	17 Non-Regular	72	Instagram Snapchat
PAM JUMPERS	19	4 Non-regular	Cartoon characters Celebrities	1,553	3 Non-Regular	3	Twitter Instagram
KASPER IVANOFF	71	14 Non-regular	Landscapes Artwork	13	54 Regular	84	None
MARY ARTFUL	28	8 Non-regular	Celebrities Landscapes	184	12 Non-regular	0	Twitter
ERIC JAMESON	72	1 Non-regular	Selfies	8	2 Non-regular	0	None
MATTHEW TYME	73	82 Regular	Landscapes	1,582	8 Non-regular	50	None

Appendix F – Example of one of the fake profile screenshots as shown to participants in STUDY 1


Message
Timeline
About
Friends
Photos
More
Intro
Models, MUA and
MODELS FIND
Miami, FL (Sale-Market
Photography : Learning and
Photos 58
Friends · 2396
English (UK) · English (US) · Polski · Español · Português (Brasil)
Privacy · Terms · Advertising · AdChoices · Cookies · More
Facebook © 2020

17.
Influencer.
Check my insta 😉

4 May

Some new exciting things coming your way guys! They've been slightly delayed due to lockdown, but I'm very close to getting the go ahead....keep your eyes peeled on my insta! 👀💋😁

4 April

Loads of special lockdown deals over on my insta to cheer you up 😘🥰

5 December 2019

It's nearly Christmas.....don't forget to check my insta. We have some AMAZING discounts for you over there! Why don't you treat your friends? Or even yourself?? 😉😘🥰

28 November 2019

Oh my days....GUYS...It's live! Go check my insta for the announcement 😉😘

3 comments

OMG girl, this is amazing! Well done, ilysm 🖤 1

YASSSSSS 1

Amazing!! So proud of you 😘 1

12 October 2019

LONGEST day rehearsing today!! Sooooo excited for you guys to see what's coming your way 😉😆😘

2 comments

You tease!! 1

Omg i'm so excited for you!! 1

Appendix G – Consent form for participants providing their real Facebook profile

Dear Sir/Madam,

I am a PhD Psychology student at Lancaster University, conducting research on online deception. The focus of my thesis is on the use of fake identities online and the creation and use of fake social media profiles. The current study involves participants looking at a mixture of real and fake Facebook profiles and making a judgement as to whether they are real or fake. To allow for this research to take place, I would like to request permission to use your Facebook profile within this research as one of the real profiles.

To protect your privacy as much as possible, the use of your Facebook profile will be in the form of a screenshot. Participants will not have free access to your profile and as such cannot click on any photos, posts, friends or links that may appear on your profile. Additionally, your name will be removed from all areas of the screenshot of your profile to further protect your privacy and anonymity.

You are not obliged to take part in this research and you should not feel any pressure to do so, participation is completely voluntary. The screenshots of your profile will be kept confidential and stored on university approved encrypted equipment and destroyed after the appropriate amount of time (in accordance with GDPR & UK Data Protection Act).

You are free to withdraw from the research at any time after you have provided your initial consent on this form, and you are not obliged to give your reasons. After initial consent has been given, you will be sent a copy of the screenshot of your profile that will be used within the study. All names and identifiable data will be omitted – this includes;

- your name as it appears in any form on your profile (i.e. on posts, comments, tags by other people)
- others' names (i.e. names of your friends on your friends list, names of anyone who has commented on your profile or tagged you in anything that appears on your profile)
- profile pictures of anyone who has commented or tagged you in anything that appears on your timeline (the profile pictures in your friends list will remain without their names)
- names of other social media profiles or email addresses that may be shown on your profile.

You will be provided with a further consent form to confirm that you are happy with the screenshot and that you are still willing for your profile to be used within the study. If you confirm this, you will be able to withdraw from the study up to one week after final consent is given, again without having to provide any reasons. If you are unhappy with the screenshot in its final stage, you will be able to withdraw from the study and the screenshot of your profile will not be used.

Please confirm that you have read the information above and confirm that:

- I understand the nature of the research
- I understand how my Facebook profile will be used
- I understand how my data will be stored and destroyed
- I understand that my participation is voluntary and I can withdraw from the research at any time after giving my initial consent on this form, without having to provide my reasoning
- I agree to send the researcher a screenshot of my Facebook profile timeline (as taken on a computer screen not a mobile phone/tablet) to allow for the names and identifiable data to be omitted

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

I hereby fully and freely consent to participate in this part of the study.

Signature: ___________________________________ Date: ______________

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself or my supervisors using the contact details below:

Grace McKenzie (Primary Researcher)
PhD Student
g.mckenzie@lancaster.ac.uk

Professor Paul Taylor
+44 (0)1524 594421
p.j.taylor@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Dr Stacey Conchie
+44 (0)1524 593830
s.conchie@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Kate Cain
Head of Psychology Department
+44 (0)1524 593990
psychology.hod@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Appendix H – Second consent form for the approval of the real profile screenshots

Dear Sir/Madam,

This is a follow up consent form regarding the use of your Facebook profile for research on online deception and the detection of fake personas/identities in the form of social media profiles. Up to this point, you have agreed to send the researcher a screenshot of your Facebook profile timeline. Attached to this form, you will find the screenshot of your profile that will be used within the research – names and identifiable data have been omitted.

Please look at all aspects of the screenshot carefully.

If you are happy with the screenshot to be used within the research, please confirm the following statements below:

- I understand the nature of the research
- I understand how my Facebook profile will be used
- I have reviewed the screenshot of my Facebook profile carefully and confirm that I am happy with the screenshot and hereby consent to its use in this study
- I understand that my participation is voluntary and I can withdraw from the research within 1 week after giving my final consent on this form, without having to provide my reasoning
- I understand how my data will be stored and destroyed

I hereby fully and freely consent for my Facebook profile to be used in this study.

Signature: ____________________________________ **Date:** ______________

If you are not happy with the screenshot of your profile and wish to withdraw from this study, please tick the box below, and sign and date.

□ I wish to withdraw my Facebook profile from this study.

I understand that any data relating to my Facebook profile will be destroyed in line with GDPR guidelines.

Signature: ___________________________________ Date: ______________

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself or my supervisors using the contact details below:

Grace McKenzie (Primary Researcher)
PhD Student
g.mckenzie@lancaster.ac.uk

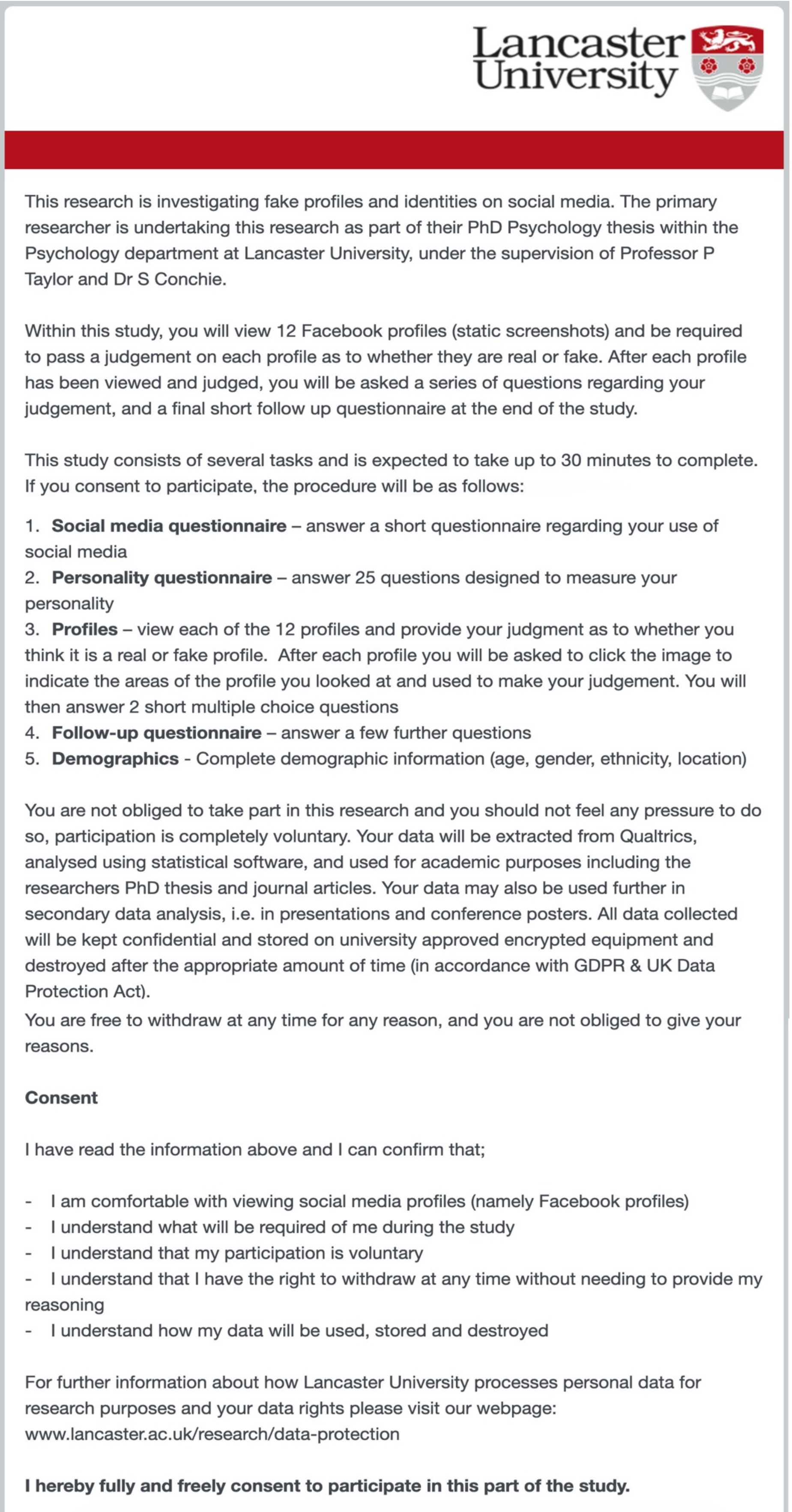
Professor Paul Taylor
+44 (0)1524 594421
p.j.taylor@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Dr Stacey Conchie
+44 (0)1524 593830
s.conchie@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Kate Cain
Head of Psychology Department
+44 (0)1524 593990
psychology.hod@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Appendix I – Participant information sheet and consent form (Qualtrics)



This research is investigating fake profiles and identities on social media. The primary researcher is undertaking this research as part of their PhD Psychology thesis within the Psychology department at Lancaster University, under the supervision of Professor P Taylor and Dr S Conchie.

Within this study, you will view 12 Facebook profiles (static screenshots) and be required to pass a judgement on each profile as to whether they are real or fake. After each profile has been viewed and judged, you will be asked a series of questions regarding your judgement, and a final short follow up questionnaire at the end of the study.

This study consists of several tasks and is expected to take up to 30 minutes to complete. If you consent to participate, the procedure will be as follows:

1. **Social media questionnaire** – answer a short questionnaire regarding your use of social media
2. **Personality questionnaire** – answer 25 questions designed to measure your personality
3. **Profiles** – view each of the 12 profiles and provide your judgment as to whether you think it is a real or fake profile. After each profile you will be asked to click the image to indicate the areas of the profile you looked at and used to make your judgement. You will then answer 2 short multiple choice questions
4. **Follow-up questionnaire** – answer a few further questions
5. **Demographics** - Complete demographic information (age, gender, ethnicity, location)

You are not obliged to take part in this research and you should not feel any pressure to do so, participation is completely voluntary. Your data will be extracted from Qualtrics, analysed using statistical software, and used for academic purposes including the researchers PhD thesis and journal articles. Your data may also be used further in secondary data analysis, i.e. in presentations and conference posters. All data collected will be kept confidential and stored on university approved encrypted equipment and destroyed after the appropriate amount of time (in accordance with GDPR & UK Data Protection Act).
You are free to withdraw at any time for any reason, and you are not obliged to give your reasons.

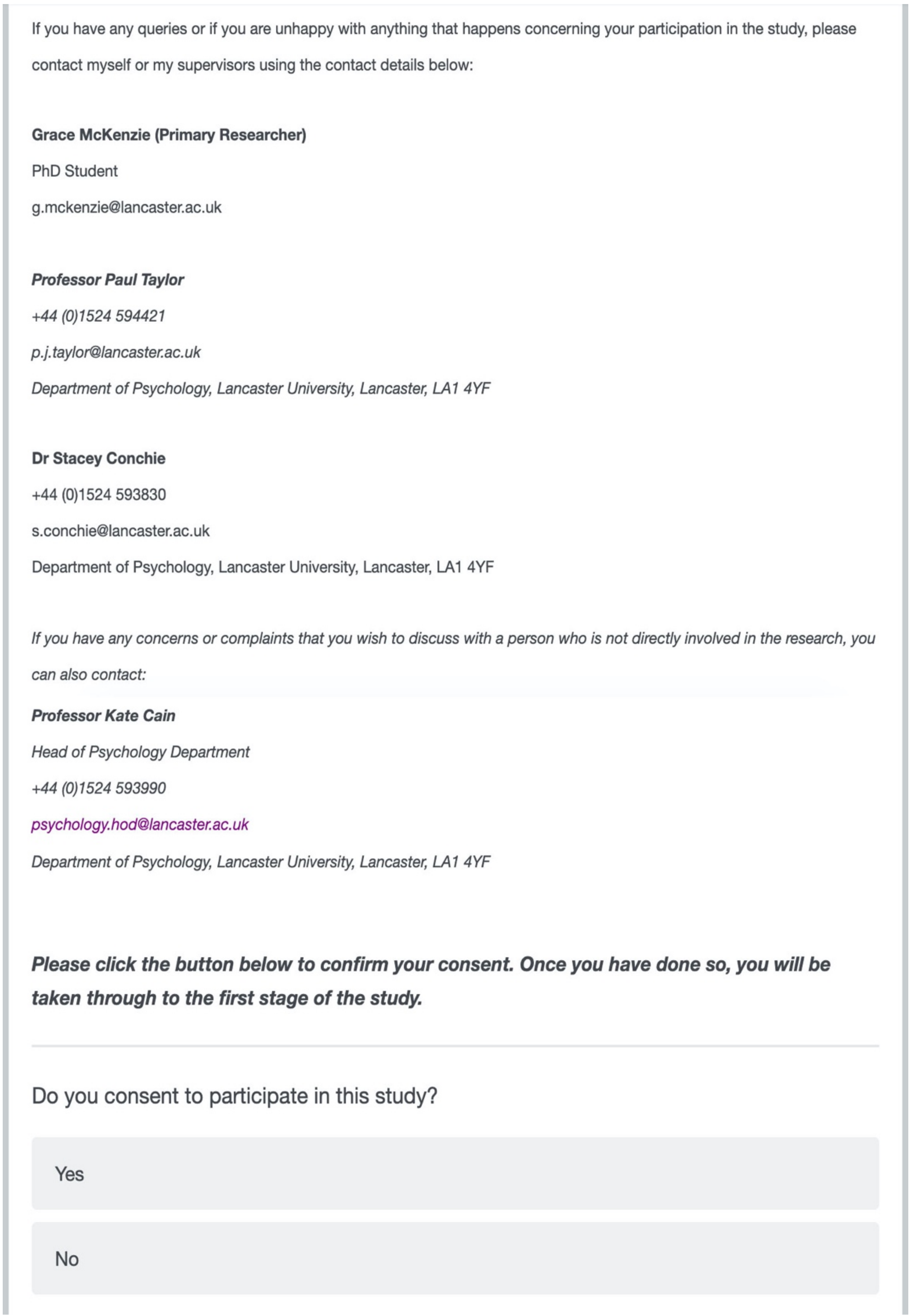
Consent

I have read the information above and I can confirm that;

- I am comfortable with viewing social media profiles (namely Facebook profiles)
- I understand what will be required of me during the study
- I understand that my participation is voluntary
- I understand that I have the right to withdraw at any time without needing to provide my reasoning
- I understand how my data will be used, stored and destroyed

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

I hereby fully and freely consent to participate in this part of the study.



If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself or my supervisors using the contact details below:

Grace McKenzie (Primary Researcher)
PhD Student
g.mckenzie@lancaster.ac.uk

Professor Paul Taylor
+44 (0)1524 594421
p.j.taylor@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Dr Stacey Conchie
+44 (0)1524 593830
s.conchie@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

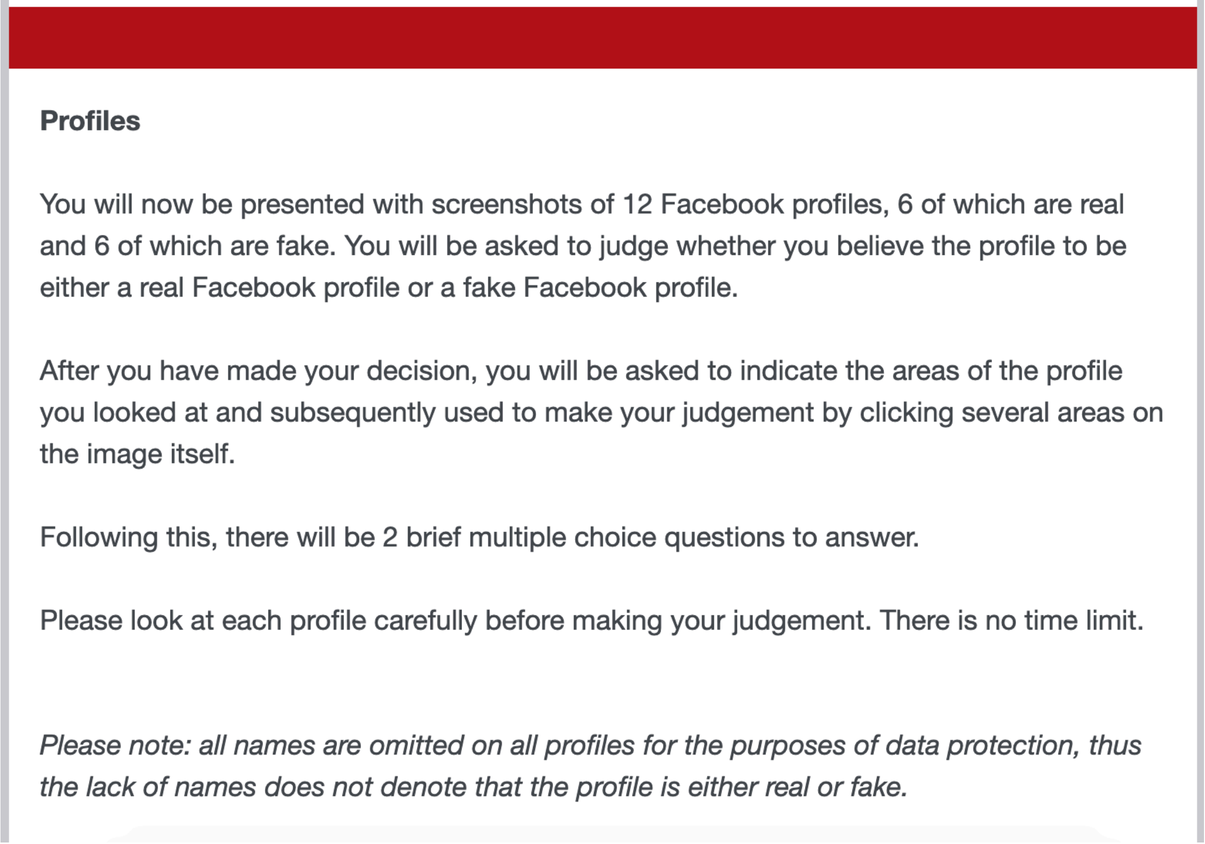
If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Kate Cain
Head of Psychology Department
+44 (0)1524 593990
psychology.hod@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Please click the button below to confirm your consent. Once you have done so, you will be taken through to the first stage of the study.

Do you consent to participate in this study?

Appendix J – Participant instructions for the profile phase of the study (Qualtrics)



Profiles

You will now be presented with screenshots of 12 Facebook profiles, 6 of which are real and 6 of which are fake. You will be asked to judge whether you believe the profile to be either a real Facebook profile or a fake Facebook profile.

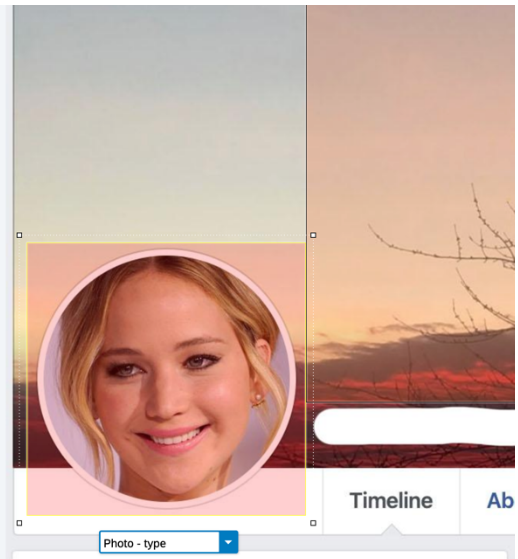
After you have made your decision, you will be asked to indicate the areas of the profile you looked at and subsequently used to make your judgement by clicking several areas on the image itself.

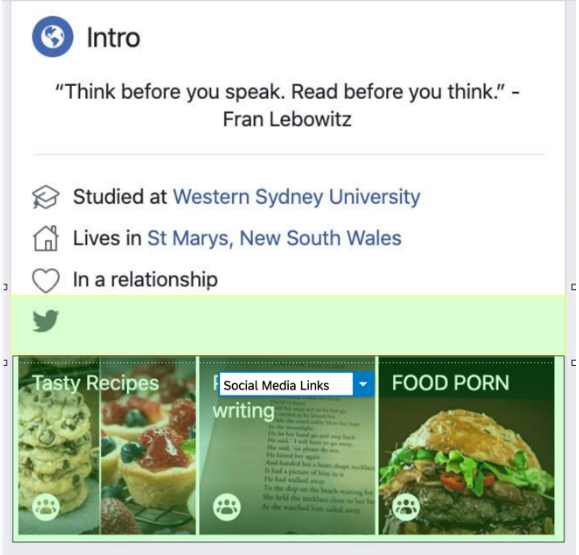
Following this, there will be 2 brief multiple choice questions to answer.

Please look at each profile carefully before making your judgement. There is no time limit.

Please note: all names are omitted on all profiles for the purposes of data protection, thus the lack of names does not denote that the profile is either real or fake.

Appendix K – An example of the Heatmap regions on the profile (visible only to the researcher)



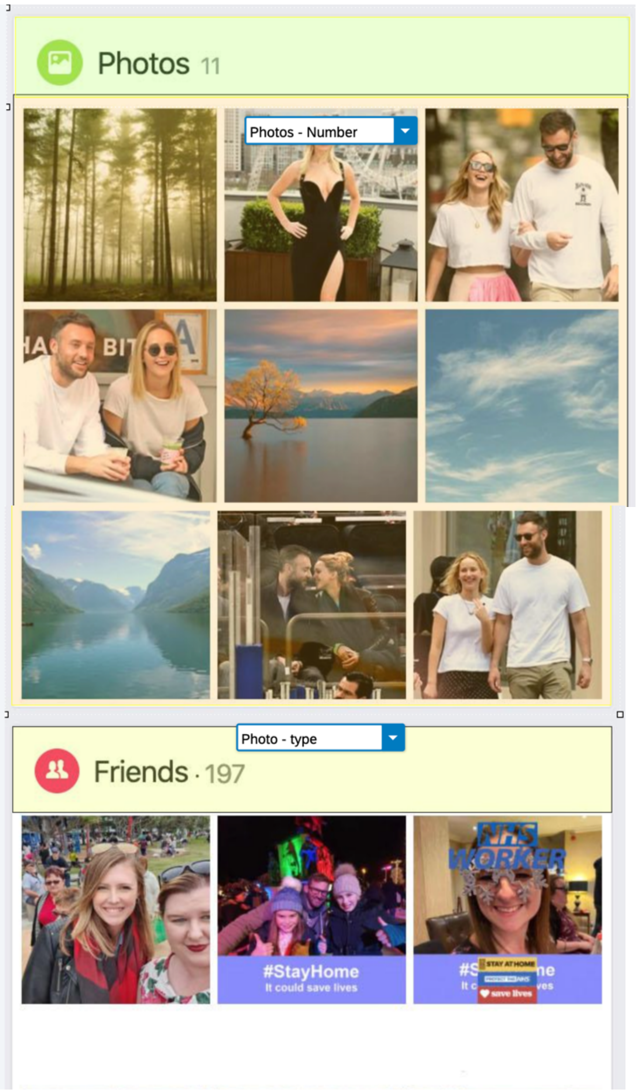

Intro

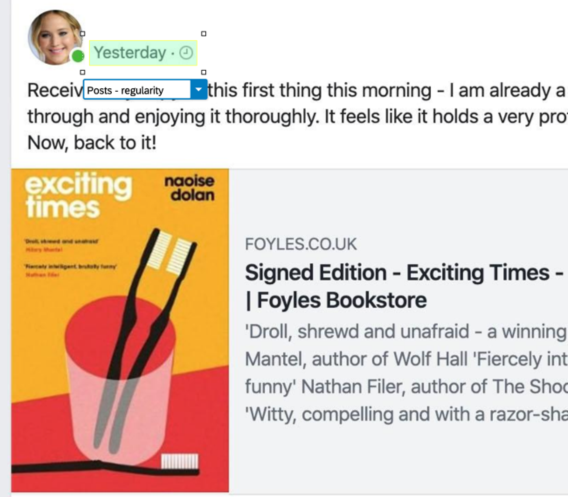
"Think before you speak. Read before you think." - Fran Lebowitz

Studied at Western Sydney University

Lives in St Marys, New South Wales

In a relationship




Yesterday ·

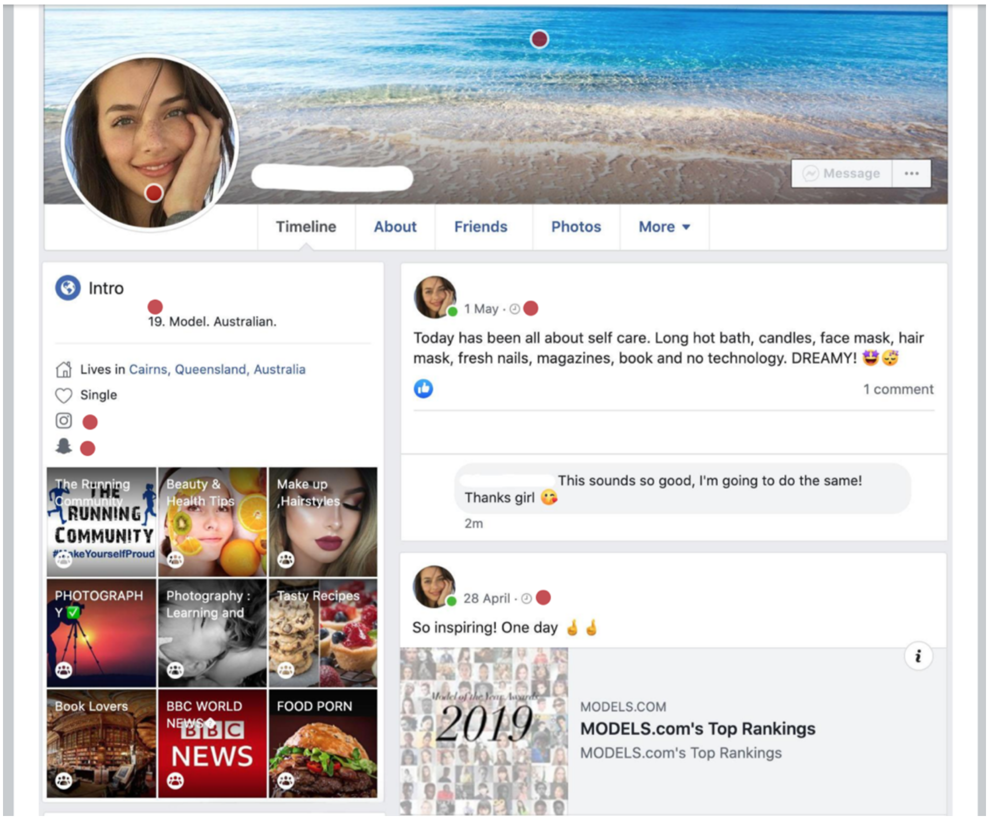
Receiv[Posts - regularity] this first thing this morning - I am already a through and enjoying it thoroughly. It feels like it holds a very pro Now, back to it!

FOYLES.CO.UK

Signed Edition - Exciting Times - | Foyles Bookstore

'Droll, shrewd and unafraid - a winning Mantel, author of Wolf Hall 'Fiercely int funny' Nathan Filer, author of The Shoc 'Witty, compelling and with a razor-sha

Appendix L – Example of the heatmap clicks on a profile as seen by the participant during the study. The red dots denote the areas they have clicked.



19. Model. Australian.

Lives in Cairns, Queensland, Australia

Single

1 May · ⊘

Today has been all about self care. Long hot bath, candles, face mask, hair mask, fresh nails, magazines, book and no technology. DREAMY! 😈😴

1 comment

This sounds so good, I'm going to do the same! Thanks girl 😘

2m

28 April · ⊘

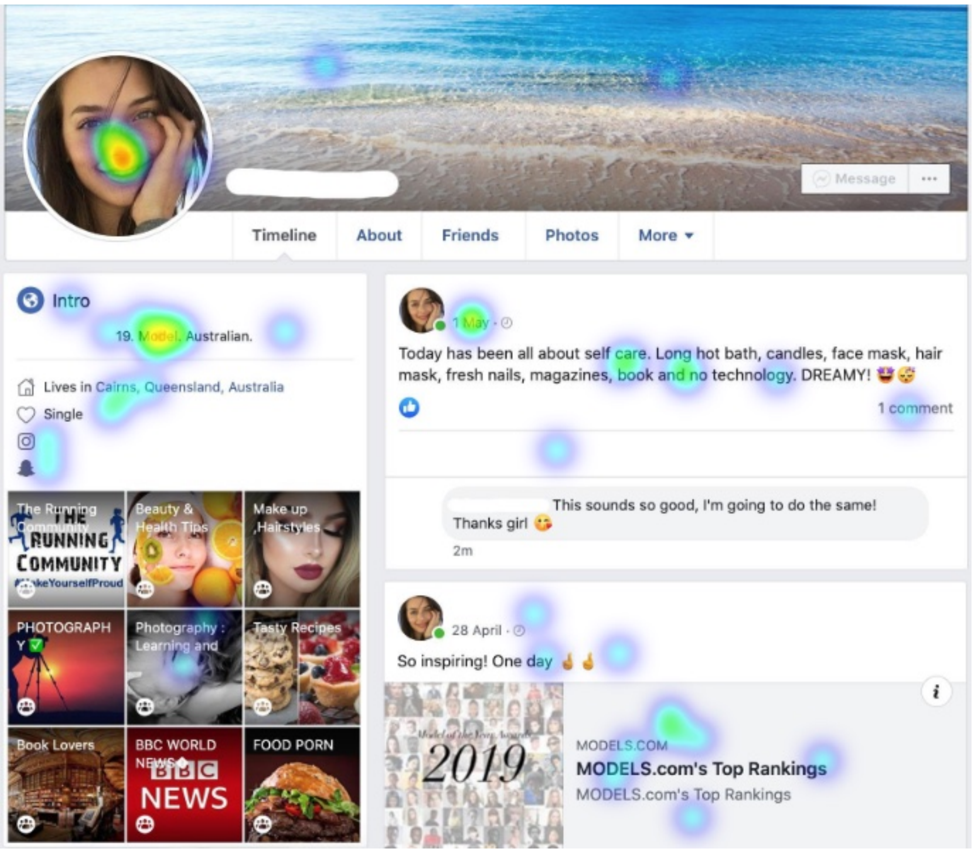
So inspiring! One day 🤞🤞

MODELS.COM

MODELS.com's Top Rankings

MODELS.com's Top Rankings

Appendix L – Example of the results of profile clicks as a heatmap image.



19. Model. Australian.

Lives in Cairns, Queensland, Australia

Single

1 May · ⊘

Today has been all about self care. Long hot bath, candles, face mask, hair mask, fresh nails, magazines, book and no technology. DREAMY! 😈😴

1 comment

This sounds so good, I'm going to do the same! Thanks girl 😘

2m

28 April · ⊘

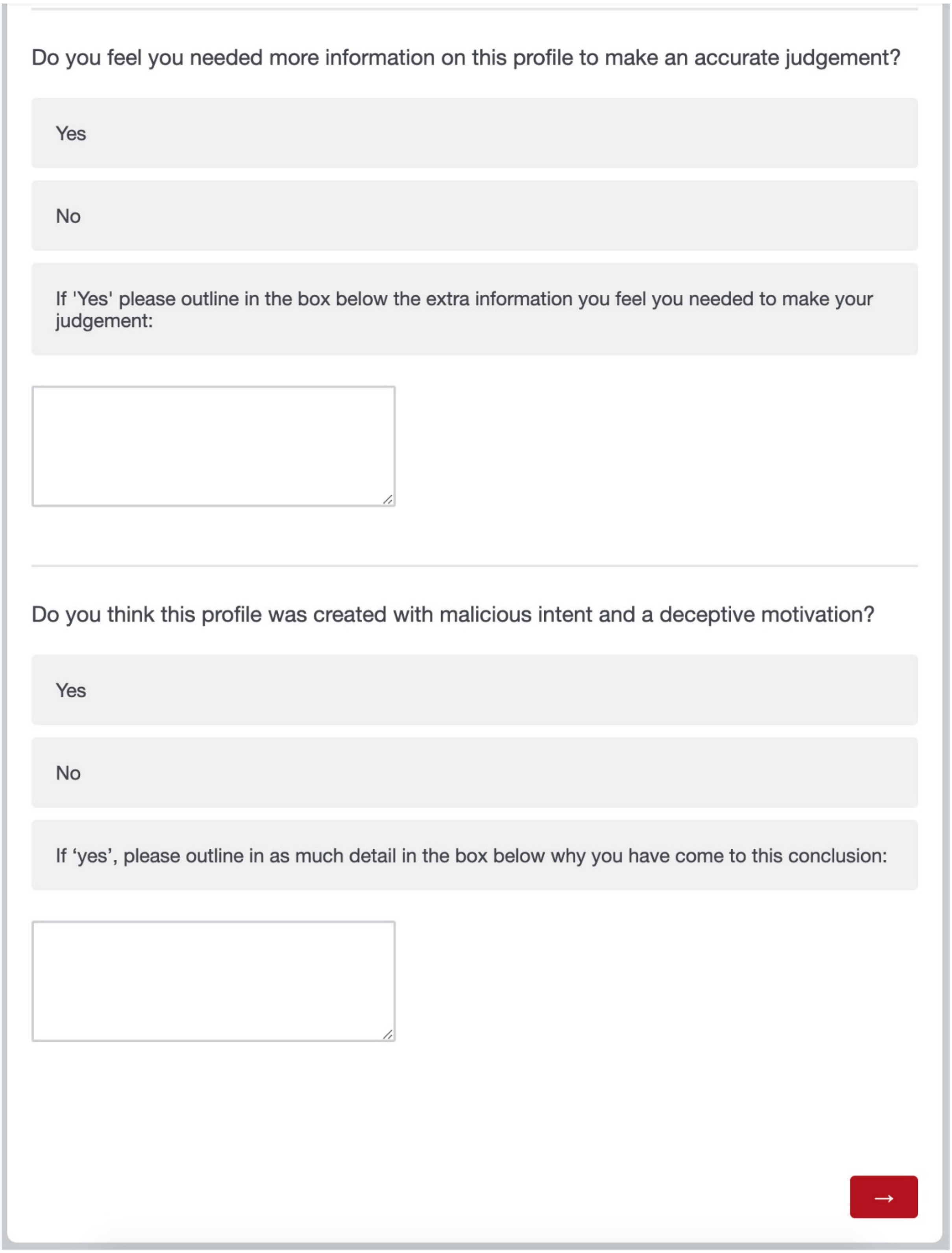
So inspiring! One day 🤞🤞

MODELS.COM

MODELS.com's Top Rankings

MODELS.com's Top Rankings

Appendix M – Questions asked after each profile had been viewed and judged (Qualtrics)



Do you feel you needed more information on this profile to make an accurate judgement?

Yes

No

If 'Yes' please outline in the box below the extra information you feel you needed to make your judgement:

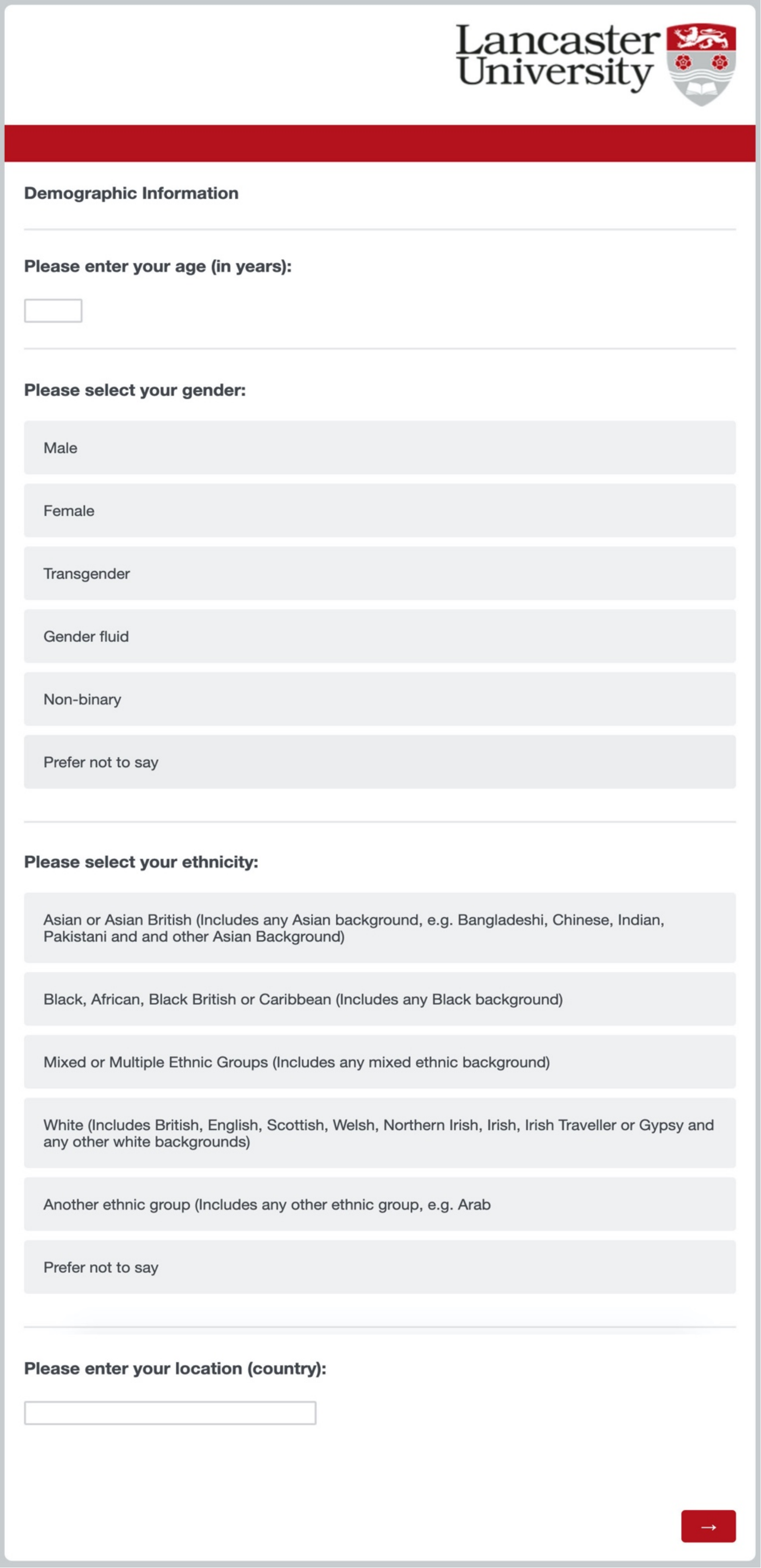
Do you think this profile was created with malicious intent and a deceptive motivation?

Yes

No

If ‘yes’, please outline in as much detail in the box below why you have come to this conclusion:

Appendix N – Demographic information



Demographic Information

Please enter your age (in years):

Please select your gender:

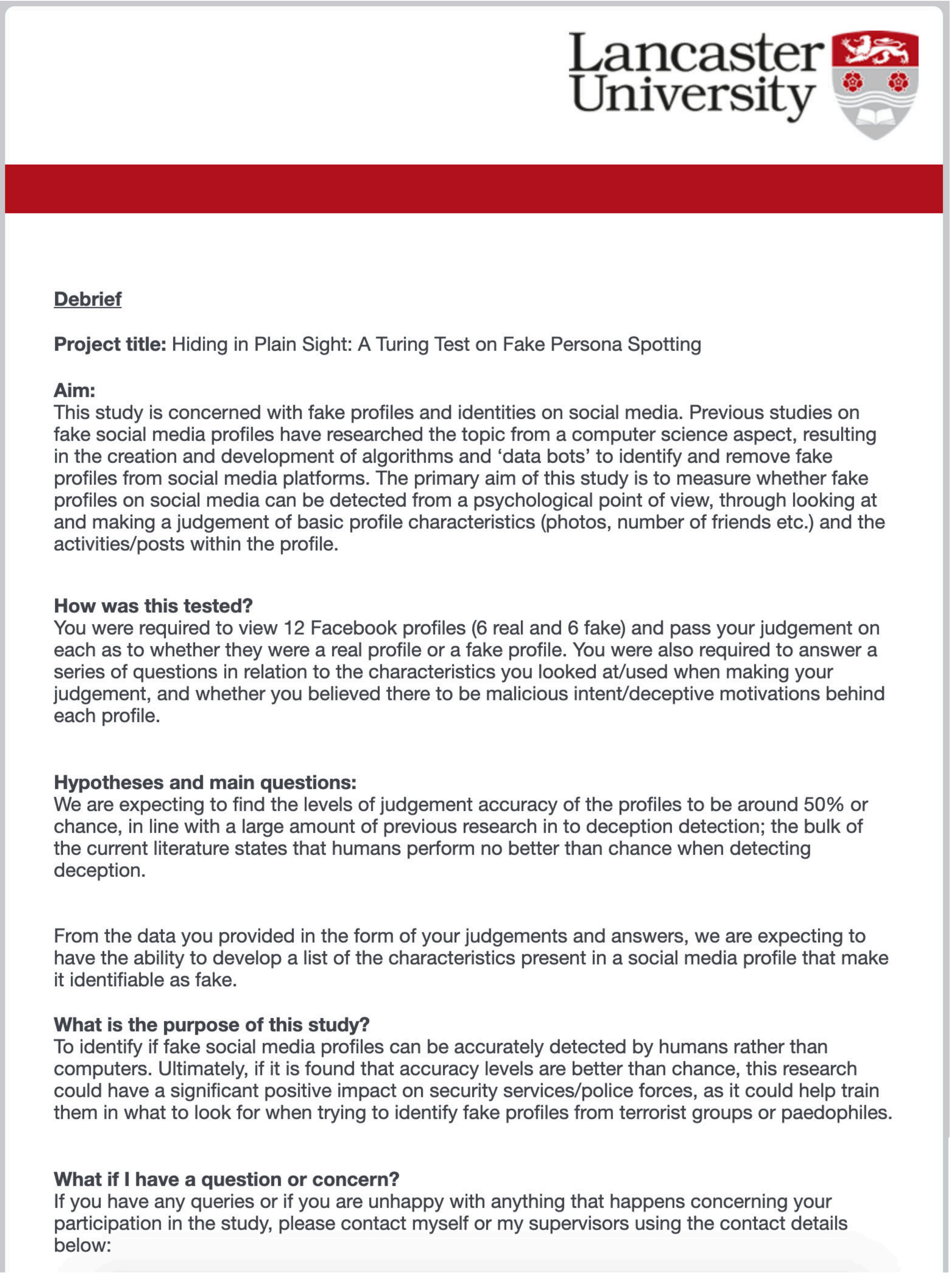
- Male
- Female
- Transgender
- Gender fluid
- Non-binary
- Prefer not to say

Please select your ethnicity:

- Asian or Asian British (Includes any Asian background, e.g. Bangladeshi, Chinese, Indian, Pakistani and and other Asian Background)
- Black, African, Black British or Caribbean (Includes any Black background)
- Mixed or Multiple Ethnic Groups (Includes any mixed ethnic background)
- White (Includes British, English, Scottish, Welsh, Northern Irish, Irish, Irish Traveller or Gypsy and any other white backgrounds)
- Another ethnic group (Includes any other ethnic group, e.g. Arab
- Prefer not to say

Please enter your location (country):

Appendix O – Participant debrief information (Qualtrics)



Debrief

Project title: Hiding in Plain Sight: A Turing Test on Fake Persona Spotting

Aim:
This study is concerned with fake profiles and identities on social media. Previous studies on fake social media profiles have researched the topic from a computer science aspect, resulting in the creation and development of algorithms and 'data bots' to identify and remove fake profiles from social media platforms. The primary aim of this study is to measure whether fake profiles on social media can be detected from a psychological point of view, through looking at and making a judgement of basic profile characteristics (photos, number of friends etc.) and the activities/posts within the profile.

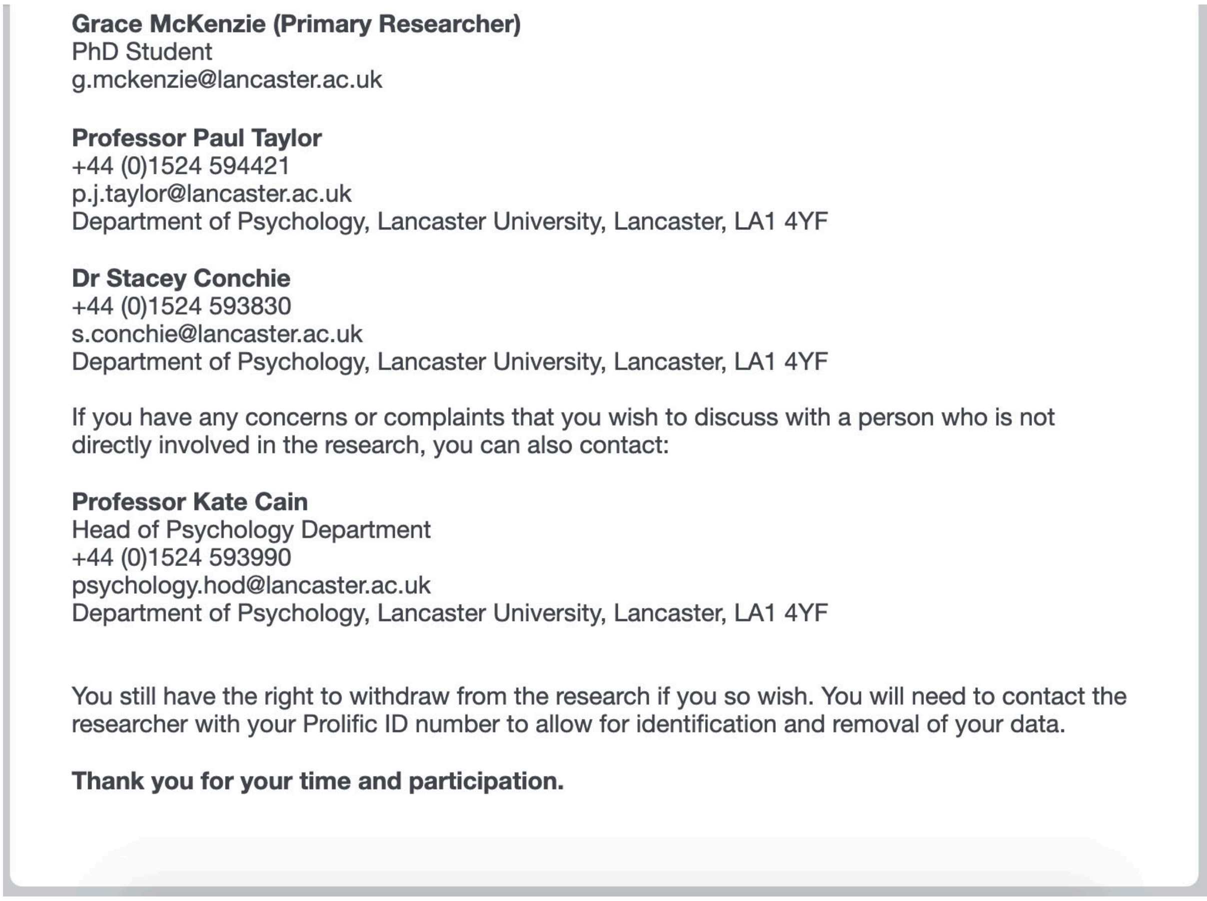
How was this tested?
You were required to view 12 Facebook profiles (6 real and 6 fake) and pass your judgement on each as to whether they were a real profile or a fake profile. You were also required to answer a series of questions in relation to the characteristics you looked at/used when making your judgement, and whether you believed there to be malicious intent/deceptive motivations behind each profile.

Hypotheses and main questions:
We are expecting to find the levels of judgement accuracy of the profiles to be around 50% or chance, in line with a large amount of previous research in to deception detection; the bulk of the current literature states that humans perform no better than chance when detecting deception.

From the data you provided in the form of your judgements and answers, we are expecting to have the ability to develop a list of the characteristics present in a social media profile that make it identifiable as fake.

What is the purpose of this study?
To identify if fake social media profiles can be accurately detected by humans rather than computers. Ultimately, if it is found that accuracy levels are better than chance, this research could have a significant positive impact on security services/police forces, as it could help train them in what to look for when trying to identify fake profiles from terrorist groups or paedophiles.

What if I have a question or concern?
If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself or my supervisors using the contact details below:



Grace McKenzie (Primary Researcher)
PhD Student
g.mckenzie@lancaster.ac.uk

Professor Paul Taylor
+44 (0)1524 594421
p.j.taylor@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Dr Stacey Conchie
+44 (0)1524 593830
s.conchie@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Kate Cain
Head of Psychology Department
+44 (0)1524 593990
psychology.hod@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

You still have the right to withdraw from the research if you so wish. You will need to contact the researcher with your Prolific ID number to allow for identification and removal of your data.

Thank you for your time and participation.

Appendix P - Framework for manipulated characteristics in each profile in STUDIES 2 & 3. Characteristics highlighted in red denote fake and green denote real.

Profiles with 0 fake characteristics:

	CHARACTERISTICS						
PROFILE NAME (Omitted on profile)	**PHOTOS** (Type)	**PHOTOS** (No.)	**BIO** (Description)	**INTRO** (Relationship, Job, Location, Uni etc.)	**POSTS** (Content)	**COMMENTS** (Number)	**LIKES** (No.)
JANE LUDLOW	Selfies	98	Yes	Location	Status Updates Links	6-8 on each post	10-12 on each post
SIMONE LAUDER	Selfies Groups	140	Yes	Location Relationship	Status Updates	12-15 on each post	8-10 on each post
ALFRED INGLOT	Selfies Groups	109	Yes	Relationship Uni/School Job Location	Status Updates	10-12 on each post	12-15 on each post
FREDERICK BROWN	Selfies Groups Pets	87	Yes	Job Relationship	Status Updates Links	20+ on each post	15-20 on each post
GERARD SIMMONS	Selfies Groups Pets	121	Yes	Location Relationship	Status Updates Links	12-15 on each post	12-15 on each post
DWAYNE MCKEY	Selfies Groups	171	Yes	Job Location Relationship	Status Updates	10-15 on each post	20-25 on each post
MAGGIE NIPPON	Selfies Groups Pets	153	Yes	Location Relationship School	Status Updates Links	20+ on each post	20-25 on each post

BELLA TIOUI	Selfies	98	Yes	Location School	Status Updates Links	12-15 on each post	8-10 on each post
SHAQ KHAN	Selfies Groups	139	Yes	Location	Status Updates	6-8 on each post	8-10 on each post
ISOBEL REDMAN	Selfies	94	Yes	Relationship Location	Status Updates Links	10-12 on each post	12-15 on each post
EDWARD PICKLE	Selfies	171	Yes	Uni Job	Status Updates Links	12-15 on each post	8-10 on each post
WILF FREIDMAN	Selfies Groups	103	Yes	Location	Status Updates	10-12 on each post	12-15 on each post

Profiles with 2 fake characteristics:

	CHARACTERISTICS						
PROFILE NAME (Omitted on profile)	**PHOTOS** (Type)	**PHOTOS** (Number)	**BIO** (Description)	**INTRO** (Relationship, Job, Location, Uni etc.)	**POSTS** (Content)	**COMMENTS** (Number)	**LIKES** (Number)
JEMIMA SHAW	Landscapes	12	Yes	Location Job Uni	Status updates	8-10 on each post	15-20 on each post
LUCY NEWT	Celebrities	56	No	Relationship Location	Status Updates Links	12-15 on each post	18 -22 on each post
ADAM HARRIS	Landscapes Artwork	125	Yes	No	Status Updates Links Sharing Videos	8-10 on each post	10-12 on each post
NOAH STEEL	Celebrities Cartoon Characters	109	Yes	Job Uni	Photo Update Posts	50+ on each post	8-10 on each post
LOIS PICKER	Celebrities	98	Yes	Location	Photo Update Posts Status Updates Links	No Comments on any posts	12-15 on each post
REUBEN JONES	Landscapes	73	Yes	Job Uni Location	Status Updates Links	6-8 on each post	1 -2 likes on each post
WENDY BROWN	Selfies Groups	8	No	Location Job	Status Updates	10-12 on each post	10-12 on each post

SHIA KHALID	Selfies	15	Yes	No	Status Updates Links	6-8 on each post	12-15 on each post
JORDAN JUDD	Selfies Groups Pets	11	Yes	Job Uni Location	Photo Update Posts	20+ on each post	8-10 on each post
KYA LEI	Selfies	9	Yes	Location Job	Status Updates Links	1-2 comments on each post	12-15 on each post
PETER STOCKSON	Groups Pets	6	Yes	Location	Status Updates Links Sharing Videos	6-8 on each post	100+ likes on each post
HEIDI LUMEN	Selfies	192	No	No	Status Updates Links	6-8 on each post	12-15 on each post
FRANCES APPLE	Selfies	87	No	Location Job	Video Shares	12-15 on each post	8-10 on each post
PHILLIP TONE	Selfies	58	No	Job Uni Location	Status Updates	2-3 comments on each post	15-20 on each post
SOFIE DRUID	Selfies Groups Pets	94	No	Job Uni Location	Status Updates Links	20+ on each post	No likes on any posts
HAZEL REDWOOD	Selfies Groups	134	Yes	No	Video Shares	6-8 on each post	8-10 on each post

KEITH GORGE	Selfies Groups	171	Yes	No	Status Updates	2-3 comments on each post	12-15 on each post
FAISAL KHAN	Selfies Groups	92	Yes	No	Status Updates Links	10-12 on each post	No likes on any posts
TORI ADAMSON	Selfies	79	Yes	Uni Job	Photo Update Posts	No comments on each post	8-10 on each post
NIALL O'SHEA	Selfies Groups	66	Yes	Job Location	Video Shares	6-8 comments on each post	100+ on each post
MARK EDGEWARE	Selfies Groups	38	Yes	Location Uni	Status Updates Video Shares	1-2 comments on each post	No likes on any posts

Profiles with 4 fake characteristics:

	CHARACTERISTICS						
PROFILE NAME (Omitted on profile)	**PHOTOS** (Type)	**PHOTOS** (No.)	**BIO** (Description)	**INTRO** (Relationship, Job, Location, Uni etc.)	**POSTS** (Content)	**COMMENTS** (Number)	**LIKES** (No.)
YANA INDIGO	Celebrities	17	No	No	Status Updates Links	6-8 on each post	10-12 on each post
AMY BESWICK	Landscapes	9	No	Location Uni	Photo Update Posts	12-15 on each post	8-10 on each post
SUSAN CHARLES	Landscapes Artwork	5	No	Job Uni	Status Updates	No Comments on any posts	12-15 on each post
MOHAMMED IQBAL	Celebrities	12	No	Location	Status Updates Links	10-12 on each post	1-2 likes on each post
PENNY WITHERS	Landscapes Artwork	6	Yes	No	Video Shares	6-8 on each post	8-10 on each post
ARNOLD GRUNIS	Celebrities	8	Yes	No	Status Updates	1-2 on each post	10-12 on each post
BRIAN YENSEN	Celebrities	15	Yes	No	Status Updates Links	6-8 on each post	1-2 likes on each post
MELISSA DEWORTH	Artwork	7	Yes	Location	Photo Update Posts	1-2 on each post	8-10 on each post

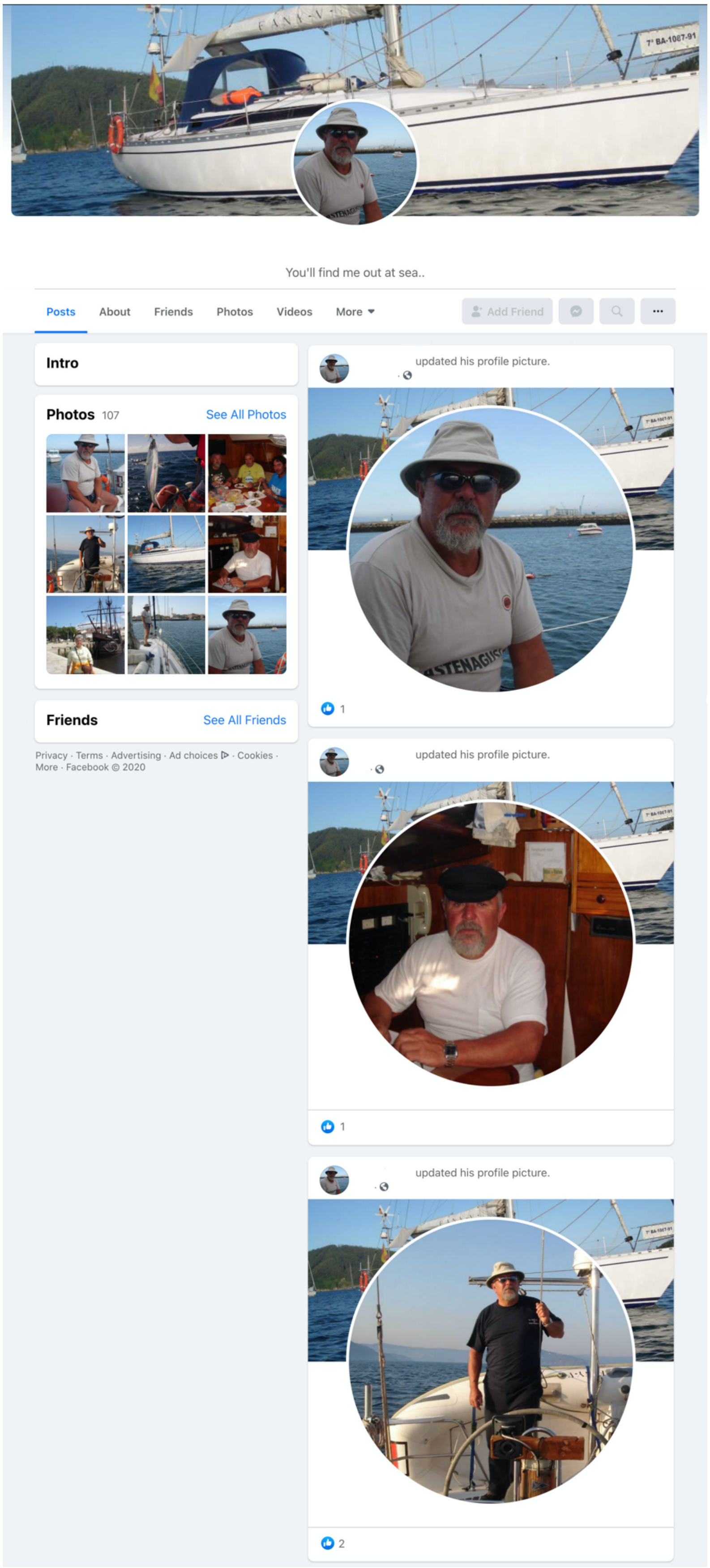
TIFFANY CHAMBERS	Artwork	17	Yes	Uni	Video Shares	10-12 on each post	No likes on any posts
MARSHA BRIGHT	Celebrities	83	No	No	Photo Update Posts	6-8 on each post	8-10 on each post
GRAHAM LINMAN	Landscapes	103	No	No	Status Updates	No Comments on any posts	12-15 on each post
REGGIE WARRINGTON	Celebrities	94	No	No	Status Updates	6-8 on each post	1-2 likes on each post
STEVE SHYBOLT	Celebrities	71	No	Job Uni Location	Video Shares	10-12 on each post	1-2 likes on each post
WARREN CHAN	Landscapes	73	Yes	No	Photo Update Posts	1-2 on each post	8-10 on each post
HASSAN OMAR	Artwork	71	Yes	No	Photo Update Posts	6-8 on each post	1-2 likes on each post
GERALDINE KNOCKWORTH	Celebrities	56	Yes	Location	Video Shares	1-2 on each post	1-2 likes on each post

MARTIN ILLINGTON	Selfies	13	No	No	Video Shares	6-8 on each post	8-10 on each post
TONY McKEE	Selfies Groups	8	No	No	Status Updates Links	1-2 on each post	10-12 on each post
MARIE ANTSON	Selfies Groups	6	No	No	Status Updates Links	12-15 on each post	No likes on any posts
EWA SLZENZA	Selfies	2	Yes	Location Job Uni	Photo Update Posts	6-8 on each post	1-2 likes on each post
GABRIELLA FONTES	Selfies Groups	6	Yes	No	Photo Update Posts	No Comments on any posts	6-8 on each post
MIRIAM PEZNIK	Selfies	4	Yes	No	Video Shares	6-8 on each post	No likes on any posts
STEFAN GRUBNER	Selfies	11	Yes	No	Status Updates	1-2 on each post	1-2 likes on each post

PAVOL KUSHELNIK	Selfies Groups	8	Yes	Location	Video Shares	1-2 on each post	1-2 likes on each post
DIMITRI DIMITRU	Selfies Pets	88	No	No	Video Shares	1-2 on each post	10-12 on each post
HELEN FIELDING	Selfies Groups	94	No	No	Photo Update Posts	6-8 on each post	1-2 likes on each post
JACOB JACOBSON	Selfies Groups	56	No	No	Status Updates	1-2 on each post	1-2 likes on each post
ISAAC BURTON	Selfies	140	No	Job	Photo Update Posts	No Comments on any posts	1-2 likes on each post
SHANE ELSTREE	Celebrities	134	No	Location	Video Shares	1-2 on each post	10-12 on each post
EDWARD KNIGHT	Selfies	17	No	Job	Video Shares	1-2 on each post	6-8 on each post
SASKIA MARIE HUNTLEY	Celebrities	18	Yes	Job	Status Updates	No Comments on any posts	1-2 likes on each post

BRITTNEY TETHERS	Artwork	94	No	Job	Status Updates	No Comments on any posts	1-2 likes on each post
STAR TEETON	Artwork	73	Yes	No	Status Updates	1-2 on each post	1-2 likes on each post
RODNEY VALE	Selfies	3	No	Job	Status Updates	1-2 on each post	No likes on any posts
ANGEL COSTAS	Selfies Groups Pets	107	Yes	No	Photo Update Posts	No comments on any post	1-2 likes on each post

Appendix Q – Example of a Facebook profile screenshot with new Facebook layout



You'll find me out at sea..

updated his profile picture.

1

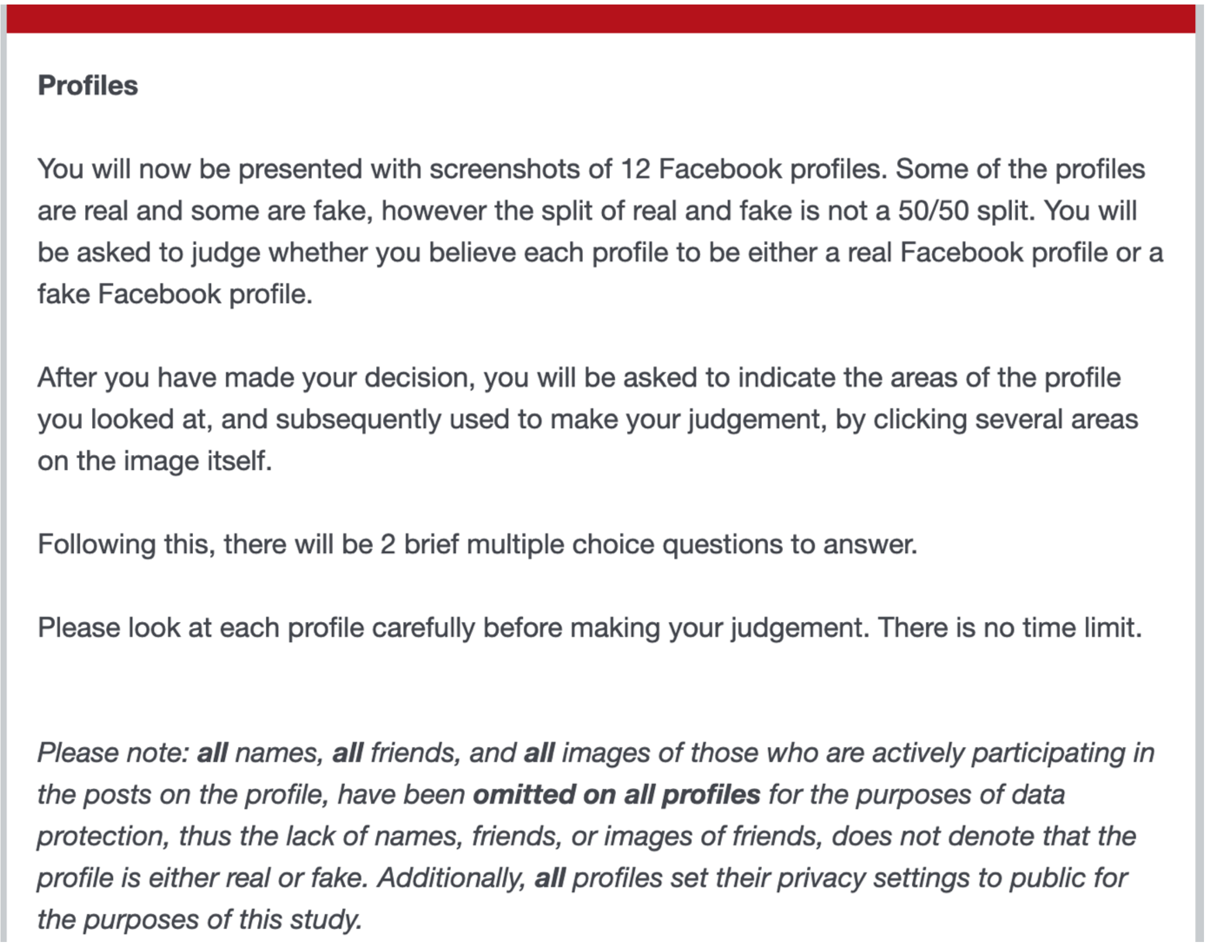
updated his profile picture.

1

updated his profile picture.

2

Appendix R – Updated participant instructions for the profile phase of the study in Study 3



Profiles

You will now be presented with screenshots of 12 Facebook profiles. Some of the profiles are real and some are fake, however the split of real and fake is not a 50/50 split. You will be asked to judge whether you believe each profile to be either a real Facebook profile or a fake Facebook profile.

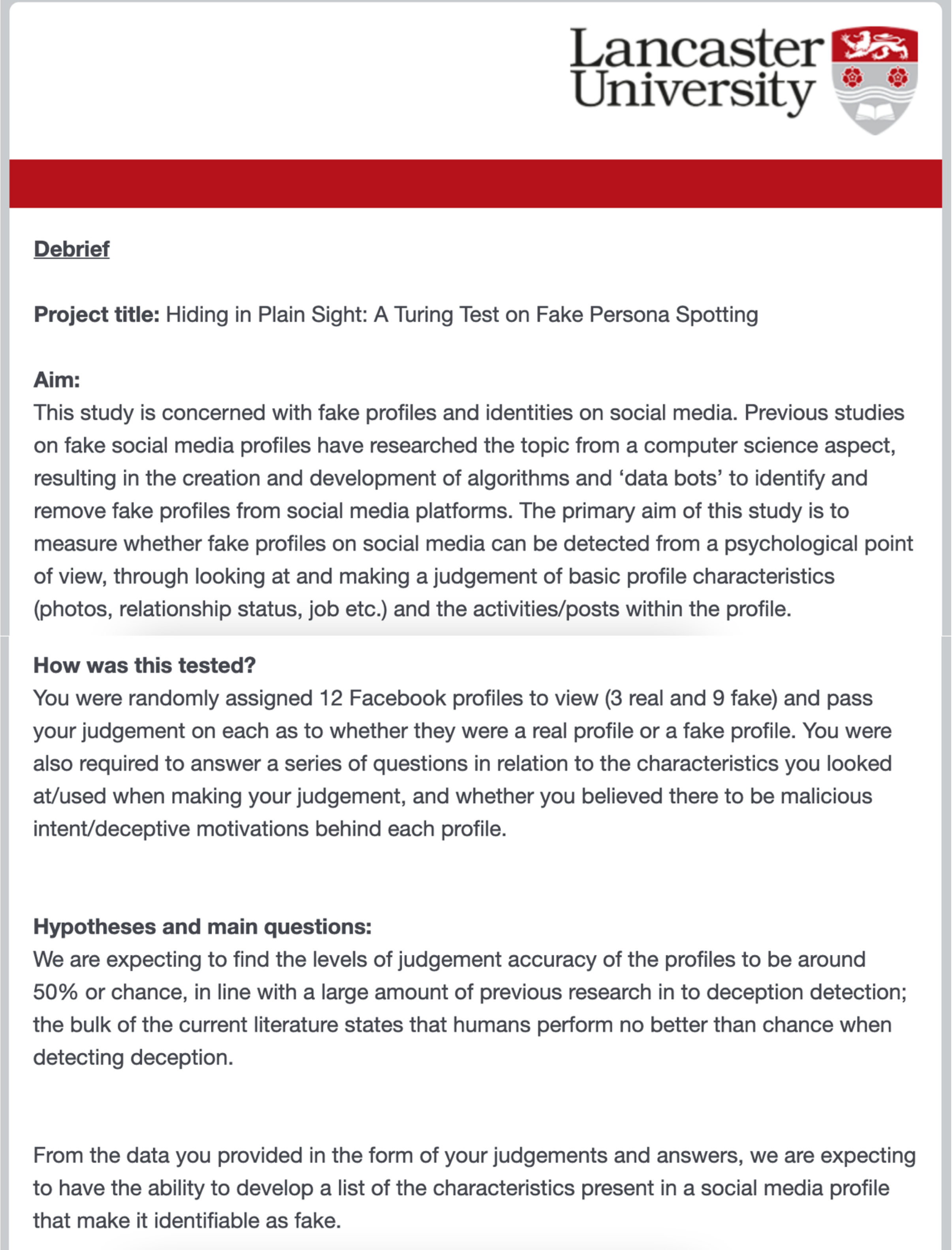
After you have made your decision, you will be asked to indicate the areas of the profile you looked at, and subsequently used to make your judgement, by clicking several areas on the image itself.

Following this, there will be 2 brief multiple choice questions to answer.

Please look at each profile carefully before making your judgement. There is no time limit.

*Please note: **all** names, **all** friends, and **all** images of those who are actively participating in the posts on the profile, have been **omitted on all profiles** for the purposes of data protection, thus the lack of names, friends, or images of friends, does not denote that the profile is either real or fake. Additionally, **all** profiles set their privacy settings to public for the purposes of this study.*

Appendix S – Update debrief form for participants in Study 3



Lancaster University

Debrief

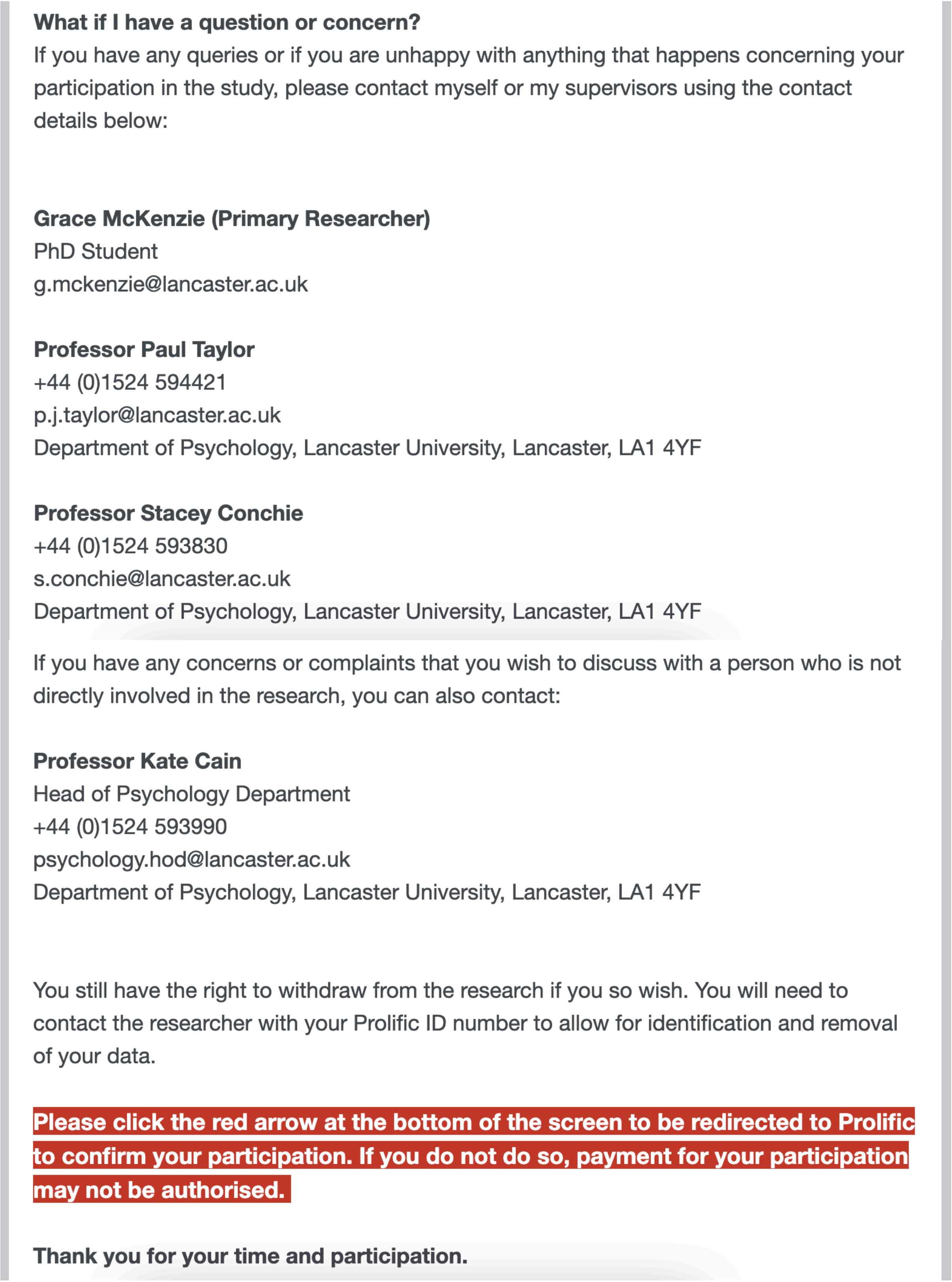
Project title: Hiding in Plain Sight: A Turing Test on Fake Persona Spotting

Aim:
This study is concerned with fake profiles and identities on social media. Previous studies on fake social media profiles have researched the topic from a computer science aspect, resulting in the creation and development of algorithms and 'data bots' to identify and remove fake profiles from social media platforms. The primary aim of this study is to measure whether fake profiles on social media can be detected from a psychological point of view, through looking at and making a judgement of basic profile characteristics (photos, relationship status, job etc.) and the activities/posts within the profile.

How was this tested?
You were randomly assigned 12 Facebook profiles to view (3 real and 9 fake) and pass your judgement on each as to whether they were a real profile or a fake profile. You were also required to answer a series of questions in relation to the characteristics you looked at/used when making your judgement, and whether you believed there to be malicious intent/deceptive motivations behind each profile.

Hypotheses and main questions:
We are expecting to find the levels of judgement accuracy of the profiles to be around 50% or chance, in line with a large amount of previous research in to deception detection; the bulk of the current literature states that humans perform no better than chance when detecting deception.

From the data you provided in the form of your judgements and answers, we are expecting to have the ability to develop a list of the characteristics present in a social media profile that make it identifiable as fake.



What if I have a question or concern?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself or my supervisors using the contact details below:

Grace McKenzie (Primary Researcher)
PhD Student
g.mckenzie@lancaster.ac.uk

Professor Paul Taylor
+44 (0)1524 594421
p.j.taylor@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

Professor Stacey Conchie
+44 (0)1524 593830
s.conchie@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Kate Cain
Head of Psychology Department
+44 (0)1524 593990
psychology.hod@lancaster.ac.uk
Department of Psychology, Lancaster University, Lancaster, LA1 4YF

You still have the right to withdraw from the research if you so wish. You will need to contact the researcher with your Prolific ID number to allow for identification and removal of your data.

Please click the red arrow at the bottom of the screen to be redirected to Prolific to confirm your participation. If you do not do so, payment for your participation may not be authorised.

Thank you for your time and participation.

Appendix T – Updated participant instructions for the profile phase of the study in Study 4

Profiles

You will now be presented with screenshots of 12 Facebook profiles. Some of the profiles are real and some are fake, however the split of real and fake is not a 50/50 split. You will be asked to judge whether you believe each profile to be either a real Facebook profile or a fake Facebook profile.

You will be shown each profile twice.

The first time you see the profile you will have a time limit of 40 seconds to view the profile. Once the 40 seconds is over, the questionnaire will automatically move on to the next profile, however you can choose to move on before the 40 seconds is over. Please look at each profile carefully before moving on.

Once the 40 seconds is over, you will be asked to make your judgement by selecting that the profile is either 'fake' or 'real'.

After you have made your judgement, you will be shown the **same profile for a second time** and asked to indicate the areas of the profile you looked at, and subsequently used to make your judgement, by clicking several areas on the image itself. **There is no time limit for you to make your clicks.**

Following this, there will be 2 brief multiple choice questions to answer.

*Please note: **all** names, **all** friends, and **all** images of those who are active*ly participating *in the posts on the profile, have been **omitted on all profiles** for the purposes of data protection, thus the lack of names, friends, or images of friends, does not denote that the profile is either real or fake. Additionally, **all** profiles set their privacy settings to public for the purposes of this study.*

Appendix U - Framework for manipulated characteristics in each profile in STUDY 5 – INDIAN PROFILES.
Characteristics highlighted in red denote fake and green denote real.

0 Fakes

	CHARACTERISTICS						
PROFILE NAME (Omitted on profile)	**PHOTOS** (Type)	**PHOTOS** (No.)	**BIO** (Description)	**INTRO** (Relationship, Job, Location, Uni etc.) ADD LANCASTER AS LOCATION/UNI IN A FEW	**POSTS** (Content)	**COMMENTS** (Number)	**LIKES** (No.)
Sheela Khatri (F)	Selfies	506	Yes	Location	Status Updates Photos	6-8 on each post	50-60 on each post
Jaya Anand (F)	Selfies Groups	612	Yes	Location Relationship-Married	Status Updates Photos	12-15 on each post	70-80 on each post
Ishaan Banerjee (M)	Selfies Groups	587	Yes	Relationship - Married Uni/School Job Location	Status Updates Photos	10-12 on each post	60-70 on each post
Anjali Patel (F)	Selfies Groups	787	Yes	Job Relationship-Married	Status Updates Links	20+ on each post	80-90 on each post
Sahil Burman (M)	Selfies Groups	821	Yes	Location Relationship - Single	Status Updates Links Photos	12-15 on each post	50-60 on each post
Ashwin Bhatt (M)	Selfies Groups	771	Yes	Job Location Relationship - Single	Status Updates Photos	10-15 on each post	90-100 on each post

Ambar Bakshi (F)	Selfies Groups	653	Yes	Location Relationship - Single School	Status Updates Links	20+ on each post	30-40 on each post
Farid Basu (M)	Selfies	998	Yes	Location School	Status Updates Links Photos	12-15 on each post	40-50 on each post
Chnader Varma (M)	Selfies Groups	839	Yes	Location	Status Updates Photos	6-8 on each post	30-40 on each post
Artha Babu (F)	Selfies Groups	694	Yes	Relationship - Married Location	Status Updates Links	10-12 on each post	60-70 on each post
Lata Arya (F)	Selfies	671	Yes	Uni Job	Status Updates Links	12-15 on each post	50-60 on each post
Johar Dalal (M)	Selfies Groups	903	Yes	Location	Status Updates Photos	10-12 on each post	40-50 on each post

2 Fakes

	CHARACTERISTICS						
PROFILE NAME (Omitted on profile)	**PHOTOS** (Type)	**PHOTOS** (Number)	**BIO** (Description)	**INTRO** (Relationship, Job, Location, Uni etc.) ADD LANCASTER AS LOCATION IN A FEW	**POSTS** (Content)	**COMMENTS** (Number)	**LIKES** (Number)
Mahesh Chowdury (M)	Landscapes	12	Yes	Location Job Uni	Photo Update Posts Links	8-10 on each post	15-20 on each post
Kamala Mangal (F)	Celebrities	556	No	Relationship - Married Location	Photo Update Posts	12-15 on each post	18 -22 on each post
Deshad Chabra (M)	Landscapes Artwork	825	Yes	No	Photo Update Posts Links	8-10 on each post	10-12 on each post
Sajan Chawla (M)	Celebrities Cartoon Characters	609	Yes	Job Uni	Status Updates – Personal thoughts	50+ on each post	8-10 on each post
Anju Malhotra (F)	Celebrities	798	Yes	Location	Photo Update Posts Links	No Comments on any posts	12-15 on each post
Indu Jha (F)	Landscapes	673	Yes	Job Uni Location	Photo Update Posts Links Check-ins	6-8 on each post	1 -2 likes on each post
Ahjit Amin (M)	Selfies Groups	78	No	Location Job	Photo Update Posts Links	10-12 on each post	10-12 on each post

Nalin Apti (M)	Selfies	15	Yes	No	Links Photos	6-8 on each post	12-15 on each post
Alka Joshi (F)	Selfies Groups Pets	91	Yes	Job Uni Location	Status Updates – Personal thoughts	20+ on each post	8-10 on each post
Shakila Kapadia (F)	Selfies	93	Yes	Location Job	Photo Update Posts Links	300-350 comments on each post	12-15 on each post
Darsha Iyer (F)	Groups Pets	56	Yes	Location	Photo Update Posts Links Check ins	6-8 on each post	100+ likes on each post
Balraj Datta (M)	Selfies	592	No	No	Photo Update Posts Links	6-8 on each post	12-15 on each post
Adry Khanna (F)	Selfies	987	No	Location Job Relationship – Single	Video Shares	12-15 on each post	8-10 on each post
Mahavir Deol (M)	Selfies	858	No	Job Uni Location	Photo Update Posts Links	0 comments on each post	15-20 on each post
Ashok Lal (M)	Selfies Groups Pets	894	No	Job Uni Location	Photo Update Posts Links	20+ on each post	No likes on any posts
Charita Kaur (F)	Selfies Groups	634	Yes	No	Video Shares	6-8 on each post	8-10 on each post

Anjay Dewan (M)	Selfies Groups	771	Yes	No	Photo Update Posts Links	1-2 comments on each post	12-15 on each post
Vanita Kashyap (F)	Selfies Groups	992	Yes	No	Photo Update Posts Links Check ins	10-12 on each post	No likes on any posts
Andal Kohli (M)	Selfies	579	Yes	Uni Job Relationship – Single	Status Updates – Personal thoughts	No comments on each post	8-10 on each post
Chandini Ghosh (F)	Selfies Groups	866	Yes	Job Location	Video Shares	6-8 comments on each post	100+ on each post
Tenaya Garg (F)	Selfies Groups	738	Yes	Location Uni	Photo Update Posts Links Check ins	200-250 comments on each post	No likes on any posts

4 Fakes

	CHARACTERISTICS						
PROFILE NAME (Omitted on profile)	**PHOTOS** (Type)	**PHOTOS** (No.)	**BIO** (Description)	**INTRO** (Relationship, Job, Location, Uni etc.) ADD LANCASTER AS LOCATION/ UNI IN A FEW	**POSTS** (Content)	**COMMENTS** (Number)	**LIKES** (No.)
Nisha Dhar (F)	Celebrities	17	No	No	Photo Update Posts Links Check-ins	6-8 on each post	90-100 on each post
Dillip Biswas (M)	Landscapes	69	No	Location Uni	Status Updates – Personal thoughts	12-15 on each post	90-100 on each post
Azha Ghandi (F)	Landscapes Artwork	35	No	Job Uni	Photo Update Posts Links	No Comments on any posts	50-60 on each post
Soma Gupta (F)	Celebrities	92	No	Location Relationship – Married	Photo Update Posts Links	10-12 on each post	200+ likes on each post
Gulshan Batra (M)	Landscapes Artwork	64	Yes	No	Video Shares	6-8 on each post	80-90 on each post
Neena Das (F)	Celebrities	87	Yes	No	Photo Update Posts	1-2 on each post	30-40 on each post
Harshad Metra (M)	Celebrities	15	Yes	No	Photo Update Posts Links	6-8 on each post	200+ likes on each post

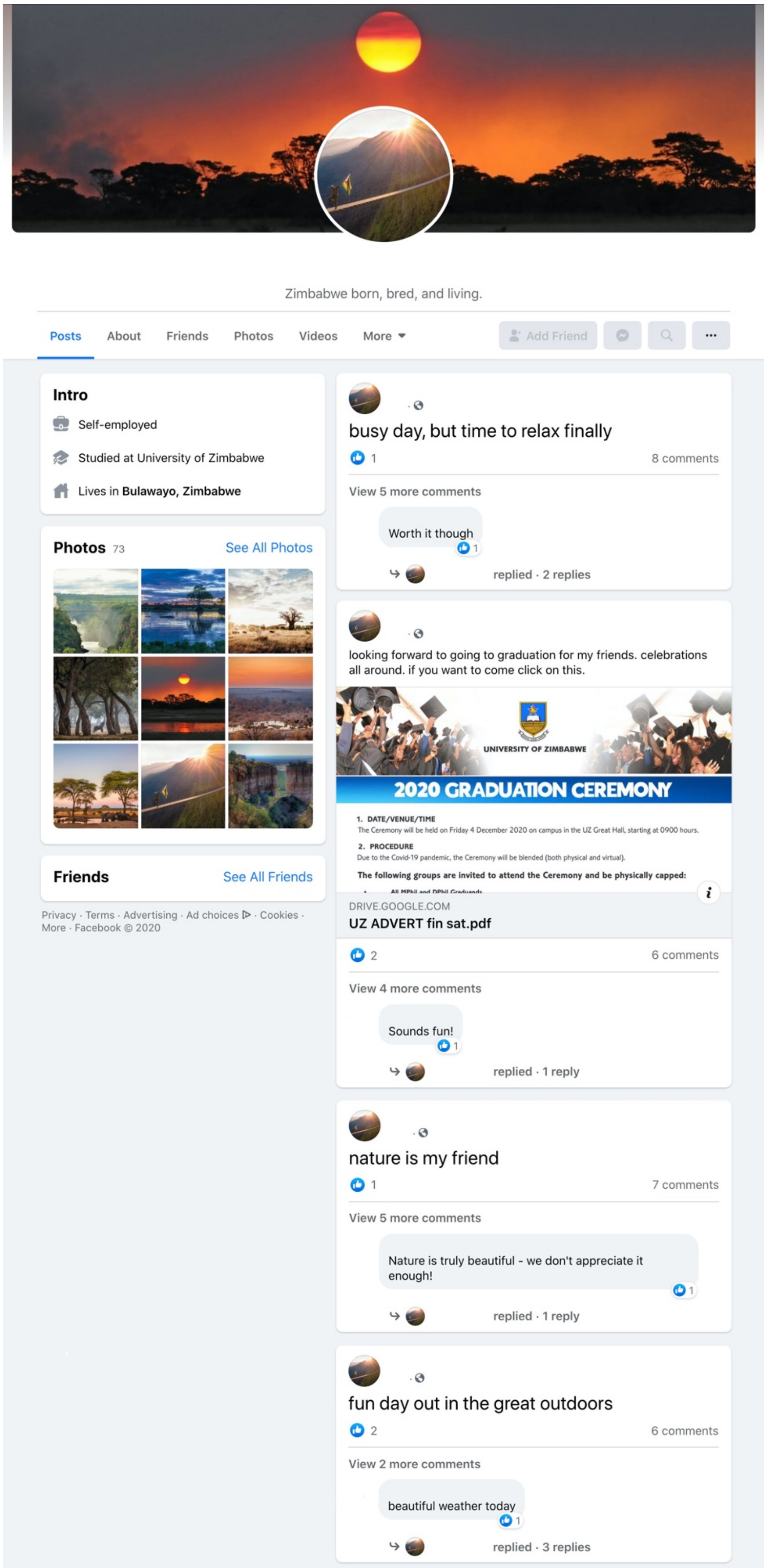
Deva Chopra (F)	Artwork	79	Yes	Location	Status Updates – Personal thoughts	100-150 on each post	80-90 on each post
Khalinda Dixit (F)	Artwork	87	Yes	Uni	Video Shares	10-12 on each post	No likes on any posts
Anura Haldar (F)	Celebrities	830	No	No	Status Updates – Personal thoughts	6-8 on each post	60-70 on each post
Jaladi Mukherjee (M)	Landscapes	903	No	No	Photo Update Posts Links Check-ins	No Comments on any posts	70-80 on each post
Nameen Saxena (M)	Celebrities	944	No	No	Photo Update Posts Links	6-8 on each post	1-2 likes on each post
Chella Kapoor (F)	Celebrities	751	No	Job Uni Location	Video Shares	10-12 on each post	200+ likes on each post
Saarik Shah (M)	Landscapes	730	Yes	No	Status Updates – Personal thoughts	200-250 on each post	60-70 on each post
Abha Madan (F)	Artwork	671	Yes	No	Status Updates –	6-8 on each post	1-2 likes on

					Personal thoughts		each post
Tanea Bajwa (F)	Celebrities	556	Yes	Location	Video Shares	No comments on each post	200+ likes on each post
Tipu Ray (M)	Selfies	93	No	No	Video Shares	6-8 on each post	80-90 on each post
Eshana Bhasin (F)	Selfies Groups	86	No	No	Photo Update Posts Links	1-2 on each post	90-100 on each post
Umen Rao (M)	Selfies Groups	46	No	No	Photo Update Posts Links	12-15 on each post	No likes on any posts
Bhavika Chandra (F)	Selfies	21	Yes	Location Job Uni	Status Updates – Personal thoughts	6-8 on each post	200+ likes on each post
Apsara Dayal (F)	Selfies Groups	66	Yes	No	Status Updates – Personal thoughts	No Comments on any posts	60-80 on each post
Din Singh (M)	Selfies	84	Yes	No	Video Shares	6-8 on each post	No likes on any posts

Anjuli Goswama (F)	Selfies	99	Yes	No	Photo Update Posts Links Check-ins	90-100 on each post	200+ likes on each post
Kirani Sharma (M)	Selfies Groups	86	Yes	Location Relationship – Single	Video Shares	300-350 on each post	1-2 likes on each post
Ajaala Goel (F)	Selfies Pets	884	No	No	Video Shares	250-300 on each post	90-100 on each post
Amoli Marrick (F)	Selfies Groups	964	No	No	Status Updates – Personal thoughts	6-8 on each post	200+ likes on each post
Valin Seth (M)	Selfies Groups	562	No	No	Photo Update Posts Links	1-2 on each post	1-2 likes on each post
Lalika Kumar (F)	Selfies	840	No	Job	Status Updates – Personal thoughts	No Comments on any posts	No likes on each post
Danverr Puri (M)	Celebrities	934	No	Location Relationship – Single	Video Shares	1-2 on each post	70-80 on each post
Bhanu Puri (M)	Selfies	77	No	Job Relationship – Single	Video Shares	1-2 on each post	80-90 on each post

Sagara Mani (F)	Celebrities	67	Yes	Job Relationship – Married	Photo Update Posts Links	No Comments on any posts	1-2 likes on each post
Babar Modi (M)	Artwork	940	No	Job	Photo Update Posts Links	No Comments on any posts	1-2 likes on each post
Hanita Gill (F)	Artwork	673	Yes	No	Photo Update Posts Links	150-200 on each post	1-2 likes on each post
Amul Thakur (M)	Selfies	3	No	Job Relationship – Single	Photo Update Posts Links Check-ins	100-150 on each post	No likes on any posts
Chintak Batra (M)	Selfies Groups Pets	507	Yes	No	Status Updates – Personal thoughts	No comments on any post	1-2 likes on each post

Appendix V – Original 'RJ' profile from Studies 2 and 3.



Zimbabwe born, bred, and living.

Intro

Self-employed

Studied at University of Zimbabwe

Lives in **Bulawayo, Zimbabwe**

Photos 73

Friends

Privacy · Terms · Advertising · Ad choices · Cookies · More · Facebook © 2020

busy day, but time to relax finally

1 · 8 comments

View 5 more comments

Worth it though · 1

replied · 2 replies

looking forward to going to graduation for my friends. celebrations all around. if you want to come click on this.

UNIVERSITY OF ZIMBABWE

2020 GRADUATION CEREMONY

1. DATE/VENUE/TIME

The Ceremony will be held on Friday 4 December 2020 on campus in the UZ Great Hall, starting at 0900 hours.

2. PROCEDURE

Due to the Covid-19 pandemic, the Ceremony will be blended (both physical and virtual).

The following groups are invited to attend the Ceremony and be physically capped:

- All MPhil and DPhil Graduands

DRIVE.GOOGLE.COM

UZ ADVERT fin sat.pdf

2 · 6 comments

View 4 more comments

Sounds fun! · 1

replied · 1 reply

nature is my friend

1 · 7 comments

View 5 more comments

Nature is truly beautiful - we don't appreciate it enough! · 1

replied · 1 reply

fun day out in the great outdoors

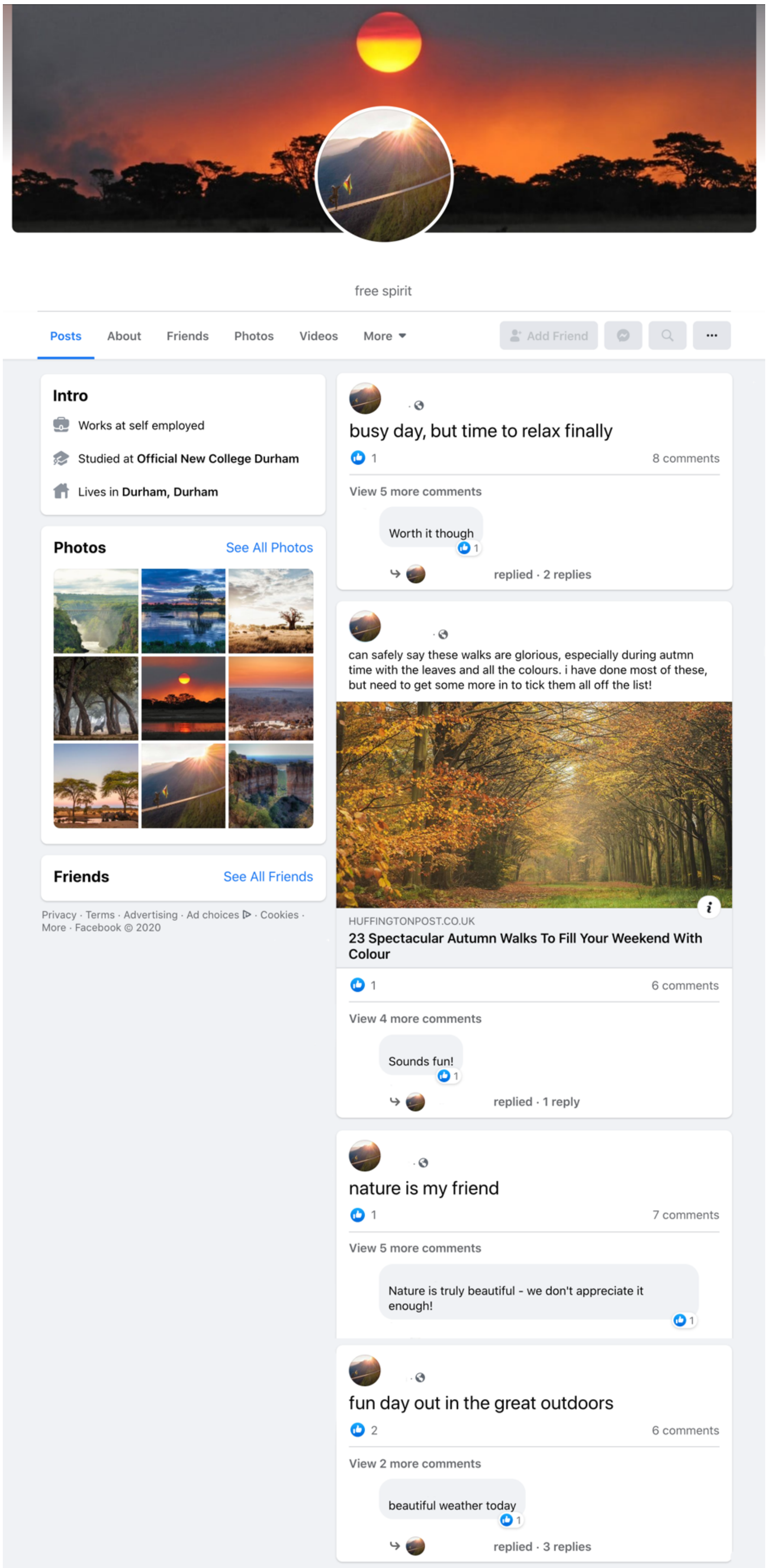
2 · 6 comments

View 2 more comments

beautiful weather today · 1

replied · 3 replies

Appendix W – Edited 'RJ' profile for Study 5.



free spirit

Intro

Works at self employed

Studied at **Official New College Durham**

Lives in **Durham, Durham**

Photos See All Photos

Friends See All Friends

Privacy · Terms · Advertising · Ad choices · Cookies · More · Facebook © 2020

busy day, but time to relax finally

1 — 8 comments

View 5 more comments

Worth it though

1

replied · 2 replies

can safely say these walks are glorious, especially during autmn time with the leaves and all the colours. i have done most of these, but need to get some more in to tick them all off the list!

HUFFINGTONPOST.CO.UK

23 Spectacular Autumn Walks To Fill Your Weekend With Colour

1 — 6 comments

View 4 more comments

Sounds fun!

1

replied · 1 reply

nature is my friend

1 — 7 comments

View 5 more comments

Nature is truly beautiful - we don't appreciate it enough!

1

fun day out in the great outdoors

2 — 6 comments

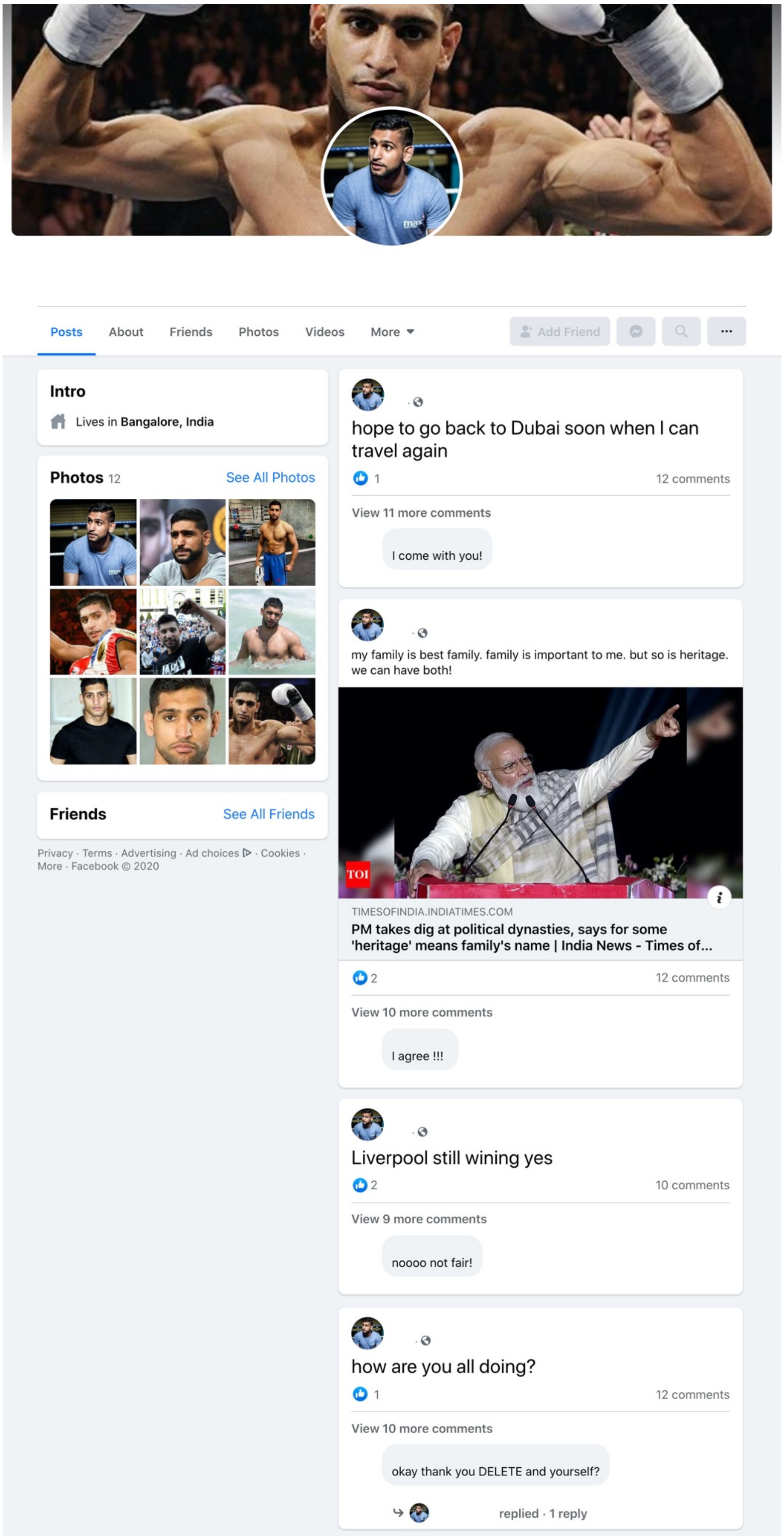
View 2 more comments

beautiful weather today

1

replied · 3 replies

Appendix X – Original 'MI' profile from Studies 2 and 3.



Intro

Lives in Bangalore, India

Photos 12

Friends

Privacy · Terms · Advertising · Ad choices · Cookies · More · Facebook © 2020

hope to go back to Dubai soon when I can travel again

1 · 12 comments

View 11 more comments

I come with you!

my family is best family. family is important to me. but so is heritage. we can have both!

TIMESOFINDIA.INDIATIMES.COM

PM takes dig at political dynasties, says for some 'heritage' means family's name | India News - Times of...

2 · 12 comments

View 10 more comments

I agree !!!

Liverpool still wining yes

2 · 10 comments

View 9 more comments

noooo not fair!

how are you all doing?

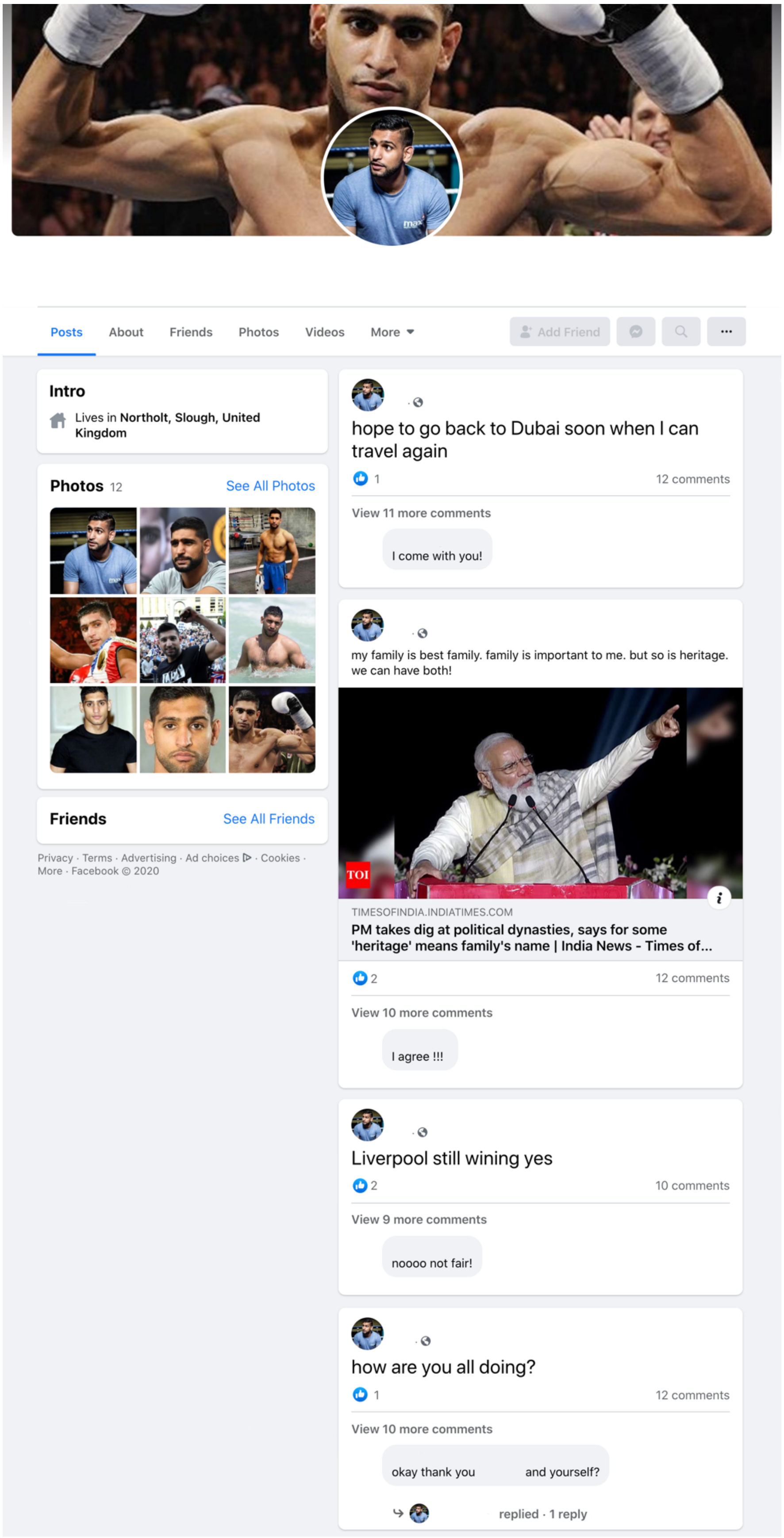
1 · 12 comments

View 10 more comments

okay thank you DELETE and yourself?

replied · 1 reply

Appendix Y – Edited 'MI' profile for Study 5.



Intro

Lives in **Northolt, Slough, United Kingdom**

Photos 12

Friends

Privacy · Terms · Advertising · Ad choices · Cookies · More · Facebook © 2020

hope to go back to Dubai soon when I can travel again

1 · 12 comments

View 11 more comments

I come with you!

my family is best family. family is important to me. but so is heritage. we can have both!

TIMESOFINDIA.INDIATIMES.COM

PM takes dig at political dynasties, says for some 'heritage' means family's name | India News - Times of...

2 · 12 comments

View 10 more comments

I agree !!!

Liverpool still wining yes

2 · 10 comments

View 9 more comments

noooo not fair!

how are you all doing?

1 · 12 comments

View 10 more comments

okay thank you and yourself?

replied · 1 reply

Appendix Z – Advert recruiting Indian participants to provide their real Facebook profiles in Study 5.

Do you use Facebook? Do you want £20? Are you an international student from **India** starting your 1st or 2nd year of study in October 2021? We could help each other!

I'm offering a £20 Amazon voucher to volunteers who can help with my PhD research on online deception. You'll be asked to provide a screenshot of your Facebook profile page, and consent to this screenshot being used in my study on fake social media profiles. Your profile will be randomly shown to participants, alongside several fake profiles, and participants will have to judge whether each profile they see is fake or real! All names on your real profile will be removed to ensure you remain anonymous.

For more detailed info, contact me on g.mckenzie@lancaster.ac.uk, or just message me on here, and I will provide further details and instructions, and answer any questions you may have. Thank you!

Appendix AA – Updated debrief form for participants in Study 4 which includes the updated total number of profiles viewed and the ratio of fake and real profiles.

Debrief – Study 4

Project title: Hiding in Plain Sight: A Turing Test on Fake Persona Spotting

Aim:

This study is concerned with fake profiles and identities on social media. Previous studies on fake social media profiles have researched the topic from a computer science aspect, resulting in the creation and development of algorithms and 'data bots' to identify and remove fake profiles from social media platforms. The primary aim of this study is to measure whether fake profiles on social media can be detected from a psychological point of view, through looking at and making a judgement of basic profile characteristics (photos, number of friends etc.) and the activities/posts within the profile. Additionally, this study aims to understand the cognitive processes employed when making profile authenticity judgements.

How was this tested?

You were required to view 16 Facebook profiles (4 real and 12 fake) and pass your judgement on each as to whether they were a real profile or a fake profile. You were also required to answer a series of questions in relation to the characteristics you looked at/used when making your judgement, and whether you believed there to be malicious intent/deceptive motivations behind each profile. You were required to judge these profiles under a time constraint. The purpose of the time constraint is to analyse the cognitive processes you used when making that judgement. The results from this study will help to contribute to the overall aim of this research by adding valuable information regarding how the judgements of social media profiles are actually made.

Hypotheses and main questions:

We are expecting to find the levels of judgement accuracy of the profiles to be around 50% or chance, in line with a large amount of previous research into deception detection; the bulk of the current literature states that humans perform no better than chance when detecting deception.

From the data you provided in the form of your judgements and answers, we are expecting to have the ability to develop a list of the characteristics present in a social media profile that make it identifiable as fake, and a better understanding of how humans make authenticity judgements in the online environment.

What is the purpose of this study?

To identify if fake social media profiles can be accurately detected by humans rather than computers. Ultimately, if it is found that accuracy levels are better than chance, this research could have a significant positive impact on security services/police forces, as it could help train them in what to look for when trying to identify fake profiles from terrorist groups or paedophiles.

What if I have a question or concern?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself or my supervisors using the contact details below:

Grace McKenzie (Primary Researcher)
PhD Student
g.mckenzie@lancaster.ac.uk

Professor Paul Taylor
p.j.taylor@lancaster.ac.uk
Department of Psychology
Lancaster University
Lancaster
LA1 4YF

Professor Stacey Conchie
s.conchie@lancaster.ac.uk
Department of Psychology
Lancaster University
Lancaster
LA1 4YF

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

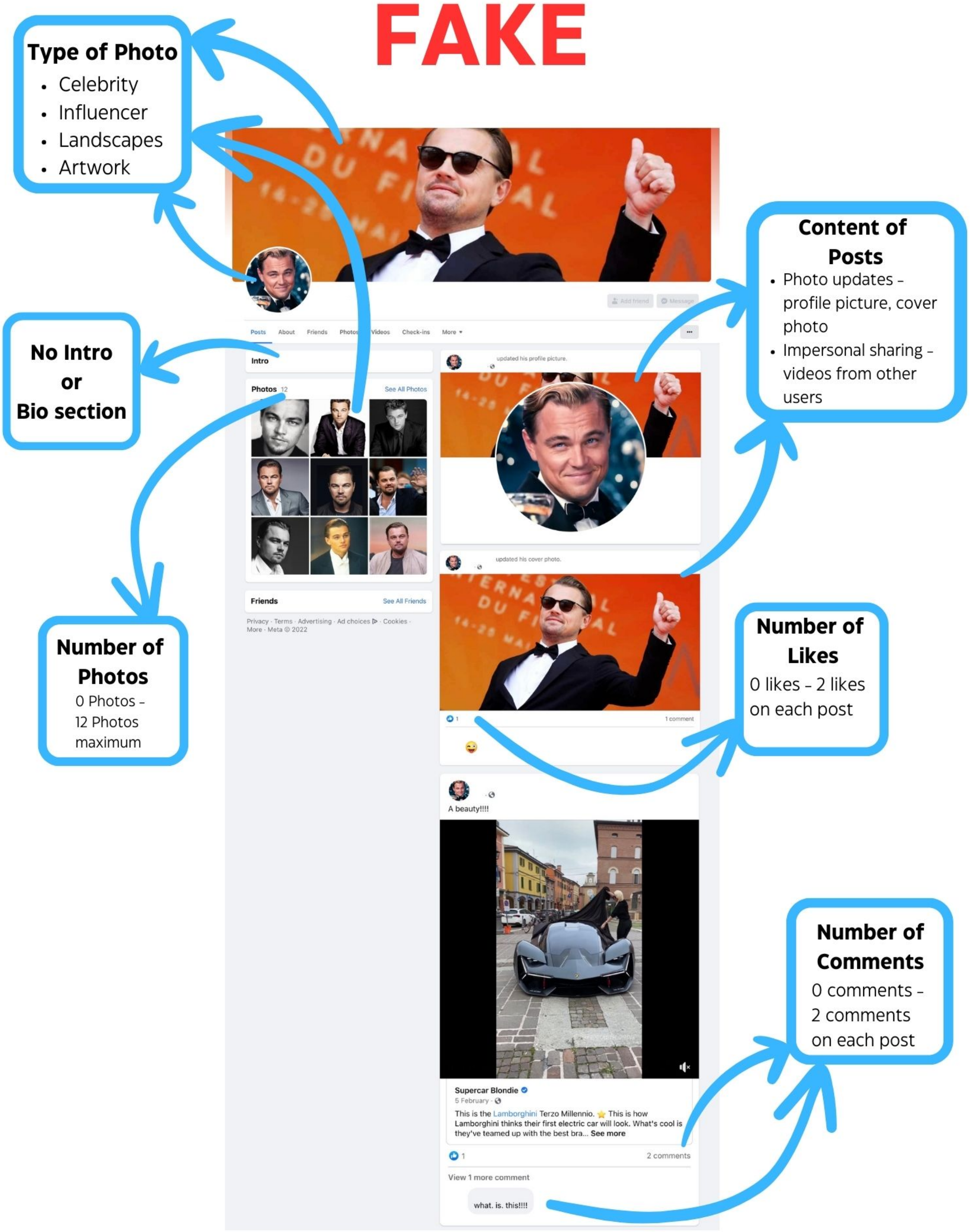
Professor Kate Cain
Head of Psychology Department
psychology.hod@lancaster.ac.uk
Department of Psychology
Lancaster University
Lancaster
LA1 4YF

You still have the right to withdraw from the research if you so wish. You will need to contact the researcher with your Prolific ID number to allow for identification and removal of your data.

Thank you for your time and participation.

Appendix AB – Training material for Study 6

FAKE



updated his profile picture.

updated his cover photo.

A beauty!!!!

Supercar Blondie

5 February

This is the Lamborghini Terzo Millennio. ⭐ This is how Lamborghini thinks their first electric car will look. What's cool is they've teamed up with the best bra... See more

what. is. this!!!!

Type of Photo

- Celebrity
- Influencer
- Landscapes
- Artwork

No Intro or Bio section

Number of Photos

0 Photos – 12 Photos maximum

Content of Posts

- Photo updates – profile picture, cover photo
- Impersonal sharing – videos from other users

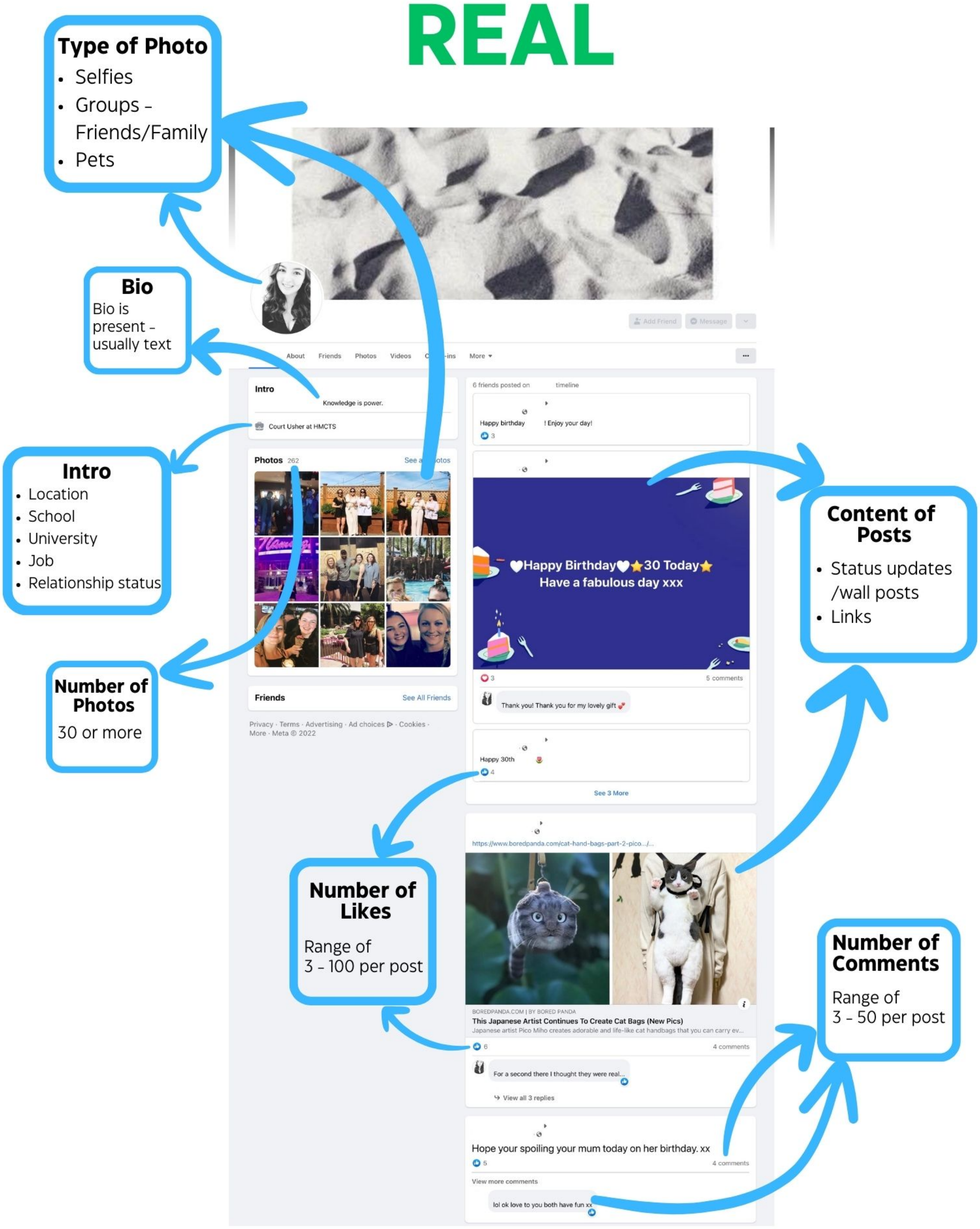
Number of Likes

0 likes – 2 likes on each post

Number of Comments

0 comments – 2 comments on each post

REAL



Type of Photo

- Selfies
- Groups – Friends/Family
- Pets

Bio

Bio is present – usually text

Intro

- Location
- School
- University
- Job
- Relationship status

Number of Photos

30 or more

Content of Posts

- Status updates /wall posts
- Links

Number of Likes

Range of 3 – 100 per post

Number of Comments

Range of 3 – 50 per post

Intro

Knowledge is power.

Court Usher at HMCTS

6 friends posted on timeline

Happy birthday ! Enjoy your day!

3

♡Happy Birthday♡⭐30 Today⭐
Have a fabulous day xxx

3 5 comments

Thank you! Thank you for my lovely gift

Happy 30th

4

See 3 More

https://www.boredpanda.com/cat-hand-bags-part-2-pico.../...

BOREDPANDA.COM | BY BORED PANDA

This Japanese Artist Continues To Create Cat Bags (New Pics)

Japanese artist Pico Miho creates adorable and life-like cat handbags that you can carry ev...

6 4 comments

For a second there I thought they were real...

View all 3 replies

Hope your spoiling your mum today on her birthday. xx

5 4 comments

View more comments

lol ok love to you both have fun xx

References

Abatecola, G., Mandarelli, G., & Poggesi, S. (2011). The personality factor: how top management teams make decisions. A literature review. *Journal of Management & Governance, 17,* 1073-1100. https://doi.org/10.1007/s10997-011-9189-y

Adikari, S., & Dutta, K. (2014, June). *Identifying fake profiles in LinkedIn* [Conference session]. Pacific Asia Conference on Information Systems (PACIS), Chengdu, China.

Ahmad, R., Wang, J., Hercegfi, K., & Komlodi, A. (2011). *Different people different styles: Impact of personality style in web sites credibility judgement* [paper presentation]. In M.J.Smith, & G. Salvendy (Eds.). Human Interface and the Management of Information, Lecture Botes in Computer Science. Springer. https://doi.org/10.1007/978-3-642-21793-7_59

Aiken, M. (2016). *The cyber effect. A pioneering cyberpsychologist explains how human behaviour changes online.* (1st ed.). London: John Murray.

Ajith, M., & Nirmala, M. (2022). Fake accounts and clone profiles identification on social media using machine learning algorithms. *International Journal of Scientific Research in Science, Engineering and Technology, 9*(3), 551-560. https://doi.org/10.32628/IJSRSET2293158

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning and Technology, 5,* 202-232. https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/c7949e6e-3d11-4a4e-8f6f-209aba7877bc/content

Algorithm. (n.d.). In *Cambridge Dictionary*. Retrieved 9 June 2020, from https://dictionary.cambridge.org/dictionary/english/algorithm

Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., & Alomari, D.M. (2023). Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining, 13*(20). https://doi.org/10.1007/s13278-022-01020-5

Allard, A., & Clavien, C. (2023). Nudging accurate scientific communication. *PLOS ONE, 18*(8). https://doi.org/10.1371/journal.pone.0290225

Alowibdi, J.S., Buy, U.A., Yu, P.S., Ghani, S. & Mokbel, M. (2014). *Detecting deception in online networks.* Symposium conducted at the IEEE/ACM International Conference on Computer Aided Design (ICAAD), China.

Alter, A., & Oppenheimer, D. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13,* 219-235. 10.1177/1088868309341564

Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgements of sexual orientation from thin slices of behaviour. *Journal of Personality and Social Psychology, 77*, 538-547. https://doi.org/10.1037/0022-3514.77.3.538

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin, 111,* 256-274. https://doi.org/10.1037/0033-2909.111.2.256

Ambady, N., & Rosenthal, R. (1993). Half a minute: predicting teacher evaluations from thin slices of nonverbal behaviour and psychical attractiveness. *Journal of Personality and Social Psychology, 64*, 431-441. https://www.sciencedirect.com/science/article/pii/S009265660700013X#aep-bibliography-id54

Amblee, N., & Bui, T. (2011). Harnessing the influence of social proof in online shopping: the effect of electronic word of mouth on sales of digital microproducts. *International Journal of Electronic Commerce, 16*, 91-114. doi: 10.2753/JEC1086-4415160205

American Psychological Association. (n.d.). Inoculation Theory. In *APA dictionary of psychology.* Retrieved March 26, 2024, from https://dictionary.apa.org/inoculation-theory

American Psychological Association. (n.d.). Speed Accuracy Tradeoff. In *APA dictionary of psychology.* Retrieved March 20, 2024, from https://dictionary.apa.org/speed-accuracy-tradeoff

Andreassen, C.S., Torsheim, T., Brunborg, G.S., & Pallesen, S. (2012). Development of a Facebook addiction scale. *Psychological Reports, 110*(2), 501-517. https://doi.org/10.1037/t74607-000

Andrew Hutchinson (2020, May 21). *Facebook adds new 'profile lock' option for users in India, simplifying data and content protection.* Social Media Today. www.socialmediatoday.com/news/facebook-adds-new-profile-lock-option-for-users-in-india-simplifying-dat/578472/

Antheunis, M.L., Valkenburg, P.M., & Peter, J. (2010). Getting acquainted through social network sites: Testing a model of online uncertainty reduction and social attraction. *Computers in Human Behavior, 26*(1)*,* 100-109. https://doi.org/10.1016/j.chb.2009.07.005

Apte, M. (1994). Language in sociocultural context. In R.E. Asher (Ed.), *The Encyclopedia of Language and Logistics, 4,* 2000-2010. Pergamon Press.

Arechar, A.A., Allen, J., Berinsky, A.J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J.G., Ross, R.M., Stagnaro, M.N., Zhang, Y., Pennycook, G., & Rand, D.G. (2022). Understanding and reducing misinformation across 16 countries on six continents. *PsyArXiv,* 1-48. https://doi.org/10.31234/OSF.IO/A9FRZ

Aslan, Salman. (2021, March 16). *80+ Facebook statistics you need to know in 2020.* Omnicore. https://www.omnicoreagency.com/facebook-statistics/

Bajwa, R.S., Batool, I., Asma, M., Ali. H, & Ajmal, A. (2016). Personality traits and decision making style among university students (Pakistan). *Pakistan Journal of Life and Social Sciences, 14*(1), 38-41.https://www.researchgate.net/publication/304944238_Personality_traits_and_decision_making_styles_among_university_students_Pakistan

Ballew, C.C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgements. *Proceedings of the National Academy of Sciences, 104*(46), 17,948-17,953. https://doi.org/10.1073/pnas.0705435104

Baron, J. & Gürçay, B. (2016). A meta-analysis of repsonse-time tests of the sequential two-systems model of moral judgement. *Memory & Cognition, 45*, 566-575. https://doi.org/10.3578/s13421-016-0686-8

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about Bad News: Gamified inoculation boosts confidence and cognitive immunity against fake news, *Journal of Cognition, 3*(1). doi: 10.5334/joc.91.

Bate S., Bennetts, R., Hasshim, N., Portch, E., Murray, E., Burns, E., & Dudfield, G. (2019). The limits of super recognition: An other-ethnicity effect in individuals with extraordinary face recognition skills. *Journal of Experimental Psychology: Human Perception and Performance 45*(3), 363-377. https://doi.org/10.1037/xhp0000607

Bate, S., Bennetts, R., Murray, E., & Portch, E. (2020). Enhanced matching of children's faces in "super recognisers" but not high-contact controls. *Perception, 11*(4). https://doi.org/10.1177/2041669520944420

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: 10.18637/jss.v067.i01.

Bayes, T. 1763. "An Essay Toward Solving a Problem in the Doctrine of Chances", *Philosophical Transactions of the Royal Society of London* 53, 370-418. https://doi.org/10.1098/rstl.1763.0053

Beer, A. (2019). *Information as a moderator of accuracy in personality judgement*. In T.D. Letzring, & J.S.Spain (Eds). The Oxford Handbook of Accurate Personality Judgement. https://doi.org/10.1093/oxfordhb/9780190912529.013.9

Berg, M., Mwambali, S.N. & Bogren, M. (2022). Implementation of a three-pillar training intervention to improve maternal and neonatal healthcare in the Democratic Republic of Congo: a process evaluation study in an urban health zone. *Global Health Action, 15*(1). 10.1080/16549716.2021.2019391

Berry, J.W. (1966). Temne and Eskimo perceptual skills. *International Journal of Psychology, 1,* 207-229. https://doi.org/10.1080/00207596608247156

Beymer, D., Orton, P.Z., & Russell, D.M. (2007). An eye tracking study of how pictures influence online reading. In C. Baranauskas, P. Palanque, J. Abascal, & S.D.J. Barbosa (Eds.), *Human-Computer Interaction – INTERACT 2007. INTERACT 2007. Lecture Notes in Computer Science* (Vol. 4663). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74800-7_41

Bilge, L., Strufe, T., Blzarotti, D., & Kirda, E. (2009). *All your contacts belong to us: automated identitiy theft attacks on social networks [*paper presentation]. Proceedings of the 18th International Conference on the World Wide Web, April 2009, Madrid, Spain.

Blackman, M.C., & Funder, D.C. (1998). The effect of information consensus and accuracy in personality judgement. *Journal of Experimental Social Psychology, 34*, 164-181. https://doi.org/10.1006/jesp.1997.1347

Blair, I.V., Judd, C.M., Chapleau, K.M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15*(10). https://doi.org/10.1111/j.0956-7976.2004.00739.x

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10, 214-234.

Bonner, S.E. (2008). Judgement and decision making in accounting. *Accounting Horizons, 13*(4), 385-398.

Boshmaf, Y., Muslukhov, I, Beznosov, K., & Ripeanu, M. (2011). *The socialbot network: when bots soicalize for fame and money* [paper presentation]. Proceedings of the 27th Annual Computer Security Applications Conference, Orlando, FL, USA. 93-102. https://doi.org/10.1145/2076732.2076746

Bowman, C.H., Evans, C.E.Y., & Turnbull, O.H. (2005). Artificial time constraints on the Iowa Gambling Task: the effects on behavioural performance and subjective experience. *Brain and Cognition, 57*(1), 21-25. https://doi.org/10.1016/j.bandc.2004.08.015

Brunken, R., Plass, J.L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning using dual-task methodology: Auditory load and modality effects. *Instructional Science, 32,* 115-132. https://doi.org/10.1023/B:TRUC.0000021812.96911.c5

Brunswik, E. (1956*). Perception and the representative design of psychological experiments.* University of California Press.

Bühler, J.L., Orth, U., Bleidorn, W., Weber, E., Kretzschmar, A., Scheling, L., & Hopwood, C.J. (2023). Life events and personality change: A systematic review and meta-analysis. *European Journal of Personality, 0*(0). https://doi.org/10.1177/08902070231190219

Byrne, K.A., Silasi-Mansat, C.D., & Worth, D.A. (2015). Who choke sunder pressue? The big five personality traits and decision-making under pressue. *Personality and Individual Differences, 74*, 22-28. 10.1016/j.paid.2014.10.009

Campbell, D.T., & Staley, J.C. (1963). *Experimental and quasi-experimental designs for research.* Rand McNally.

Carney, D.R., & Harrigan, J.A. (2003). It takes one to know one: interpersonal sensitivity is related to accurate assessments of others' interpersonal sensitivity. *Emotion 3*(2), 194-200. https://doi.org/10.1037/1528-3542.3.2.194

Carney, D. R., Randall Colvin, C., & Hall, J.A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality, 41,* 1054-1072. DOI: 10.1016/j.jrp.2007.01.004

Caspi, A., & Gorsky, P. (2006). Online deception: prevalence, motivation, and emotion. *CyberPsychology & Behavior, 9*(1), 54-59. http://doi.org/10.1089/cpb.2006.9.54

Cella, M., Dymond, S., Cooper, A., & Turnbull, O. (2007). Effects of decision-phase time constraints on emotion-based learning in the Iowa Gambling Task. *Brain and Cognition, 64*(2), 164-169. https://doi.org/10.1016/j.bandc.2007.02.003

Chaffey, D. (2020). Global social media research summary 2020. Retrieved 8 June 2020, from https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/#:~:text=More%20than%204.5%20billion%20people,since%20this%20time%20last%20year.

Chavoshi, N., Hamooni, H., & Mueen, A. (2016). Identifying correlated bots in Twitter. In: Spiro E., Ahn YY. (eds) *Social Informatics*. SocInfo 2016: Lecture Notes in Computer Science, vol 10047. Springer, Cham

Chegeni, M., Shahrbabaki, P.M., Shahrbabki, M.E., Nakhaee, N., & Haghdoost, A. (2021). Why people are becoming addicted to social media: A qualitative study. *Journal of Education and Health Promotion, 10*(1), 175. 10.4103/jehp.jehp_1109_20

Cheng, J. (n.d.). *Ten-Item Personality Inventory Scoring Syntax SPSS* [Computer software]. http://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). Who is tweeting on Twitter: human, bot, or cyborg? [paper presentation]. Proceedings of the 26th Annual Computer Security Applications Conference, New York, NY, USA. 21-30. https://doi.org/10.1145/1920261.1920265

Conger, K. (2022, August 4). *Musk says Twitter committed fraud in dispute over fake accounts.* The New York Times. https://www.nytimes.com/2022/08/04/technology/musk-twitter-fraud.html

Congiu, L., & Moscati, I. (2021). A review of nudges: Definitions, justifications, effectiveness. *Journal of Economic Surveys, 36,* 188—213. https://doi.org/10.1111/joes.12453

Cornwell, B., & Lundgren, D. C. (2001). Love on the Internet: Involvement and misrepresentation in romantic relationships in cyberspace vs. realspace. *Computers in Human Behavior, 17*(2), 197–211. https://doi.org/10.1016/S0747-5632(00)00040-6

Cotte, J., Chowdury, T.G., Ratneshwar, S., & Ricci, L.M. (2006). Pleasure or utility? Time planning style and web usage behaviours. *Journal of Interactive Marketing, 20*(1), 45-57. https://doi.org/10.1002/dir.20055

Cipresso, P., & Riva, G. (2016). Personality assessment in ecological settings by means of virtual reality. In U. Kumar (Ed.) *The Wiley Handbook of Personality Assessment,* (pp. 240-248). *John Wiley & Sons, Ltd.* https://doi.org/10.1002/9781119173489.fmatter

Cummins, D.D., & Cummins, R.C. (2012). Emotion and deliberative reasoning in moral judgement. *Frontiers in Psychology, 3(328),* 1-16. Doi: 10.3389/fpsyg.2012.00328

Darbyshire, D., Kirk, C., Wall, H.J., & Kaye, L.K. (2016). Don't judge a (Face)Book by its cover: Exploring judgement accuracy of other' personality on Facebook. *Computers in Human Behaviour, 58*, 380-387. http://dx.doi.org/10.1016/j.chb.2016.01.021

Davdison, B.I., & Joinson, A.N. (2021). Shpae shifting across social media. *Social Media + Society, 7*(1). https://doi.org/10.1177/2056305121990632open_in_new

de Castro Bellini-Leite, S. (2013). The embodied embedded character of system 1 processing. *Mens Sana Monographs, 11*(1), 239-252. Doi: 10.4103/0973-1229.109345.

DeDonno, M.A., & Demaree, H.A. (2008). PErcieved time pressure and the Iowa Gambling Task. *Judgement and Decision Making, 3*(8), 636-640. https://doi.org/10.1017/S1930297500001583

Dennis, A.R., & Minas, R.K. (2018). Security on autopilot: Why current security theories hijack our thinking and leas us astray. *The DATA BASE for Advances in Information Systems, 49*, 15-37. https://doi.org/10.1145/3210530.321.0533

DePaulo, B.M., Lindsay, J.J., Malone, B.E., Muhlenbruck, L., Charlton, K. & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118.

Donath, J.S. (1995). Identity and deception in the virtual community. In P. Kollock & M. Smith (Eds.), *Communities in cyberspace.*

Dou, X. (2011). The influence of cultures on SNS usage: comparing Mixi in Japan and Facebook in the US. *Public Relations Journal, 5*(4). Retrieved from https://prjournal.instituteforpr.org/wp-content/uploads/2011Dou.pdf

Driskell, J.E. (2011). Effectiveness of deception detection training: a meta-analysis. *Psychology, Crime & Law, 18*(8), 713-731. https://doi.org/10.1080/1068316X.2010.535820

Driskell, T., Sclafani, S., & Driskell, J.E. (2014). Reducing the effects of game day pressures through stress exposure training. *Journal of Sport Psychology in Action, 5*(1), 28-43. https://doi.org/10.1080/21520704.2013.866603

Drouin, M., Miller, D., Wehle, S.M.J., Hernandez, E. (2016). Why do people lie online? "Because everyone lies on the internet". *Computers in Human Behavior, 64,* 134-142. https://doi.org/10.1016/j.chb.2016.06.052

Duchaine, B., & Nakayama, K. (2006a). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted performance and prosopagnosic subjects. *Neuropsychologia, 44*(4)*,* 576-585.

Dunaway, J., Searles, K., Sui, M., & Paul, N. (2018). News attention in a mobile era. *Journal of Computer-Mediated Communication, 23,* 107-124. https://doi.org/10.1093/jcmc/zmy004

Dunn, J.D., Towler, A., Kemp, R.I., & White, D. (2023). Selecting police super-recognisers. *PLoS ONE, 18*(5). https://doi.org/10.1371/journal.pone.0283682

Dyson, M., & Haselgrove, M. (2000). The effects of reading speed and reading patterns on the understanding of text read from a screen. *Journal of Research in Reading, 23*(2), 210-223. DOI:10.1111/1467-9817.00115

Eagleton, T. (2000). *The idea of culture.* Blackwell Publishing.

Eagly, A.H., & Chaiken, S. (1993). *The psychology of attitudes.* Harcourt Brace. Retrieved from www.psycnet.apa.org/record/1992-98849-000

Edwards, C., Stoll, B., Faculak, N., & Karman, S. (2015). Social presence on LinkedIn: Percieved credibility and interpersonal attractiveness based on user profile picture. *Online Journal of Communication and Media Technologies, 5*(4), 102-115. https://doi.org/10.29333/ojcmt/2528

Ekman, P. (1972). Universal and cultural differences in facial expressions of emotion. In J.K. Cole (Ed.), *Nebraska Symposium on Motivation,* 207-283. University of Nebraska Press.

Ekman, P. (2009). *Telling lies: Clues to deceit in the marketplace, politics and marriage.* New York: NY, W.W. Norton & Company.

Elaad, E. (2003). Effects of feedback on the overestimated capacity to detect lies and the underestimated ability to detect lies. *Applied Cogntive Pscyhology, 17,* 349-363. https://doi.org/10.1002/acp.871

Elfenbein, H.A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128,* 203-235. https://doi.org/10.1037/0033-2909.128.2.203

Elyusufi, Y., Elyusufi, Z., & Kbir, M.A. (2020). *Social networks fake profiles detection usinf machine learning algorithms* [paper presentation]. In: M Ben Ahmed, A. Boudhir, D. Santos, M. El Aroussi, I. Karas. (Eds.). Proceedings of the International Conference on Smart City Applications, Casablanca, Morocco. 30-40. https://doi.org/10.1007/978-3-030-37629-1_3

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: the role of personal photos in Airbnb. *Tourism Management, 55,* 62-73. https://doi.org/10.1016/j.tourman.2016.01.013

Fatehi, K., Priestly, J.L., & Taasoobshirazi, G. (2020). The expanded view of individualism and collectivism: One, two, or four dimensions? *International Journal of Cross Cultural Management, 20*(1), 7-24. https://doi.org/10.1177/1470595820913077

Faul, F., Erdfelder, E., Lang, A.G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fazio, L.K., Brashier, N.M., Payne, B.K., & Marsh, E.J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General, 144,* 993-1002. Doi: 10.1037/xge0000098

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96-104. https://doi.org/10.1145/2818717

Ferrera, E., Wang, WQ., Varol, O., Flammini, A. & Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. In: Spiro E., Ahn YY. (eds) *Social Informatics*. SocInfo 2016: Lecture Notes in Computer Science, vol 10047. Springer, Cham

Fischer, R., & Derham, C. (2022). Is in-group bias culture-dependent? A meta-analysis across 18 societies. *Springerplus, 5*, 70. 10.1186/s40064-015-1663-6

Frank, M.G., & Feeley, T.H. (2003). To catch a liar: Challenges for research in lie detection training. *Journal of Applied Communication Research, 31*(1), 58-75. https://doi.org/10.1080/00909880305377

Fred Tabsharani (n.d.). *Support vector machine (SVM).* Tech Target. https://www.techtarget.com/whatis/definition/support-vector-machine-

SVM#:~:text=A%20support%20vector%20machine%20(SVM)%20is%20a%20type%20of%20supervised,data%20set%20into%20two%20groups.

Fu, G., Lee, K., Cameron, C.A., & Xu, F. (2001). Chinese and Canadian asults' categorisation and evaluation of lie- and truth-telling about prosocial and antisocial behaviours. *Journal of Cross-Cultural Psychology, 32*(6), 740-747. https://doi.org/10.1177/0022022101032006005

Funder, D.C. (1995). On the accuracy of personality judgement: A realistic approach. *Psychological Review, 102*(4), 652-670. https://doi.org/10.1037/0033-295X.102.4.652

Funder, D.C. (1999). *Personality judgment; A realistic approach to person perception.* Academic Press.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter? [Conference session]. ACM SIGMETRICS/IFIP Performance, Antibes Juan-les-Pins, France. https://inria.hal.science/hal-01281190

Gass, R. H., & Seiter, J. S. (2014) *Persuasion: Social influence and compliance gaining* (5th ed.). Pearson.

Gelfand, M.J.(2018). *Rule makers, rule breakers: How culture wires our minds, shpaes our nations, and drives our differences.* Robinson. DOI: 10.1126/science.1197754

Gelfand, M.J, Nishii, L.H., & Raver, J.L. (2006). On the nature and importance of cultural tightness-looseness. *Journal of Applied Psychology, 91*(6), 1225–1224. https://doi.org/10.1037/0021-9010.91.6.1225

Gelfand, M.J., Raver, J., Nishii, L., Leslie, L.M., Lun, J., Lim, B.C., … & Aycan, Z. (2011). Differences between tight and loose cultures: A 33-nation study. *Science, 332*(6033), 1100-1104.

George, J.F., Marett, K., Crews, J., Cao, J., Lin, M., Biros, D.P., & Burgoon, J.K. (2004). *Training to detect deception: an experimental investigation* [paper presentation]. 37th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA. 10.1109/HICSS.2004.1265082

Getov, S., Kanai, R., Bahrami, B., & Rees, G. (2015). Human brain structure predicts individual differences in preconscious evaluation of facial dominance and trustworthiness. *Social Cognitive and Affective Neuroscience*, *10*(5), 690-699. https://doi.org/10.1093/scan/nsu103

Griffiths, M.D. (2010). The role of context in online gaming excess and addicition: Some case study evidence. *International Journal of Mental Health and Addiction, 8,* 119-125. https://doi.org/10.3109/16066350009005587

Gupta, A., & Kaushal, R. (2017, January 29 – February 1). Towards detecting fake user accounts in Facebook [Conference session]. ISE Asia Security and Privacy (ISEAP), Surat, India. https://doi.org/10.1109/ISEASP.2017.7976996

Haile, T. (2014). *What you think you know about the web is wrong.* Time. Retrieved March 14, 2024, from (https://time.com/12933/what-you-think-you-know-about-the-web-is-wrong/#:~:text=If%20you%27re%20an,We%20mistake%20sharing%20for%20reading.)

Hall, J. A., & Andrzejewski, S. A. (2008). Accuracy of interpersonal perception and social sensitivity: Revisiting concepts, methods, and findings. *Personality and Social Psychology Review, 12*(3), 349–363. https://doi.org/10.1177/1088868308316890

Hanley, J., Herron, C., & Cole, S. (1995). Using video as advance organizer to a written passage in the FLES classroom. *The Modern Language Journal, 79*, 57-66. https://doi.org/10.2307/329393

Hauch, V., Sporer, S.L., Michael, S.W., & Meissner, C.A. (2016). Does training improve the detection of deception? A meta-analysis. *Communication Research, 43*(3), 283-343. DOI: 10.1177/0093650214534974.

Hinds, J., & Joinson, A. (2019). Human and computer personality prediction from digital footprints. *Current Directions in Psychological Science, 28*(2), 204-211. https://doi.org/10.1177/0963721419827849

Hinkle, S., & Brown, R.J. (1990). Intergroup comparisons and social identity: some links and lacunae. In: D. Abrams, M.A. Hogg (Eds.), *Social identity theory: constructive and critical advances,* 48-70. Harvester-Wheatsheaf.

Hofstede, G. (1980). *Culture's Consequences: International differences in work-related values.* Sage.

Hofstede, G. (1991). Empirical models of cultural differences. In N. Bleichrodt & P.J.D. Drenth (Eds.), *Contemporary issues in cross-cultural psychology* (pp.4-20). Swets & Zeitlinger Publishers.

Hopkins, N., & Reicher, S. (2011). Identity, culture and contestation: Social identity as cross-cultural theory. *Pscyhological Studies, 56*, 36-42. https://doi.org/10.1007/s12646-011-0068-z

Huang, C-M., & Park, D. (2013). Cultural influences on Facebook photographs. *International Journal of Psychology, 48*(3), 334-343, DOI: 10.1080/00207594.2011.649285

Humphreys, L. (2018). *The qualified self: Social media and the accounting of everyday life.* MIT Press.

Ivcevic, Z., & Ambady, N. (2012). Personality impression from identity claims on Facebook. *Psychology of Popular Media Culture, 1*(1), 38-45. https://doi.org/10.1037/a0027329

Izard, C.E. (1971). *The face of emotion.* Appelton-Centrury-Crofts.

Jackson, P.C. Jr. (2018). Thoughts on bands of action [Conference session]. Annual International Conference on Biologically Inspired Cognitive Architectures (BICA). Prague, Czech Republic.

James Chen (2024, July 28). *What is a neural network?* Investopedia. https://www.investopedia.com/terms/n/neuralnetwork.asp

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). Human or not? A gamified approach to the Turing test. *arXiv,* 10.48550/arXiv.2305.20010

Jarecki, A., Schulman, A., Schmidt, E., … Maroney, J. (Executive producers) (2012 - present). *Catfish: The Tv Show* [TV series]. Catfish Picture Company Relativity Media: MTV Entertainment Studios.

Johnson, T.J., & Kaye, B.K. (2015). Reasons to believe: Influence of credibility on motivations for using social networks. *Computers in Human Behavior, 50,* 544-555. https://doi.org/10.1016/j.chb.2015.04.002

Kagan, D., Elovichi, Y., & Fire, M. (2018). Generic anomalous vertices detection utilizing a link prediction algorithm. *Social Network Analysis and Mining, 8*(27). https://doi.org/10.1007/s13278-018-0503-4

Kahneman, D. (2003) Maps of bounded rationality: Psychology for behavioural economics. *The American Economic Review, 93*(5), 1449-1475. https://www.jstor.org/stable/3132137

Kahneman, D. (2011). *Thinking, fast and slow.* Penguin Random House, UK.

Kaplan, M.F., Wanshula, L.T., & Zanna, M.P. (1993). Time pressure and information integration in social judgement . In O.Svenson & A.J. Maule, *Time pressure and stress in human judgement and decision making,* 255-267. Plenum.

Karelaia, N., & Hogarth, R.M. (2008). Determinants of linear judgements: A meta-analysis of lens model studies. *Psychological Bulletin, 134*(3), 404-426. doi: 10.1037/0033-2909.134.3.404.

Kashima, Y., & Gelfand, M.J. (2012). A history of culture in psychology. In A.W. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology,* 499-520. Psychology Press.

Kegelaers, J., Wylleman, P., Bunigh, A. & Oudejans, R.D., (2021) A mixed methods evaluation of a pressure training intervention to develop resilience in female basketball players. *Journal of Applied Sport Psychology,* 33(2), 151-172. https://doi.org/10.1080/10413200.2019.1630864

Kenny, R., Fischhoff, D., Davis, A., Carley, K.M., & Canfield, C. (2024). Duped by bots: Why some are better than others at detecting fake social media personas. *Human Factors, 66*(1), 88-102. https://doi.org/10.1177/00187208211072642

Kessler, L., & Ashton, R.H. (1981). Feedback and prediction achievement in financial analysis. *Journal of Accounting Research, 19*(1), 146-162. https://doi.org/10.2307/2490966

Khairullah, D.H., & Khairullah, Z.Y. (2013). Cultural values and decision-making in China. *International Journal of Business, Humanities and Technology, 3*(2), 1-12.

Kim, E.S., & Willson, V.L. (2010). Evaluation pretest effects in pre-post studies. *Educational and Psychological Measurement, 70*(5), 744-759. doi: 10.1177/001316441066687.

Kitajima, M., & Toyota, M. (2011). Four processing modes of in situ human behaviour [Conference session]. Annual International Conference on Biologically Inspired Cognitive Architectures (BICA). Amsterdam, The Netherlands.

Kitajima, M., & Toyota, M. (2013). Decision-making and action selection in two minds: An analysis based on model human processor with realtime constraints (MHP/RT). *Biologically Inspire Cognitive Architectures, 5,* 82-93. https://doi.org/10.1016/j.bica.2013.05.003

Kluger, A.N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284. https://doi.org/10.1037/0033-2909.119.2.254

Köbis, N.C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: people cannot detect deepfakes but they think they can. *iScience, 24*(11) 10.1016/j.isci.2021.103364

Kocher, M.G., & Sutter, M. (2006). Time is money – Time pressure incentives, and the quality of decision making. *Journal of Economic Behaviour & Organisation, 61*(3), 375-392. https://doi.org/10.1016/j.jebo.2004.11.013

Krackow, E. (2010). Narratives distinguish experiences from imagined childhood events. *The American Journal of Psychology, 123,* 71-80.

Kroeber, A.L., & Kluckhohn, C. (1952). Culture: a critical review of concepts and definitions. *Papers. Peabody Museum f Archaeology & Ethnology, 47*(1). Retrieved from https://psycnet.apa.org/record/1953-07119-001

Krombholz, K., Merkl, D., & Weippl, E. (2012). Fake identities in social media: A case study on the sustainability of the Facebook business model. *Journal of Service Science Research, 4*, 175-212. doi: 10.1007/s12927-012-0008-z

Kruglanski, A.W., & Freund, T. (1983). The freezing andf unfreezing of lay-inferences: effects of impressional primacy, ethnic stereotyping, and numerical anchoring.

Journal of Experimental Social Psychology, 19, 448-468. https://doi.org/10.1016/0022-1031(83)90022-7

Kudugunta, S. & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences, 467*, 312-322. Retrieved 9 June 2020, from https://arxiv.org/pdf/1802.04289.pdf

Kumar, A., Beniwal, R., Jain, D. (2023). Personality detection using kernel-based ensemble model for leveraging social psychology in online networks. *Association for Computing Machinery Transactions on Asian and Low-Resource Language Information Processing.* https://doi.org/10.1145/3571584

Kumar, S., & Shah, N. (2018). False information on web and social media: A survey. *arXiv, 1*(1), 1-35. https://doi.org/10.48550/arXiv.1804.08559

Leakey, R.E., & Lewin, R. (1978). *People of the lake: Mankind and its beginnings.* Anchor Press.

Lee, K., Caverlee, J., & Webb, S. (2010, July 19-23). *Uncovering social spammers: social honeypots and machine learning* [Conference session]. Proceedings of the 33rd International Association for Computing Machinery SIGIR'10 conference on Research and Development in Information Retrieval, Geneva, Switzerland.

Lee, K., Lee, B., & Oh, W. (2015). Thumbs up, sales up? The contingent effect of Facebook likes on sales performance in social commerce. *Journal of Management Information Systems, 32*(4), 109-143. doi: 10.1080/07421222.2015.1138372

Lewandowsky, S., Ecker, U.K.H., Seifert, C.M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13,* 106-131. https://doi.org/10.1177/1529100612451018

Leviathan, Y., & Matias, Y. (2018, May 8). Google Duplex: An AI system for accomplishing real-world tasks over the phone. *Google Research.* https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Levine, T.R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378-392.

Levine, T.R. (2020). *Duped: Truth Default Theory and the social science of lying and deception.* University of Alabama Press.

Levine, T.R., Park, H.S., & McCornack, S.A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect." *Communication Monographs, 66,* 125-144.

Levine, T.R., Serota, K.B., Shulman, H., Clare, D.D. , Park, H.S., Shaw, A.S., Shim, J.C., & Lee, J.H. (2011). Sender demeanor: Individual differences in sender believability have a powerful impact on deception detection judgements. *Human Communication Research, 37*, 377-403. https://doi.org/10.1111/j.1468-2958.2011.01407.x

Li, Y., & Xie, Y. (2019). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research, 57*(1), 1-19. https://doi.org/10.1177/0022243719881113

Lindsay, D.S., Hagen, L., Don Read, J., Wade, K.A., & Garry, M. (2004). True photographs and false memories. *Psychological Science, 15*(3), 149-154. https://doi.org/10.1111/j.0956-7976.2004.01503002.x

Liu, X., Zhan, Y., Jin, H., Wang, Y., & Zhang, Y. Research on the classification methods of social bots. *Electronics*, *12*(4), 3030. https://doi.org/10.3390/electronics12143030).

Louth, S.M., Williamson, S., Alpert, M., Pouget, E.R. & Hare, R.D. (1998). Acoustic distinctions in the speech of male psychopaths. *Journal of Psycholinguistic Research, 27*(3), 375-384.

Lu, B., Fan, W., & Zhou, M. (2016). Social presence, trust, and social commerce purchase intention: an empirical research. *Computers in Human Behaviour, 56*, 225-237. https://doi.org/10.1016/j.chb.2015.11.057

Luo, M., Hancock, J.T., & Markowitz, D.M. (2020). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research, 49*(2). https://doi.org/10.1177/0093650220921321

Lutkevich, B., & Gillis, A.S. (2022, March). *Definition: Bot.* What Is. Retrieved January 24th, 2023, from https://www.techtarget.com/whatis/definition/bot-robot.

Machiavelli, N. (1532). *The Prince*. Antonio Blado d'Asola.

Maheswari, S., & Mukherjee, T. (2015). How does academic performance increase virtual popularity? A case study of Facebook usage among Indian college students. *Contemporary Educational Technology, 13*(1), 1-11. https://doi.org/10.30935/cedtech/8709

Manjunatha, S. (2013). The usage of social networking sites among the college students in India. *International research journal of social sciences, 2*(5), 15-21.

Marshall, R., Huan, T.C., Xu, Y., & Nam, I (2011). Extending prospect theory cross-culturaly by examining witching behaviour in consumer and business-to-business contexts. *Journal of Business Research, 64*(8), 871-878. 10.1016/j.jbusres.2010.09.009

Masthi, N.R.R., Cadabam, S.R., & Sonakshi, S. (2015). Facebook addiction among health university students in Bengalaru. *International Journal of Health & Allied Sciences, 4*(1), 10.4103/2278-344X.149234

Masuda, T., Wang, H., Ishii K., & Ito, K. (2012). Do surrounding figures' emptions affect judgement of the target figure's emotion? Comparing the eye-movement patterns of European Canadians, Asian Canadians, Asian international students, and Japanese. *Frontiers in Integrative Neuroscience, 6*(72). https://doi.org/10.3389/fnint.2012.00072

Matsumoto, D., Hwang, H., & Yamada, H. (2010). Cultural differences in the relative contributions of face and context to judgements of emotions. *Journal of Cross-Cultural Psychology, 43*(2), 198-218. https://doi.org/10.1177/0022022110387426

Maxwell, S.E., & Delaney, H.D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Psychology Press.

Mayer, R.E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Mayer, R.E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38* 43-52. https://data-lessons.github.io/instructor-training/files/papers/mayer-reduce-cognitive-load-2003.pdf

Mbona, I., & Eloff, J.H.P. (2023). Classifying social media bots as malicious or benign using semi-supervised machine learning. *Journal of Cybersecurity, 9*(1), 1-12. https://doi.org/10.1093/cybsec/tyac015

McKenzie, G. P. S. F. (2016). *Are psychopath's good liars?: A linguistic analysis of their deceptive speech.* [Unpublished master's thesis]. Goldsmiths, University of London.

Merriam-Webster (n.d.). *Merriam-Webster.com dictionar.* Retrieved March 17, 2024, from
https://www.merriam-webster.com/dictionary/hedonism#:~:text=1,suggesting%20the%20principles%20of%20hedonism

Meta (2024a). *Account integrity and authentic identity.* https://transparency.fb.com/en-gb/policies/community-standards/account-integrity-and-authentic-identity

Meta (2024b). *Fake Accounts, Community Standards Enforcement Report.* https://transparency.fb.com/data/community-standards-enforcement/fake-accounts/facebook/#content-actioned

Millar, P.R., Serbun, S.J., Vadalia, A., & Gutchess, A.H. (2013). Cross-cultural differences in memory specifity. *Culture and Brain, 1*(2-4), 138-157. https://doi.org/10.1007/s40167-013-0011-3

Mityamoto, Y., Yoshikawa, S., & Kitayama, S. (2011). Feauture and configuration in face processing: Japanese are more configural than Americans. *Cognitive Science, 35*(3), 563-574. https://doi.org/10.1111/j.1551-6709.2010.01163.x

Moravec, P.L., Kim, A., Dennis, A.R. (2018). Appealing to sense and sensibility: System 1 and System 2 interventions for fake news on social media. *Kelley School of Business Research Paper, 18.* http://dx.doi.org/10.2139/ssrn.3269902

Morton, C. (2016, July 21). The negative effects of false identity and how it can be overcome. Retrieved June 09, 2020, from https://www.virtualkollage.com/2016/07/false-identity.html

Murphy, N.A., Hall, J.A., Randall Colvin, C. (2003). Accurate intelligence assessments in social interactions: Mediators and Gender Effects. *Journal of Personality, 71*(3), 465-493. https://doi.org/10.1111/1467-6494.7103008

Muscanell, N.L, Guadagno, R.E., & Murphy, S. (2014). Weapons of influence misused: A social influence analysis of why people fall prey to internet scams. *Social and Personality Psychology Compass, 8*(7), 388-396. https://doi.org/10.1111/spc3.12115.

Na, J., Kosinski, M., & Stillwell, D.J. (2015. When a new tool is introduced in different cultural contexts: Individualsism-collectivism and social newtokr on Facebook. *Journal of Cross-Cultural Psychology, 46*(3), 355-370. Doi: 10.1177/0022022114563932

Nas, E., & de Kleijn, R. (2024). Conspiracy thinking and social media use are associated with ability to detect deepfakes. *Telematics and Informatics, 87*. https://doi.org/10.1016/j.tele.2023.102093

Nau, J., Halfens, R., Needham, I., & Dassen, T., Student nurses' de-escalation of patient aggression: a pretest-posttest intervention study. *International Journal of Nursing Studies, 47*(6), 699-708. https://doi.org/10.1016/j.ijnurstu.2009.11.011

Nadkarni, A., & Hofmann, S. (2012). Why do people use Facebook? *Personality and Individual Differences, 52*(3), 243-249. https://doi.org/10.1016/j.paid.2011.11.007

Nador, J.D., Zoia, M., Pachai, M.V., & Ramon, M. (2021). Psychophysical profiles in super-recognizers. *Scientific Reports*, *11,* 1-11. doi: https://doi.org/10.1038/s41598-021-92549-6.

Natale, S. (2021). *Deceitful media: Artificial intelligence and social life after the Turing Test.* Oxford University Press. https://doi.org/10.1093/oso/9780190080365.001.0001

Naveh Ben Dror (2024, February 29). *Locking your Facebook account on an iPhone, Android, or computer.* https://www.wikihow.com/Lock-Facebook-Profile#:~:text=While%20Facebook%20doesn't%20list,Ukraine%20can%20lock%20their%20profiles.

Nielsen, J. (2006). F-shaped pattern for reading web content. Nielsen Norman Group. https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content-discovered/

Nightingale, S.J., Wade, K.A., Farid, H., & Watson, D.G. (2019). Can people detect errors in shadows and reflections? *Attention, Perception, & Psychophysics, 81,* 2917-2943. https://doi.org/10.3758/s13414-019-01773-w

Nightingale, S.J., Wade, K.A., & Watson, D.G. (2017). Can people identify original and manipulated photos of real world scenes? *Cognitive Research: Principles and Implications, 2*(1), 30. 10.1186/s41235-017-0067-2

Nightingale, S.J., Wade, K.A., & Watson, D.G. (2022). Investigating age-related difference in ability to distinguish between original and manipulated images. *Psychology and Aging, 37,* 326-337. https://doi.org/10.1037/pag0000682

Nisbett, R.E., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences, 100*(19), 11163-11170. https://doi.org/10.1073/pnas.1934527100

OpenAI. (2022). ChatGPT [Large language model]. https://chat.openai.com/chat

Papacharissi, Z. (2009). The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. *New Media & Society, 11*, 199-220. 10.1177/1461444808099577

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1993). *The adaptive decision maker.* Cambridge University Press. 10.1017/cbo9781139173933

Pelled, A., Zilberstein, T., Pick, E., Patkin, Y., Tsironlikov, A., & Tal-Or, N. (2016, July). *Which post will impress me the most? Impression formation based on visual and textual cues in Facebook profiles* [paper presentation]. Proceedings of the 7th International Conference on Social Media & Society, 25, 1-10, London, UK. https://doi.org/10.1145/2930971.2930997

Pelto, P. (1968). The difference between 'tight' and 'loose' societies. *Transcation, 5,* 37- 40. https://doi.org/10.1007/BF03180447

Pennycook, G., & Rand, D.G. (2022). Nudging social media towards accuracy. *The ANNALS of the Americal Academy of Political and Social Science, 700*(1), 152-164. https://doi.org/10.1177/00027162221092342

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., & Rand, D.G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592,* 590-595. https://doi.org/10.1038/s41586-021-03344-2

Peters, A.N., Winschiers-Theophilus, H., & Mennecke, B.E. (2015). Cultural influences on Facebook practices: A comparative study of college students in Namibia and the United States. *Computers in Human Behavior, 49,* 259-271. https://doi.org/10.1016/j.chb.2015.02.065

Pitsea, M., & Thau, S. (2013). Compliant sinners, obstinate saints: how power and self-focus determine the effectiveness of social imfluences in ethical decision making.

Academy of Management Journal, 56(3), 636-658. https://doi.org/10.5465/amj.2011.0891

Potter, M.C., Wyble, B., Hagmann, C.E., & McCourt, E.S. (2014). Detecting meaning in RSVP at 13ms per picture. *Attention, Perception, & Psychophysics, 76,* 270-279. https://doi.org/10.3758/s13414-013-0605-z

R Core Team (2020). *R: A Language and Environment for Statistical Computing* (Version 1.3.1073) [Computer software]. Vienna, Austria. Retrieved from https://www.R-project.org/

Ramon, M. (2021). Super-recognizers – a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia, 158*, 1-11. doi: https://doi.org/10.1016/j/neuropsychologia.2021.107809.

Rohith Gandhi (2018, May 5). *Naïve bayes classifier.* Medium. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Romanov, A., Semenov, A., Mazhelis, O., & Veijalainen, J. (2017, April 25-27). *Detection of fake profiles in social media* [Conference session]. International Conference on Web Information Systems and Technologies (WEBIST 2017), Porto, Portugal. DOI: 10.5220/0006362103630369

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educ Psychol Meas, 81*(2), 340-362. doi: 10.1177/0013164420940378.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*(65). doi: 10.1057/s41599-019-0279-9.

Roozenbeck, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8*(34). DOI: 10.1126/sciadv.abo6254

Roozenbeck, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School (HKS) Misinformation Review, 1*(2). doi: 10.37016//mr-2020-008.

Riggio, R.E. (2002). *Social Skills Self Description Inventory* (2nd ed.). Mind Garden Inc.

Riggio, R.E., & Carney, D.R. (2003). *Social Skills Inventory Manual* (2nd ed.). Mind Garden Inc.

Rubin, V.L. (2017). Deception detection and rumour debunking for social media. In Sloan, L. & Quan-Haase, A. (Eds.). *The SAGE Handbook of Social Media Research Methods,* London:SAGE.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review, 16*(2), 252-257. doi: 10.3758/PBR.16.2.252.

Sandi, C., Rusconi, P., & Li, S. (2017). *Can humans detect the authenticity of social media accounts? On the impact of verbal and non-verbal cues on credibility judgements of twitter profile* [Conference session]. Proceedings of the 3rd IEEE International Conference on Cybernetics (CYBCONF), Exeter, UK. doi: 10.1109/CYBConf.2017.7985764.

Saunders, T., Driskell, J.E, Johnston, J.H., & Salas, E., (1996), The effect of stress inoculation training on anxiety and performance. Journal of Occupational Health Psychology, 1(2), 170-186. 10.1037//1076-8998.1.2.170

Schooler, L.J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112,* 612-628. 10.1037/0033-295X.112.3.610

Schramm-Nielsen, J. (2001). Cultural dimensions of decision-making: Denmark and France compared. *Journal of Managerial Psychology, 16*(6), 404-423. https://doi.org/10.1108/02683940110402389

Scott, W. (1808). *Marmion: A tale of Flodden field* (1st ed.). Edinburgh: George Ramsay and Company.

Scott, G.G., & Hand, C.J. (2016). Motivation determines Facebook viewing strategy: An eye movement analysis. *Computers in Human Behaviour, 56*, 267-80. https://doi.org/10.1016/j.chb.2015.11.029

Shao, C., Ciampaglia, G.L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking misinformation [paper presentation]. Proceedings of the 25th International Conference Companion on World Wide Web, Montréal, Québec, Canada. 745-750. https://doi.org/10.1145/2872518.2890098

Sharma, R., & Gautam, V. (2023). *New age digital media consumption in India: A survey of social. Media, OTT content and online gaming.* Esya Centre. https://www.esyacentre.org/documents/2023/8/18/age-digital-consumption-in-india-a-survey-of-social-media-ott-content-and-online-gaming

Shelley, M. (1818). *Frankenstein: The 1818 text.* Penguin Classics.

Shepherd, J.L., Lane, D.J., Tapscott, R.L., & Gentile, D.A. (2011). Susceptible to social influence: risky “driving” in response to peer pressure. *Journal of Applied Social Psychology, 41*(4), 773-797. https://doi.org/10.1111/j.1559-1816.2011.00735.x

Short, C. (Executive producer) (2019-2022). *For Love or Money* [TV series]. BBC.

Shrestha, S., Lenz, K., Chaparro, B., & Owens, J. (2007). “F” pattern scanning of text and images in web pages. *Proceedings of the Human Factors and Ergonomics*

Society Annual Meeting, 51(18), 1200-1204. https://doi.org/10.1177/154193120705101831

Sime, M., & Boyce, G. (1969). Overt responses, Knowledge of Results and Learning. *Programmed Learning and Educational Technology, 6*(1), 12-19.

Similarweb (n.d.). chatgpt.com. Retrieved March 27, 2024, from https://www.similarweb.com/website/chatgpt.com/#overview

Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129-138. https://doi.org/10.1037/h0042769

Simon, H.A. (1990). Bounded Rationality. In: J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability* (1st ed., pp 15-18. Palgrave Macmillan. https://doi.org/10.1007/978-1-349-20568-4_5

Starrett, G. (2003). Violence and the rhetoric of images. *Cultural Anthropology, 18*(3), 398-428.

Statista. (2023). *India – Statistics & facts.* Retrieved March 22, 2024, from https://www.statista.com/topics/754/india/#topicOverview

Statista. (2024a). *Number of internet and social media users worldwide as of January 2024*. Retrieved March 27, 2024 https://www.statista.com/statistics/617136/digital-population-worldwide/

Statista. (2024b). Actioned fake accounts on Facebook worldwide from 4th quarter 2017 to 4th quarter 2023. Retrieved March 27, 2024, from https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/#:~:text=Facebook%3A%20fake%20account%20removal%20as%20of%20Q4%202022&text=In%20the%20fourth%20quarter%20of,the%20first%20quarter%20of%202019.

Statista. (2024c). *Most popular social networks worldwide as of January 2024, ranked by number of monthly active users.* Retrieved March 22, 2024, from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Statista. (2024d). *Leading countries based on Facebook audience size as of January 2024.* Retrieved March 22, 2024, from https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

Stecher, K., & Counts, S. (2008, March 30 – April 2). *Thin slices of online profile attributes* [Conference session]. Proceedings on the 2nd International Conference on Web and Social Media (ICWSM), Seattle, Washington, USA.

Stocking, G., & Sumida, N. (2018). Social media bots draw public's attention and concern. *Pew Research Center's Journalism Project.* Retireved from https://policycommons.net/artifacts/617096/social-media-bots-draw-publics-attention-and-concern/1597862/

Stollak, M.J., Vandenberg, A., Burklund, A., & Weiss, S. (2011). *Getting social: The impact of social networking usage on grades amongst college students* [paper presentation]. American Society of Business and Behavioral Sciences, 18(1), 859-865. Las Vegas, NV, United States of America.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46*(2), 137-150. doi: 10.1037/h0062958.

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: systematic review. *Journal of Medical Internet Research, 23*(1). doi: 10.2196/17187

Sutcliffe, A., & Namoun, A. (2012). Predicting user attention in complex web pages. *Behaviour & Information Technology, 31*(7), 679-695. DOI: 10.1080/0144929X.2012.692101

Tajfel, H. (1978). Social categorization, social identity and social comparisons. In H. Tajfel (Ed.), *Differentiation between social groups*, 61-76. Academic Press.

Talib, A. & Saat, M. (2017). Social proof in social media shopping: An experimental design research. *SHS Web of Conferences 34*(1): 1-6. doi:10. 1051/shsconf/20173402005

Thaler, R.H., & Sunstein, C.R. (2009). *Nudge. Improving decisions about health, wealth, and happiness.*Penguin.

The Global Deception Research Team (2006). A world of lies. *Journal of Cross-Cultural Psychology, 37*(1), 60-74. https://doi.org/10.1177/0022022105282295

The Economic Times of India. (n.d.). *Most of Elon Musk's 153 million X followers fake, just 453,000 subscribe to X premium.* https://economictimes.indiatimes.com/tech/technology/most-of-elon-musks-153-million-x-followers-fake-just-453000-subscribe-to-x-premium/articleshow/102891718.cms?from=mdr

The Mindless Philosopher (2020). *Some Lies Really Aren't So Terrible: On Socrates' Noble Lie In American Political Thought*. Retrieved 9 June 2020, from https://themindlessphilosopher.wordpress.com/2011/02/04/some-lies-really-arent-so-terrible-on-socrates-noble-lie-in-american-political-thought/

Timberg, C., & Dwoskin, E. (2023, July 6). *Twitter is sweeping out fake accounts like never before, putting user growth at risk.* The Washington Post. https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/

Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, D. (2016). Intuition and moral decision-making – The effect of time pressure and cognitive load on moral judgement and altruistic behaviour. *PLOS ONE, 11*(10), 1-19. https://doi.org/10.1371/journal.pone.0164012

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27(*6), 813-833. DOI: 10.1521/soco.2009.27.6.813

Toma, C.L., & D'Angelo, J.D. (2014). Tell-tale words: linguistic cues used to infer the expertise of online medical advice. *Journal of Language and Social Psychology, 34*(1), 25-45. https://doi.org/10.1177/0261927X14554484

Traberg, C.S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The Annals of the American Academy of Political and Social Science, 700*(1). https://doi.org/10.1177/00027162221087936

Triandis, H.C. (1989). The self and social behaviour in differing cultural contexts. *Psychological Review, 96,* 506-520. https://doi.org/10.1037/0033-295X.96.3.506

Triandis, H.C. (1995). *Individualism and collectivism.* Westview press.

Tsikerdekis, M., & Zeadally, S. (2014). Online deception in social media. *Communications of the ACM, 57*(9), 72-80. https://doi.org/10.1145/2629612

Turing, A.M. (1950). Computing machinery and intelligence. *Mind: A Quarterly Review of Psychology and Philosophy, 59*(236), 433-460. https://doi.org/10.1093/mind/LIX.236.433

Turner, M., & Chin, E. (2017). *A cross-cultural comparison of the accuracy of personality judgements made through social media* [paper presentation]. International Conference on Applied Human Factors and Ergonomics, Los Angeles, California, USA. https://doi.org/10.1007/978-3-319-60747-4_1

Turner, M., & Hunt, N. (2014, June 22-27). *What does your profile picture say about you? The accuracy of thin-slicing personality judgments from social networking sites made at zero-acquaintance* [Paper presentation]. Social Computing and Social Media: 6th International Conference, SCSM 2014, Heraklion, Crete, Greece. https://link.springer.com/chapter/10.1007/978-3-319-07632-4_48

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. DOI: 10.1126/science.185.4157.1124

Twain, M. (1907). Chapters from my autobiography X. *North American Review, 184*(607), 113-119.

Tylor, E.B. (1871). *Primitive culture* (1996 ed., pp. 26-40). John Murray.

U.K. Department for Digital, Culture, Media & Sport. (2022, March 17). *World-first online safety laws introduced in Parliament* [Press release].

https://www.gov.uk/government/news/world-first-online-safety-laws-introduced-in-parliament

Vainuku, T., & Duffy, R. (Directors) (2022). *Untold: The Girlfriend Who Didn't Exist* [Film]. Players' Tribune.

van der Heide, B., D'Angelo, J.D., & Schumaker, E.M. (2012). The effects of verbal versus photographic self-presentation on impression formation in Facebook. *Journal of Communication, 62,* 98-116. Doi: 10.1111/j.1460-2466-2011-01617.x

van der Linden, S., Leiserowitz, A., Rosenthal, & S., Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2). https://doi.org/10.1002/gch2.201600008

van der Walt, E. (2018). *Identity deception detection on social media platforms.* Ph.D dissertation, Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, Pretoria, South Africa.

van der Walt, E. & Eloff, J. (2019). Identity deception detection: requirements and a model. *Information and Computer Security, 27*(4), 562-574. https://doi.org/10.1108/ICS-01-2019-0017

Varol, O., Ferrara, E., Davis, C.A., Menczer, F. & Flammini, A. (2017). *Online human-bot interactions: Detection, estimation, and characterization* [Conference session]. Proceedings of the 11th International Conference on Web and Social Media (ICWSM), Montréal, Quebec, Canada.

Vraga, E., Bode, L., Troller-Renfree, S. (2016). Beyond self-reports: Using eye-tracking to measure topic and style difference in attention to social media content. *Communication Methods and Measures, 10*(2-3), 149-164. https://doi.org/10.1080/19312458.2016.1150443

Vrij, A. (2000). *Detecting Lies and Deceit.* Wiley & Sons.

Walters, S.B. (2000). *The truth about lying: How to spot a lie and protect yourself from deception.* Sourcebooks, Inc.

Walther, J.B., & Parks, M.R. (2002). Cues filtered out, cues filtered in; computer mediated communication and relationships. In M.L. Knapp & J.A. Daly (Eds), *Handbook of Interpersonal Communication* (3rd ed., pp. 529-561). Sage.

Wang, Y., Norcie, G. & Cranor, L. (2011). *Who is concerned about what? A study of America, Chinese and Indian users' privacy concerns on social network sites* [Conference session]. International Conference on Trust and Trustworthy Computing, Pittsburgh, PA, United States. https://doi.org/10.1007/978-3-642-21599-5_11

Wani, M.A., Jabin, S., Yazdani, G., & Ahmad, N. (2018). Sneak into Devil's colony – A study of fake profiles in online social networks and the cyber law. *arXiv.* https://doi.org/10.48550/arXiv.1803.08810

Walter, N., & Murphy, S.T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs, 85*(3), 423-441. DOI: 10.1080/03637751.2018.1467564)

Weizenbaum, J. (1966). ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45. https://doi.org/10.1145/365153.365168

Willis, J., & Todorov, A. (2006). First impressions: Making your mind up after a 100-ms exposure to face. *Psychological Science, 17*(7)*,* 592-598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Wright, G.N., & Phillips, L.D. (1980). Cultural variation in probabilistic thinking: Alternative ways of dealing with uncertainty. *International Journal of Psychology, 15,* 239-257. https://doi.org/10.1080/00207598008246995

Yamagishi, T., Jin, N., & Miller, A.S. (1998). In-group bias and culture of collectivism. *Asian Journal of Social Psychology, 1*(3), 315-328. https://doi.org/10.1111/1467-839X.00020

Yang, C., Harkreader, R.C., & Gu, G. (2011, September 20-21). *Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers* [Conference session]. Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection (RAID' 11), Menlo Park, California, USA. DOI:10.1007/978-3-642-23644-0_17

Zaki, J., & Ochsner, K. (2011). Reintegrating accuracy into social cognition research: The empirical case for a broader approach. *Social Cognition, 29*(6), 599–631. https://doi.org/10.1521/soco.2011.29.6.599

Zhao, C., & Jiang, G. (2011). *Cultural differences on visual self presentation through social networking site profile images* [paper presentation]. Annual Conference on Human Factord in Computing Systems, Vancouver, British Columbia, Canada. 10.1145/1978942.1979110

Zhao, S., Shchekoturov, A., & Shchekoturova, S.D. (2017). Personal profile settings as cultural frames: Facebook versus Vkontakte. *Journal of Creative Communications, 12*, 171-184. 10.1177/0973258617722003

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in asynchronous computer-mediated communications. *Group Decision and Negotiation*, 13, 81–106.