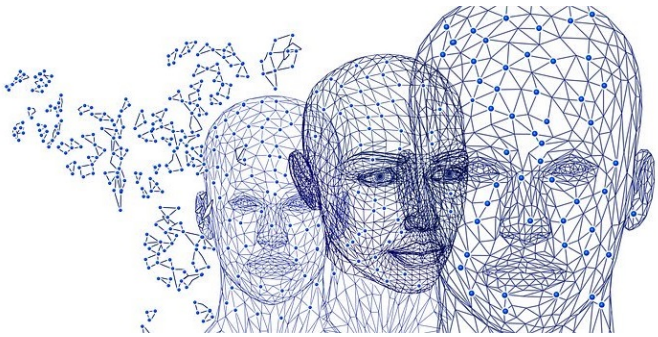
December 2020, Issue Code NL-2021-3

DDD (Digital Data Deception) Technology Watch Newsletter

Table of Contents



"*All warfare is based on deception. Hence, when we are able to attack, we must seem unable; when using our forces, we must appear inactive; when we are near, we must make the enemy believe we are far away; when far away, we must make him believe we are near.*"

— Sun Tzu, *The Art of War*

Editors: Enes Altuncu, Virginia Franqueira, Sanjay Bhattacherjee, Li Qin and Shujun Li
Affiliation: Kent Interdisciplinary Research Centre in Cyber Security (KirCCS), University of Kent, UK
Contact Us: dd-newsletter@kent.ac.uk

Editorial

In this third issue of the Digital Data Deception (DDD) Technology Watch Newsletter, we cover deception in different settings: (1) recommender systems, (2) communication (from a psychological perspective), and (3) cyber-physical systems.

We used a venue-based approach to select recent papers for sections (1) and (3), and we selected chapters from *The Palgrave Handbook of Deceptive Communication* related to two parts of the book deemed relevant. We also used a research group based approach to select recent papers from a set of five research groups in the UK, active in adversarial AI.

In total 26 research papers were summarised in this issue, all published since 2019. Four other supporting papers were cited as well, when applicable. This issue is the first to have an addendum Chinese section (NL-2021-3-C) where the scope of the DDD technology watch is extended to research papers published in Chinese; this section is available upon request.

We hope you enjoy reading this issue. Feedback is always welcome, and should be directed to dd-newsletter@kent.ac.uk.

List of Acronyms

- ADDR: Automatic Dyadic Data Recorded
- AFT: Adversarial Tensor Factorization
- ALIED: Adaptive Lie Detector Theory
- AR: Augmented Reality
- CIT: Concealed Information Test
- CNN: Convolutional Neural Network
- COLD: Contextual Organization of Language and Deception
- CPS: Cyber-Physical Systems
- DCGAN: Deep Convolutional Generative Adversarial Network
- DNN: Deep Neural Network
- FGSM: Fast Gradient Sign Method
- FID: Fréchet Inception Distance
- GLR: Generalized Likelihood Ratio
- HMI: Human-Machine Interaction
- ICS: Industrial Control Systems
- JSMA: Jacobian-based Saliency Map Attack
- LIWC: Linguistic Inquiry and Word Count
- LSTM: Long Short-Term Memory
- NIST: National Institute of Standards and Technology
- MCTS: Monte-Carlo Tree Search
- PGD: Projected Gradient Descent
- RBF: Radial Basis Function
- ReLU: Rectified Linear Unit
- RNN: Recurrent Neural Network
- ROC: Receiver Operating Characteristic
- SIFT: Scale Invariant Feature Transform
- STL: Stereolithography
- SUE: Strategic Use of Evidence
- SUS: Small Unmanned Systems
- SVM: Support Vector Machine
- VR: Virtual Reality
- WER: Word Error Rate
- WFM: Weighted Fusion Model

Deception in Recommender Systems

Introduction

Recommender systems have been widely used in many real-world applications. Since such systems use collaborative filtering algorithms to provide recommendations based on normal people's activities as direct input, they can be attacked by malicious users who inject false information and deceptive data into the system, known as "shilling attacks" in the literature [11]. It is therefore interesting to investigate how recommender systems can be misled by false and deceptive inputs, and how their robustness to such inputs can be improved. In this section, we cover four recent research papers on this topic, all published at RecSys (ACM Recommender Systems Conference), the top-tier conference on recommender systems in 2019 or 2020.

Attacks

Christakopoulou and Banerjee [6] proposed to apply adversarial machine learning to automate attacks on an oblivious recommender system, i.e., one that interacts with the attacker but is oblivious to the attacker's existence. The attack tries to create fake user profiles to meet two goals: (1) they are *indistinguishable* from real user profiles based on some reasonable metrics; (2) they have a *malicious intent* to affect the output of the recommender system, e.g., putting down the positions of the target profiles so that they drop out of the top profiles recommended. The first goal is to make sure that the generated fake user profiles are not noticeable by the system and other users, so they are deceptive. The attack is formulated as a general-sum game between the recommender and the attacker, each trying to minimise a different loss function. The generation of fake profiles is achieved via the use of the Deep Convolutional Generative Adversarial Networks (DCGANs) architecture. In the attack, the adversary does not have access to the gradient information of the recommender, so the authors proposed a zero-order optimisation method to approximate the gradient information needed. The authors conducted a wide range of experiments under two setups: 1) the adversary targets unrated user-item entries (i.e., candidates for recommendation); 2) the adversary targets a small subset of (user, item, rating) tuples gathered from the training set of the recommender. The experimental results showed that the adversary can successfully target a number of areas of the recommender's outputs, such as top predicted items of a user, top users of an item, a user's hit rate, an item's prediction error, and modelling gaps between user groups. This paper demonstrates that more research is needed to build recommender systems that are more robust against such machine learning attacks.

Tang et al. [25] looked at the scenario where the attacker uses a local surrogate model to learn how to inject fake user behaviours that will be used to attack the recommender system (see Figure 1 for the general architecture of such attacks). They criticised the "white-box" assumption of many other researchers used for the local surrogate model, and considered a more realistic scenario where the local surrogate model differs from the target model so there is a problem of attack transfer. The authors modelled the

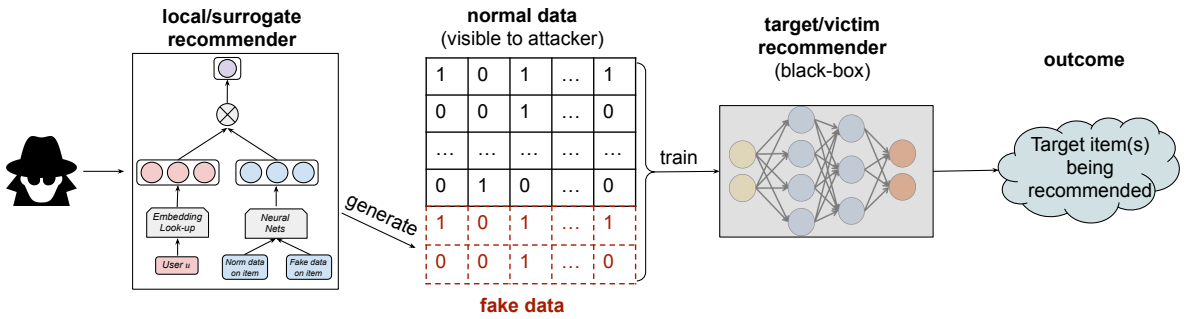


Figure 1: The general architecture of the injection attack against recommender systems considered by Tang et al. [25].

attack as a bi-level optimisation problem, with an inner objective describing consumption of injected fake user behaviours by the local surrogate model and an outer objective describing how the attack achieves its goal on normal users' predictions after the fake data is consumed. They pointed out that past studies had under-estimated the power of such an attack due to less accurate estimate of gradient computation, and studied the exact solution with two ways to calculate the approximate. Using a real-world dataset on user-venue check-ins (available at `http://snap.stanford.edu/data/loc-Gowalla.html`), the authors conducted experiments to show the transferability from surrogate models to different types of target recommenders. They also discussed limitations of the attack considered, including 1) the attack is less effective on "cold" items and 2) the fake user learned can still be detected by a recommender system that is aware of such an attack. While the results do not lead to practical attacks, they can help inform develop defensive mechanisms against such attacks. The authors made the source code of their implemented attacks available at `https://github.com/graytowne/revisit_adv_rec`.

Defence against Attacks

Aktukmak et al. [1] proposed a method for detecting fake user profiles that are injected to mislead recommender systems. Their method is based on a probabilistic matrix factorisation model that can embed observed ratings and attributes of genuine users into a low-dimensional space and provide useful anomaly statistics for detecting new fake users. The detection capability is based on the assumption that a genuine user's profile and ratings match the statistics learned from genuine users better than fake users because the latter often involve some randomness (either on the generation of the profile or the ratings). Such mismatches over time can be captured by a cumulative anomaly statistic, which can be used to issue an alarm if the statistic exceeds a threshold. Using the MovieLens 100K dataset with 943 users and 1,682 items, the authors generated 94 fake user profiles by mixing three well-studied attack types (random, average and bandwagon). Comparing with four baseline detection methods, the authors showed that the proposed method outperformed all of them with a significant margin in terms of the area under the ROC (Receiver Operating Characteristic) curve, as shown in Figure 2. In addition, in terms of the detection speed, the authors showed that their method was very efficient, comparing with an existing method known to be very efficient – the GLR (Generalized Likelihood Ratio) detector proposed by Li and Wang [16].

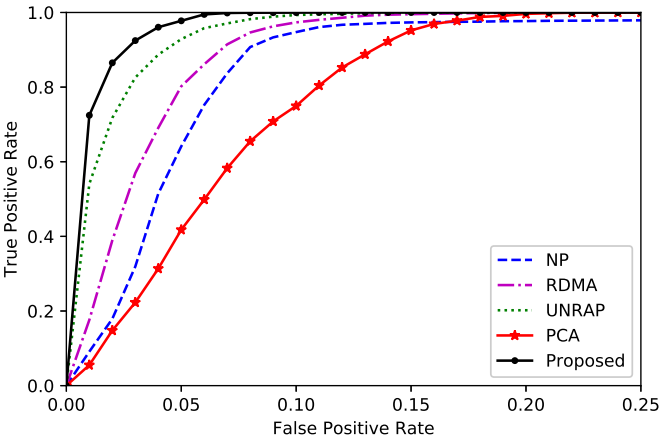


Figure 2: The performance of the fake user profile detection method proposed by Aktukmak et al. [1], compared against several state-of-the-art methods.

Chen and Li [4] looked at the robustness of context-aware recommender systems against adversarial samples. They proposed ATF (Adversarial Tensor Factorization), a model that combines tensor factorization and adversarial learning to improve the robustness of context-aware recommendations. The basic idea is that adversarial perturbations are used to simulate attacks on model parameters, while the model is trained in such a way to defend against such perturbations for self-improvement. This is achieved via a unified objective function considering both adversarial perturbations and model parameters. Using two real-world datasets, MovieLens and Last.fm each with around 2,000 users and over 10k tags and items, the authors showed that their proposed method outperformed three standard tensor models in tag recommendations.

The Palgrave Handbook of Deceptive Communication

Introduction

The Palgrave Handbook of Deceptive Communication [8] is relevant for DDD. This book was identified in our ad-hoc search for related work. In this section of the newsletter, we cover most chapters from two parts of the book, titled (1) "Deception Theories, Frameworks, and Approaches", and (2) "Detecting Deceptive Communication". The first chapter of the part on detecting deceptive communication has already been covered in Issue 2 (NL-2021-2). The content of this book can inform and enable the design and development of more effective deceptive technologies and detection techniques against them.

Deception Theories, Frameworks, and Approaches

Powers [21] distinguished deception from error. In particular, *discursive deception* was defined as the intentional use of language to mislead, misdirect, or misinform another person in order to induce them to follow a flawed path of thinking, belief, or behaviour. The author analysed situations when the sender of a message has reason to distort the truth for their personal benefit. The conscious use of language to mislead others was explored. The author admitted that it is potentially difficult to prove a sender's *intent* to deceive. Hence, the theories discussed in the chapter could also be applied to understanding a sender's linguistically induced error. The author described a framework of concepts and principles for analysing deceptive discursive practices that can occur at four different levels of language, moving from the smaller units of discursive communication to the larger ones. The chapter looked at deception arising at the lexical (i.e., individual word choice) level of discourse, and explored the propositional or syntactic bases of deception. The authors surveyed the speech act dimensions of deceptive discursive practices. Finally, they featured the types of deception that can arise from the macro-semantic dimensions of discursive communication such as descriptions, narratives, and argument structures. Each section introduced a small number of principles related to a particular level as a starting point for identifying the modes of deception that are made possible by abusing the key principles of language (at that level).

Markowitz and Hancock [19] pointed out that a growing body of research suggests that language cues for deception are not universal. The effect of deception on word patterns varies with changes in the context and settings. This makes it difficult to draw conclusions about the overall impact of deception on the use of language. The authors outlined the problems in considering a universal approach to deception and language. They considered how research has revealed the effect of various moderators (e.g., the mode of message production, the valence of the situation) on how language is affected by deception. This suggested that deception is a context-contingent phenomenon. They also addressed how language, independent of deception, is highly context dependent. Finally, they outlined the Contextual Organization of Language and Deception (COLD) framework, which proposes that psychological dynamics (e.g., emotional and cognitive processes modified by deception), pragmatic goals (e.g., what the speaker is trying to accomplish with their deception), and genre conventions (e.g., the norms of each discourse community that shape how language is produced) profoundly and systematically influence the effect of deception on language. They applied the COLD framework to a database of deceptive political speeches from six US presidents (George W. Bush, Lyndon B. Johnson, Bill Clinton, Richard Nixon, John F. Kennedy, and Ronald Reagan), finding support for the idea that false language patterns are reliably modified by the deception type. Figure 3 summarised their findings through raw Linguistic Inquiry and Word Count (LIWC) percentages of the total word count.

Street et al. [24] proposed a shift towards a theory-driven approach to lie detection research. They explored the reasons behind *truth bias* and the *lie bias* – the tendencies of people to believe and disbelieve others. The *adaptive lie detector theory*, or ALIED theory, recognises that these biases are adaptive and functional, rather than a sign of error. They briefly reviewed recent tests of the ALIED theory. The authors also made critical comments regarding the troubling trend of the lack of theoretical progress in lie detection at the moment. The practice has been to observe effects, tag an explanation onto them, and hold it as a theory. The authors called for a shift toward theories that would be falsifiable. Predictions

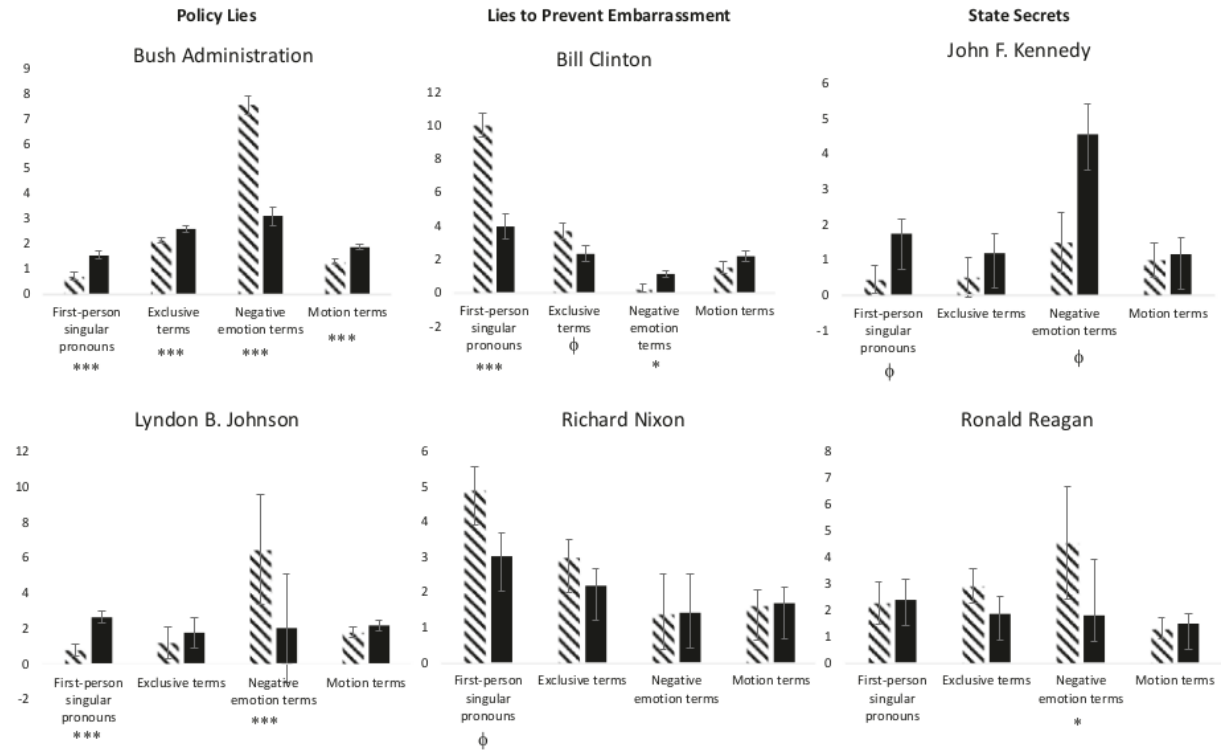


Figure 3: Evaluation of six US presidents using the COLD framework proposed by Markowitz and Hancock [19]. Vertical axes represent raw LIWC percentages of the total word count. Striped bars indicate deceptive statements while solid bars indicate truthful control statements. $*p < 0.5$; $***p < 0.001$; $\phi = p < 0.8$. Error bars are standard errors.

should emerge from such a theory, which can then be tested. To take examples from the field of astronomy, Newton's theory of gravitation predicted the existence of Uranus before we had telescopes powerful enough to observe it. We have moved further ahead from Newtonian physics to that of Einstein through repeated falsification. Any theory changes and shifts over time, but can maintain its core. In doing so, it develops original predictions that should be verified to ascertain the correctness and robustness of the theory. The field of lie detection currently lacks this ability to generate clearly defined predictions that can then be ratified or falsified.

Williams and Muir [28] focused on the fact that a crucial factor when successfully deceiving individuals is to make them trust that the concerned scenario and communication are genuine. Deceivers often manipulate established norms and trust mechanisms to help in the deception process. This chapter explored some of the methods used by individuals to develop trust in communication, and to signal that the positive expectations held by another individual will be met. These included *linguistic mechanisms*, like self-disclosure and verbal mimicry, and *situational mechanisms*, like heuristic based strategies. The latter leads to automatic biases (such as pre-conceived stereotypes, expectations, or emotional responses) in decision-making or more resource-intensive, systematic, processing strategies. They proceeded to explore scenarios after trust has been built, when a deceiver can exploit these communicative mechanisms. They presented an initial model of trust manipulation described in Figure 4 that brings all of the above factors together to consider how elements of communication, such as building rapport and the use of authenticity cues, may be used to invoke trust in order to effectively deceive others.

Carr et al. [3] noted that researchers have historically attempted to use their latest ideas and technology to detect lies. However, such techniques were difficult to apply in the real world because the methods and variables used were often not well defined. The most critical question for any deception researcher interested in designing a valid deception scenario should be – "*Does my artificial interview accurately reflect an interview in the real world?*" Such con-

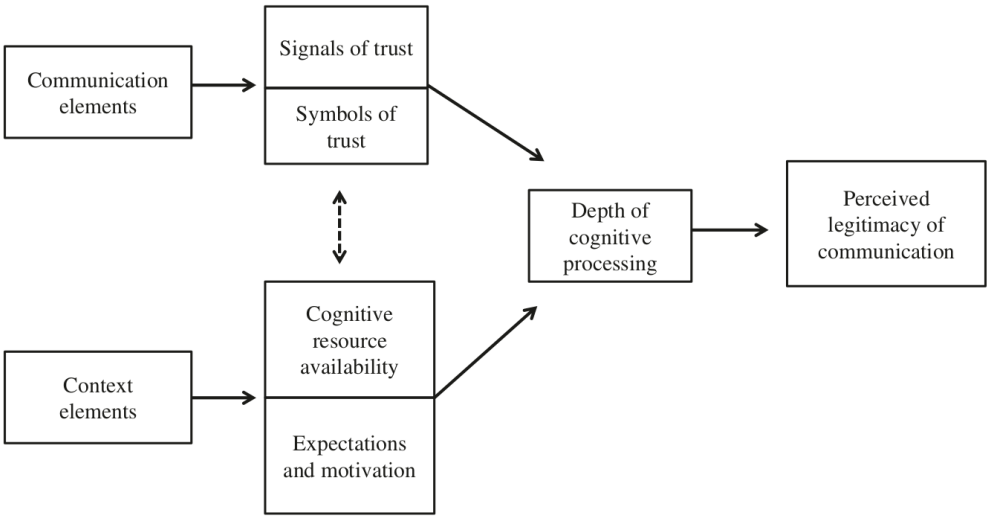


Figure 4: An initial model proposed by Williams and Muir [28] of how receiver trust may be manipulated by deceivers.

trolled experiments are typically inaccurate in replicating real situations. So, a more appropriate question to ask would be – "*How close can I get my artificial interview to reflect a real-world interview?*" Researchers should then report results and discuss limitations when attempting to generalise those results, as they too will not often perfectly match a real-world high-stakes interview. This chapter proposed that researchers make five important considerations when designing any deception study: (i) their specific definitions of phenomena, (ii) their use of stakes or incentives, (iii) their allowance of participants to choose to lie, (iv) their use of sanctioned or unsanctioned lies, and (v) their appreciation of the power of the interview process itself in generating behaviours associated with truth or lie. They concluded that researchers should approach the study of deception in a more united, clearly defined methodological fashion for the betterment of our collective scholarly knowledge and for those professionals who rely on it.

Detecting Deceptive Communication

Sternglanz et al. [23] reviewed and synthesised meta-analytic studies about deception detection. The authors examined the scope, methodology, and findings of meta-analyses on the following topics: deception detection accuracy, moderators of accuracy, perceived verbal and nonverbal cues to deception, actual verbal and nonverbal cues to deception, physiologically based techniques for detecting deception (including polygraphs and brain-imaging tools), cognitive/interrogative techniques for detecting deception, and the effectiveness of training to detect deception. They discussed meta-analytic findings about deception detection techniques commonly used by law enforcement as well as techniques used by common people in their interpersonal interactions. They also briefly discussed useful topics for future meta-analyses about deception-related topics, as well as methodological strengths and limitations of meta-analyses about deception detection.

Larner [14] pointed out that previous research into deception detection argued that deception is more cognitively demanding than truth-telling. This additional cognitive load can lead to changes in linguistic and non-linguistic behaviour, which, in turn, can be considered cues to deception. While the majority of deception research was rooted in Psychology, this chapter approached deception from a linguistics perspective by proposing and empirically testing a feature of language used to manage the cognitive demands of interpersonal communication. *Formulaic sequence* is an umbrella term for sequences of words including metaphors, clichés, collocations, and routine phrases. Sequences are stored holistically as single lexical items, and this makes the act of producing language less cognitively demanding. The hypothesis, therefore, was that individuals may seek to compensate for the additional cognitive demands of lying by increasing their reliance on formulaic sequences. To test this assertion,

the authors identified formulaic sequences in a corpus of 1600 deceptive and truthful hotel reviews totalling 239,113 words. After removing non-formulaic matches, a total of 2279 formulaic sequences were identified across all the data. These sequences were composed of 525 different types ranging from one to six words. Table 5 provides examples of such sequences. The authors used an automated procedure based on a specially compiled dictionary of formulaic sequences. The results shed light on the relationship between formulaic sequences and deceptive language, their potential role in detecting deception, and the generalisability of findings to other types of texts.

N	Matches	Examples
1	276	okay, on-the-phone, plain-as-day, state-of-the-art
2	1260	above average, final straw, in future, no brainer
3	504	blew me away, down to earth, in my opinion, in the meantime
4	217	bump in the road, icing on the cake, spur of the moment, set my sights on
5	16	a piece of my mind, as hard as a rock, at our beck and call, bad taste in my mouth
6	6	you get what you pay for, to make a long story short, cost an arm and a leg

Figure 5: Examples of *formulaic sequences* identified by Larner [14], where N is the number of words constituting a formulaic sequence, ranging from one to six.

Dianiska et al. [7] described and discussed research that documents how the act of lying can influence the content of the liar's memories of the occasions when they lied and memories of the original experience. The authors presented two types of lying – *confabulations* (lies that involve a person describing a specific event or experience as if it had occurred), and *false denials* (lies in which a person says that an event never occurred, although the event took place). False denials require lesser effort compared to confabulations because no new details have to be fabricated. They also elaborated on how an understanding of memory processes can be a tool for uncovering deception. For instance, the content of memories of actual and fabricated events differ in characteristic ways, and people can be trained to utilise these features to discriminate between them. Furthermore, it is possible to magnify differences in reports of liars and truth-tellers to increase detection. Therefore, memory can play a critical role in catching liars.

Geven et al. [9] pointed out that the bodily responses of a lying individual could be very similar to the ones of an individual who is experiencing increased stress when facing a lie detection test. The importance of avoiding wrongful incarceration leads us to techniques detecting memory rather than lies. The Concealed Information Test (CIT), proposed by the authors, aimed to detect the recognition of concealed knowledge in an interviewee by presenting a series of multiple-choice questions while measuring several psycho-physiological (e.g., skin conductance) or behavioural (e.g., reaction time) responses. When a suspect was subjected to the critical (e.g., crime-related) items and consistently showed distinct responses, compared to the neutral control items, inferences were drawn. This chapter provided an overview of memory detection using various response measures, including research findings and the underlying mechanisms. Available data confirmed the validity of the CIT. However, there was quite a gap between these laboratory studies and realistic criminal investigations. The authors discussed possible ways to tackle challenges on the topic, including field validity, leakage of critical information to innocent suspects, and discovering intentions.

Mac Giolla and Granhag [18] raised attention that most research on deception detection was traditionally focused on statements about the past. In this chapter, the authors provided an overview of studies in true and false intentions, where the focus changed to statements about the future. A statement of *true intent* refers to a future action that a speaker intends to carry out, while a statement of *false intent* refers to a future action that a speaker claims, but does not intend to carry out. An ability to distinguish between such statements holds great practical value for a myriad of professions. The chapter defined key

terms in the field, summarised the extent of research, and highlighted recent theoretical developments and areas for future research.

Kleinberg et al. [13] considered the problem of detecting deceptive intent as in the last chapter, but they formed a complementary perspective of large-scale applications. They outlined a set of criteria that an applied system should meet from a practitioner's perspective to evaluate deception theories, interviewing approaches, information elicitation methods, and verbal deception cues that may be of use for large-scale applications, such as for prospective airport passenger screening. Their findings indicated as promising approaches: (i) the cognition-based deception theory, (ii) the information-gathering interviewing approach, (iii) the unanticipated questions method and the model statement technique, and (iv) verbal cues, especially the verifiability of details and stylometric cues. They concluded the chapter with an illustration of how this combination of elements can be put to use.

Wilson and Rule [29] reviewed the accuracy of impressions of dispositional deceptive tendencies. In other words, they looked at evaluation of the accuracy and inaccuracy in predicting deception from facial appearance. They pointed to research that has recently started to address the question whether indirect cues, such as facial appearance, can predict people's deceptive behaviour. The answer was unclear, and results from different domains were mixed. The authors attempted to identify circumstances in which deception was accurately perceived. They aimed to clarify distinctions between deceptive versus non-deceptive untrustworthy behaviours that can be detected from the face. The authors suggested that the review they provided may foster better hypotheses about the specific cues that predict deceptive behaviour.

Introduction

This section provides a selection of recent research where deception potentially happens via Cyber-Physical Systems (CPS). NIST (National Institute of Standards and Technology, of the US's Department of Commerce) defines CPS [10] as "smart systems that include engineered interacting networks of physical and computational components". The first sub-section covers everyday objects that may become touchable interfaces or interactive, raising innovative opportunities for digital deception. The second sub-section discusses papers where deception may happen via wearable devices. In particular, it covers an engineered bracelet and smart shoes available commercially which may, deceptively, be used to obstruct microphones nearby or gather intelligence about a physical environment.

Deception via Physical Objects

Tejada et al. [26] presented a technique, called *AirTouch*, to manufacture touch-sensitive objects using a consumer-level resin-based stereolithography (STL) 3D printer. It leverages on principles of fluid dynamics, i.e, the principle of continuity (the total flow of air entering and exiting an object must be equal) and the Bernoulli's principle (blocking the flow of air from an outlet causes an increase in pressure), to recognise which outlets on the surface of an 3D printed object is touched; this is illustrated in Figure 6. Their approach has 4 main components: (1) an object (e.g., an interactive animal) built with a 3D printer containing outlets, i.e., air passage holes, of specific diameter on its surface; (2) an internal structure (including a flow-distribution chamber connected to cylindrical tubes linked to outlets) built inside the object also using 3D printer; (3) a setup to connect an air compressor tubing to valves, and a barometric sensor to the object (air pressure is generated using an Arduino board); and (4) software for recognition of touch events and identification of corresponding outlet. The design of the 3D components (a) and (b) are facilitated by an Autodesk Meshmixer script. The feasibility of their approach for recognition of touch was evaluated for AirTouch-enabled objects of different shapes and outlet configurations: an interactive bar chart (example of *data physicalization*), an interactive animal, a grasp-sensing sphere, and a color hue selector. The authors trained a Support Vector Machine (SVM) model with a Radial Basis Function (RBF) kernel using 1000 samples to classify results of touch events. The lowest accuracy was obtained for the "grasp-sensing sphere" (91.6%), and the highest was obtained for the "interactive animal" (100%). A couple of limitations were discussed of this **early stage** proposal to turn static objects into interactive ones, such as the need for an air compressor powering the 3D objects. However, as this field progresses and more sophisticated interactions are achieved, deceptive objects can start collecting and processing data.

Iravantchi et al. [12] proposed a *digital ventril-*

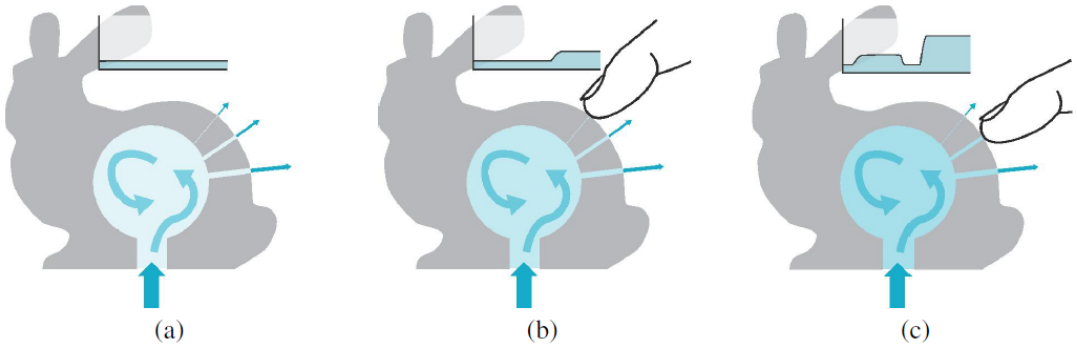


Figure 6: The *AirTouch* technique proposed by Tejada et al. [26] relies on fluid dynamics for touch recognition; (a) shows a flow-distribution chamber built inside a 3D printer generated object, and 3 outlets (i.e., holes on the object surface) of different diameters but same configuration; (b) and (c) show different outlets being pressed causing different barometric pressure responses.

oquism prototype to enhance ordinary inanimate objects, located within the same environment as a smart speaker, with the ability to render sound and give the illusion that they can interact with humans. This approach allows, e.g., a plant to remind people to water it, or a picture frame to tell stories related to the occasion depicted. The prototype leverages from "directed ultrasonic beams that, when modulated with an input signal, are inaudible in flight and demodulate when striking a surface, allowing the sound to emanate from the target rather than the speaker itself". A number of non-expensive components were used to build the prototype, as shown in Figure 7, and the pre-trained YOLO (v3) algorithm was used for real-time object detection and bounding. The authors reported on 3 types of experiments and studies to inform and validate their approach. The first type involved *physical studies* to understand the effect of different materials and geometry of objects in relation to the sound emanation scheme. The second type was a *real objects study* where the authors tested the reflected acoustic power of their ventriloquism scheme using different angles and distances between the speaker and 24 objects typically found in an office, domestic, workshop and outdoor environments (without any mechanism in place to control "normal" background noise). Overall, they noticed that, although larger objects outperformed (e.g., dishwasher), objects with complex geometry and asymmetries performed well on an off-angle setup such as 90° or −90° (e.g., plant pot). In all cases, however, the sound remained intelligible. The third type was a *user study* with 5 participants aimed at testing recognition of the "talking" object and its conveyed message in the office, domestic and workshop environments again. Results indicated a high degree of correct localisation (92%) and comprehensibility (100%). The main limitations of the implemented prototype are: ultrasound cannot pass through walls and large objects, therefore, the ventriloquism illusion is disturbed if a user walks past the ultrasound beam; objects made of absorptive materials do not reflect sound well; the generated sound operates at a frequency which does not sound as natural as typical speakers do (such as Amazon's Alexa); and the apparatus used to direct the ultrasound beam is cumbersome and alternatives would blend better into the environments studied and remains as future work.

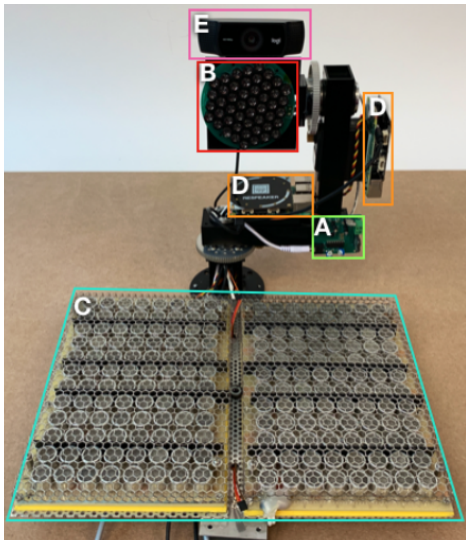


Figure 7: Components of the *Digital Ventroloquism* prototype proposed by Iravantchi et al. [12]: (A) Driver Board, (B) Small Speaker Array, (C) Large Speaker Array, (D) Raspberry Pi Microphone Array, and (E) Webcam.

Deception via Wearable

Chen et al. [5] leveraged from the fact that we are surrounded by devices with microphone capabilities potentially able to listen and record conversations (e.g., smartphones, voice assisted smart speakers, and smartwatches) due to default features, misconfiguration or misuse by attackers. To counter this threat, the authors proposed a *bracelet-like ultrasound jammer* (as illustrated in Figure 9) to make conversations incomprehensible to surrounding (smart) devices, therefore protecting users' privacy and security while not disturbing them since this technology is inaudible to humans. The self-contained wearable jammer prototype has been created using a 3D printer model of the bracelet, which can be turned on/off, and the following components: ultrasonic transducers, a signal generator, a microcontroller, a battery, a voltage regulator and a 3W audio amplifier. A number of experiments and Matlab simulations to inform the jammer design were conducted. Results evaluating the wearable jammer, against a planar jammer (proposed in the literature) and the i4 jammer (according to the authors, available at Amazom.com), indicated that the prototype outperforms (1) in angular coverage, (2) in mitigating blind spots, (3) in jamming effectiveness (measured in terms of Word Error Rate (WER) for the transcribed speech), and (4) in with-

standing noise cancellation attacks. The authors also validated how the bracelet jammer would perform against hidden microphones, covered by different materials; results showed that its performance remains unaffected by materials such as paper tissue, paper sheet, foam windshield and cloth (WER of 99%), but is highly affected by materials such as plastic box and cardboard box (WER 41.01% and 46.76%, respectively). A "user study" with 12 participants, and 4 smartphones was also conducted. Overall, it indicated a positive perception of the bracelet by users as a mechanism to protect their privacy (M = 5.4; SD = 1.1), and the perception that it would be useful in the context of sensitive conversations although very noticeable. However, the authors noted that their results are not easily generalisable to different types of smart devices in the environment, and the bracelet may cause undesirable side-effects like jamming the user's own smartphone, hearing aid devices, and emergency response devices. The simulation source code, hardware design, firmware and schematics to allow replication of results is available at `https://github.com/y-x-c/wearable-microphone-jamming`.

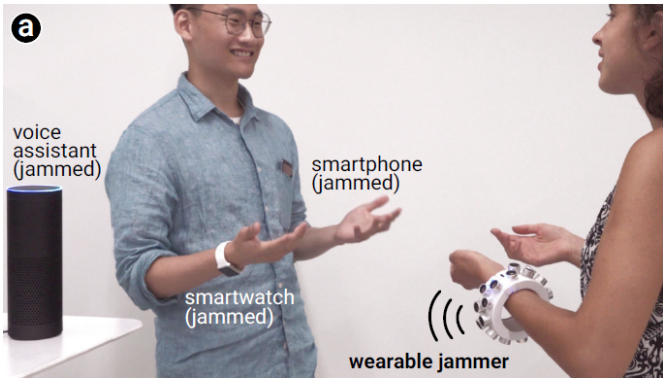


Figure 9: Illustration of the wearable ultrasound jammer designed by Chen et al. [5].

Yu and Nahrstedt [30] explored the possibility of an attacker exploiting "foot force data" collected by smart shoes, usually uploaded by users to cloud servers for statistical analysis, to reconstruct the corridors layout of a building and locate the building on a map (such as Google Map). Therefore, the authors proposed an attack in the context of a potential victim being tracked and, ultimately, physically located inside the building by an attacker (if the foot force data is available in real time). To demonstrate the feasibility of such an attack, the paper proposed the *ShoesHacker* prototype system which leverages from 5 main stages with a number of tasks, as illustrated in the system's architecture in Figure 8. The "walking path estimator" and the "corridor map estimator" stages use Support Vector Machine (SVM) classifiers to extract features from footsteps and determine direction changes; 5 algorithms were also presented to achieve some tasks scattered across the 5 stages. The prototype was evaluated with the help of 10 volunteers, with different heights and weights, using the smart shoes called "ReTiSense Stridalyzer" which contains 8 foot force sensors transmitting data to a phone via Bluetooth. Several settings were used to evaluate the performance of (1) Stair Landing Detection, (2) Angle Regression, (3) Walking Path Estimator, (4) Corridor Map Estimator, and (5) Building Recognizer. Limitations were discussed such as the assumption of typical office buildings with 90° angled corridors and squared staircases. Other types of staircase, such as spiral ones, would result in inaccurate direction (U-turn) labels extracted from the training data, and irregular corridor structures would result in inaccurate angle regression. Also, the use of elevators is currently not handled by the system. Nevertheless, this paper showed the potential of leveraging data from smart shoes for deceptive purposes via valuable intelligence gathered simply by walking through an indoor space.

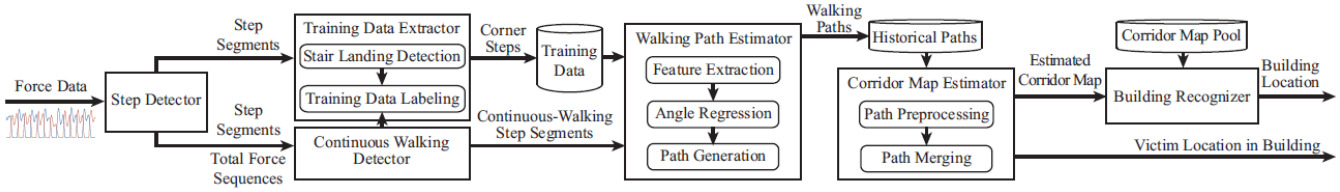


Figure 8: Architecture of the *ShoesHacker* prototype system proposed by Yu and Nahrstedt [30].

Selected Adversarial AI Research Groups in the UK

Introduction

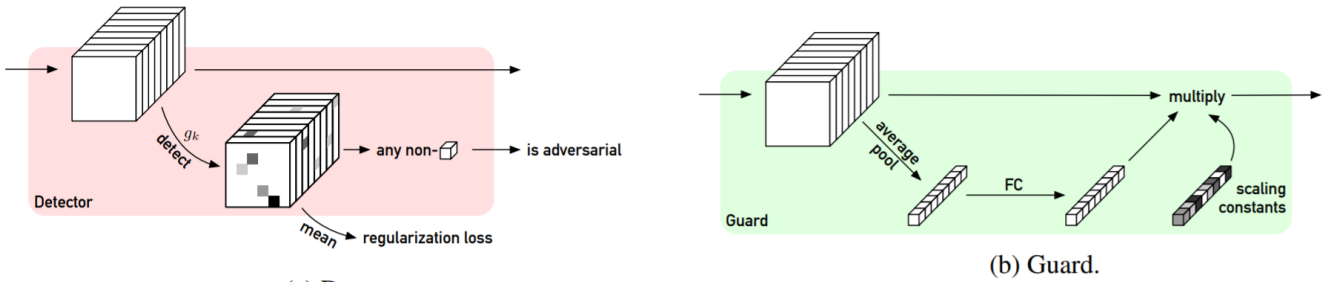
This section provides information about a number of research groups in the UK, having a strong interest in Adversarial AI. The selected groups have been chosen according to their strength and recent activities, i.e., recent publications.

Security Group, University of Cambridge

The Security Group at the University of Cambridge (`https://www.cl.cam.ac.uk/research/security/`) is a world leading research group focusing on security engineering and hosts the world-renowned cyber security researcher Professor Ross Anderson (`https://www.cl.cam.ac.uk/~rja14/`). They recently have published the following two papers on adversarial AI.

Shumailov et al. [22] introduced a system they designed to block the transfer of adversarial samples, which was called *Sitatapatra*. The idea behind the system was that adversarial samples are portable, meaning that the devices using the same CNN are vulnerable against the same adversarial samples. Therefore, to avoid the transferability of adversarial samples, the authors were inspired by cryptography and introduced a notion of *key* into Convolutional Neural Networks (CNNs) that causes each network of the same architecture to be internally different enough. As shown in Figure 10, each convolutional layer with Rectified Linear Unit (ReLU) activation is sequentially extended with a guard layer (Figure 10b) and a detector (Figure 10a). Intuitively, the guard encourages the gradient to disperse among differently initialised models, limiting sample transferability. If this fails, the detector works as our second line of defence by raising an alarm at potentially adversarial samples. The authors described multiple ways of embedding the *key*s and evaluated them using the MNIST and CIFAR10 datasets. Based on these data, they proposed a scheme to select *keys*. The proposed system is capable of tracing the detected adversarial samples back to the individual device which was used to develop the samples, according to the authors. In addition, the authors claimed that Sitatapatra can be used on constrained systems due to its minimal run-time overheads (0.6-7%).

Van der Zee et al. [27] investigated whether *full body motion* can be a signal for detecting deception. Therefore, rather than focusing on specific gestures such as fidgeting and gaze aversion, they examined the effectiveness of detecting deception from a full body motion which includes position, velocity, and orientation of 23 points in the subject's body. For the experiments, the authors recruited 60 South Asian and 30 White British interviewees and asked them to either tell the truth or a lie regarding two specific tasks. Then, they measured the participants' full body movements by using Xsens MVN motion capture suits, as shown in Figure 11. The results of the experiments showed that full body motion, i.e. the sum of joint displacements, was indicative of lying 74.4% (truths: 80.0%, lies: 68.9%). Further analyses indicated that including individual limb data in our full body motion measurements can increase its discriminatory power to 82.2% (truths: 88.9%, lies: 75.6%). Furthermore, the authors reported that movement was guilt- and penitential-related, and oc-



(a) Detector.

(b) Guard.

Figure 10: High-level view of the module extensions, proposed by Shumailov et al. [22], and added to CNNs to stop and detect transferred adversarial samples.

curred independently of anxiety, cognitive load, and cultural background. Regarding the investigation on cultural background, they also noted that the participants they tested did not differ enough from a cultural perspective to elicit distinctive behavioural patterns since all participants were students or employees of Lancaster University which means that, while South Asian participants were born and raised in South Asian countries, they have also spent a significant amount of time in the UK. Other than that, this study provides quite promising results on detecting deception.

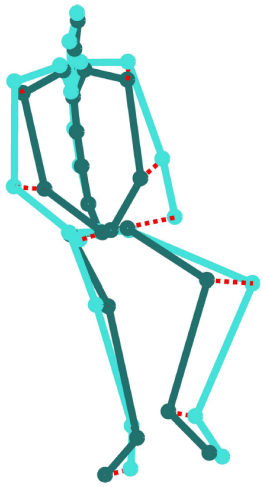


Figure 11: Illustration of absolute measure for full body motion in the study by Van der Zee et al. [27]. The figure shows two poses in shades of blue with the distance between pairs of joints indicated by dashed red lines.

Systems Security Research Lab (S2Lab), King's College London

The S2Lab (`https://s2lab.kcl.ac.uk/`) is part of the Cybersecurity group of the Department of Informatics at King's College London. The lab works at the intersection of program analysis and machine learning for systems security, and more recently on adversarial AI. The lead of the lab is Professor Lorenzo Cavallaro (`https://kclpure.kcl.ac.uk/portal/lorenzo.cavallaro.html`). One recent paper on adversarial AI from this lab is summarised below.

Pierazzi et al. [20] focused on test-time evasion attacks in the so-called *problem space*, where the challenge lies in modifying real input-space objects that correspond to an adversarial feature vector. The authors proposed a formalisation of problem-space attacks, which lays the foundation for identifying key requirements and commonalities among different domains. They identified four major categories of constraints to be defined at design time, including (1) which problem-space *transformations* are available to be performed automatically while looking for an adversarial variant; (2) which object *semantics* must be preserved between the original and its adversarial variant; (3) which non-ML *preprocessing* the attack should be robust to (e.g., image compression and code pruning); and (4) how to ensure that the generated object is a *plausible* member of the input distribution, especially upon manual inspection. Building on their formalisation, the authors proposed a problem-space attack for the Android malware domain which they claimed to overcome the limitations of existing attacks. To evaluate the effectiveness of the proposed attack, they implemented a prototype, available on request (`https://s2lab.kcl.ac.uk/projects/intriguing`). The authors used the DREBIN classifier, based on a binary feature space and a linear SVM, and its hardened variant, Sec-SVM, which requires the attacker to modify more features to perform an evasion. The experiments used a dataset with 170K Android apps from 2017 and 2018, collected from AndroZoo. Results of the experiments showed that thousands of realistic and inconspicuous adversarial applications can be automatically generated at scale, where it takes often less than two minutes to generate an adversarial app. Since the authors shared their codes and the data, further research on defence in the problem space can be conducted by making use of the resources provided by the authors.

Intelligent Systems Lab, University of Oxford

The Intelligent Systems Lab (`http://www.cs.ox.ac.uk/people/thomas.lukasiewicz/isg-index.html`) directed by Professor Thomas Lukasiewicz (`https://www.cs.ox.ac.uk/people/thomas.lukasiewicz/`) is part of the Artificial Intelligence and Machine Learning theme of the Department of Computer Science at the University of Oxford. They have been working on adversarial AI recently. One selected paper is summarised below.

Li et al. [15] proposed a lightweight generative adversarial network for efficient image manipulation using natural language descriptions; examples of

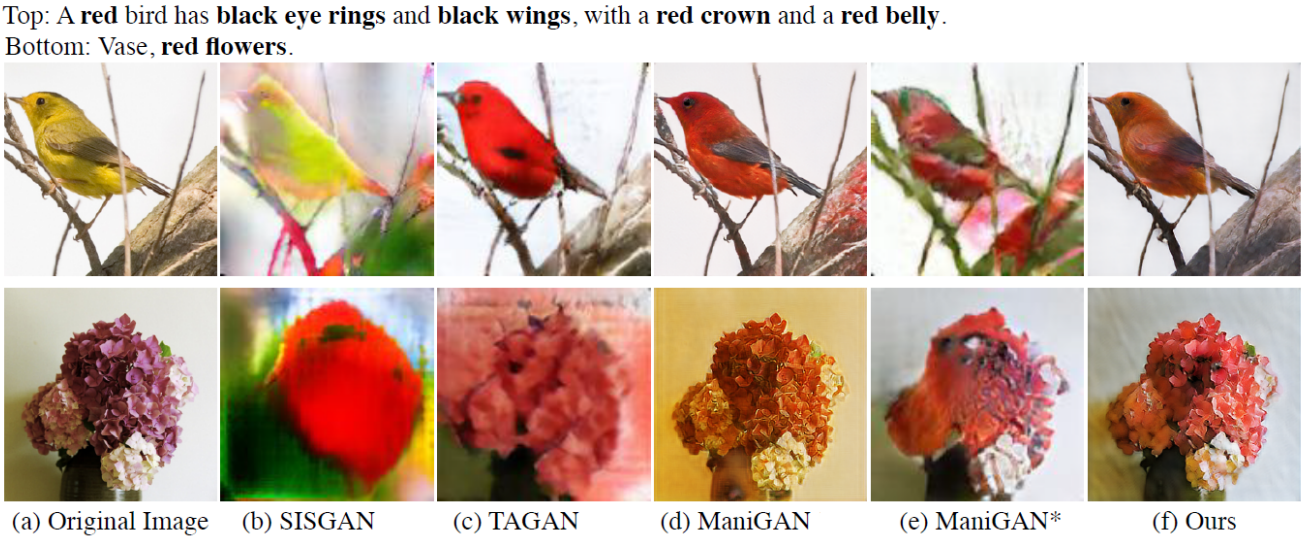


Figure 12: Examples of image manipulation using natural language descriptions from the study by Li et al. [15]. "ManiGAN*" denotes the baseline model for which the authors reduced the number of stages and parameters.

which are shown in Figure 12. More precisely, the authors aimed to semantically modify parts of an image (e.g., colour, texture, and global style) according to user-provided text descriptions, where the descriptions contain desired visual attributes that the modified image should have. They also presented a word-level discriminator to fully explore the information contained in text features and build an effective independent relation between each visual attribute and the corresponding semantic word. The lightweight generator in the system, on the other hand, contains a text encoder, which is a pre-trained bidirectional Recurrent Neural Network (RNN), and two image encoders, which are pre-trained using the Inception-v3 and VGG-16 networks, respectively. The authors evaluated their system using the CUB bird dataset which contains 8,855 training images, 2,933 test images and 10 text descriptions per image, and the COCO dataset which contains 82,783 training images, 40,504 validation images and 5 text descriptions per image. For the evaluation of the system, the authors adopted the Fréchet Inception Distance (FID) as a quantitative measure. Moreover, they conducted a user study in which, for each dataset, they randomly selected 30 images with randomly chosen descriptions. Then, they asked participants to compare two results after looking at the input image, given text and outputs based on accuracy and realism. The 1380 results obtained from 23 participants returned accuracy rates of 65.94% and 77.97%, and realism rates of 57.82% and 67.53% for CUB bird and COCO datasets, respectively. In addition, FID was reported to be lower than the existing ManiGAN solution for both datasets, which is better. The authors shared their codes publicly (`https://github.com/mrlibw/Lightweight-Manipulation`) for further research.

Airbus Centre of Excellence in Cyber Security Analytics, Cardiff University

This centre (`https://www.cardiff.ac.uk/research/explore/research-units/airbus-centre-of-excellence-in-cyber-security-analytics`) was founded as a result of a collaboration between the Cardiff University and Airbus. It covers areas of mutual interest to the Cyber Operations team at Airbus and the Cardiff University, including data science, big data analytics and AI. The centre mainly focuses on the interpretation and effective communication of applied data science and AI methods through interdisciplinary insights into cyber risk, threat intelligence, attack detection and situational awareness. The Director of the centre is Professor Pete Burnap (`https://www.cardiff.ac.uk/people/view/44219-burnap-pete`). One recent paper this centre published on adversarial AI is summarised below.

Anthi et al. [2] investigated how adversarial learning can be used to target supervised models in the context of Industrial Control Systems (ICS). In such context, they used the Jacobian-based Saliency Map Attack (JSMA) to generate adversarial samples and explored classification behaviours of the target models. The authors investigated how such samples can support the robustness of supervised models using adversarial learning by including a random sample of 20% of the generated adversarial data points in the original training dataset and retraining the models. For the experiments, a power system dataset was used that consisted of 55,663 malicious and 22,714 benign data points. It was constructed by using the power system framework implemented by Mississippi State University and Oak Ridge National Laboratory. Results showed that the classification performance of the Random Forest (RF) and J48 classifiers decreased by 16% and 20%, respectively, when adversarial samples were present. In addition, when the models were retrained after adding the adversarial samples to the training data, the RF model had a higher classification performance compared to the J48 model, which means the former is a more robust model towards classifying adversarial samples for all combinations of the JSMA parameters on the given dataset, according to the authors.

Artificial Intelligence Network, Imperial College London

The Artificial Intelligence Network (`https://www.imperial.ac.uk/artificial-intelligence`) spans all faculties of Imperial College London, from Engineering and Natural Sciences to Medicine and the Business School with over 200 academics working in the area of AI. The network covers all AI-related research activity within Imperial College London, including several AI-related research groups and Centres of Doctoral Training. One recent paper on adversarial AI from this network is summarised below.

Liu and Lomuscio [17] introduced a black-box adversarial training algorithm, called *MRobust*, leveraging from the Scale Invariant Feature Transform (SIFT) algorithm. The proposed algorithm used Monte-Carlo Tree Search (MCTS) to generate adversarial examples for adversarial training. Given an arbitrary input to a Deep Neural Network (DNN), MRobust searches small regions around the input that have significant potential to generate adversarial samples. The authors stated that the algorithm does not require access to the internal layers of the DNN, and thus falls in the realm of a black-box adversarial attack. The authors evaluated MRobust on the MNIST and CIFAR10 datasets. By comparing against Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) adversarial attack methods, the results showed that the resulting DNNs synthesised via the proposed method are less susceptible to attack transferability. Furthermore, the authors showed that the proposed method significantly reduced the number of adversarial examples required for adversarial training.

References

[1] Mehmet Aktukmak, Yasin Yilmaz, and Ismail Uysal. 2019. Quick and Accurate Attack Detection in Recommender Systems through User Attributes. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, 348–352. https://doi.org/10.1145/3298689.3347050

[2] Eirini Anthi, Lowri Williams, Matilda Rhode, Pete Burnap, and Adam Wedgbury. 2020. Adversarial Attacks on Machine Learning Cybersecurity Defences in Industrial Control Systems. , 9 pages. arXiv:2004.05005 [cs.LG] https://arxiv.org/abs/2004.05005

[3] Zachary M. Carr, Anne Solbu, and Mark G. Frank. 2019. Why Methods Matter: Approaches to the Study of Deception and Considerations for the Future. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 267–286. https://doi.org/10.1007/978-3-319-96334-1_14

[4] Huiyuan Chen and Jing Li. 2019. Adversarial Tensor Factorization for Context-Aware Recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, 363–367. https://doi.org/10.1145/3298689.3346987

[5] Yuxin Chen, Huiying Li, Shan-Yuan Teng, Steven Nagels, Zhijing Li, Pedro Lopes, Ben Y. Zhao, and Haitao Zheng. 2020. Wearable Microphone Jamming. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3313831.3376304

[6] Konstantina Christakopoulou and Arindam Banerjee. 2019. Adversarial Attacks on an Oblivious Recommender. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, 322––330. https://doi.org/10.1145/3298689.3347031

[7] Rachel E. Dianiska, Daniella K. Cash, Sean M. Lane, and Christian A. Meissner. 2019. The Reciprocal Nature of Lying and Memory: Memory Confabulation and Diagnostic Cues to Deception. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 347–365. https://doi.org/10.1007/978-3-319-96334-1_18

[8] T. Docan-Morgan. 2019. *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan. https://doi.org/10.1007/978-3-319-96334-1

[9] Linda Marjoleine Geven, Gershon Ben-Shakhar, Merel Kindt, and Bruno Verschuere. 2019. Memory Detection: Past, Present, and Future. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 367–383. https://doi.org/10.1007/978-3-319-96334-1_19

[10] Edward R. Griffor, Christopher Greer, David A. Wollman, and Martin J. Burns. 2017. *Framework for Cyber-Physical Systems: Volume 1, Overview*. Technical Report NIST SP 1500-201. National Institute Standards and Technology (NIST). https://doi.org/10.6028/NIST.SP.1500-201

[11] Ihsan Gunes, Cihan Kaleli, Alper Bilge, and Huseyin Polat. 2012. Shilling Attacks Against Recommender Systems: A Comprehensive Survey. *Artificial Intelligence Review* 42 (2012), 767––799. https://doi.org/10.1007/s10462-012-9364-9

[12] Yasha Iravantchi, Mayank Goel, and Chris Harrison. 2020. Digital Ventriloquism: Giving Voice to Everyday Objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–10. https://doi.org/10.1145/3313831.3376503

[13] Bennett Kleinberg, Arnoud Arntz, and Bruno Verschuere. 2019. Detecting Deceptive Intentions: Possibilities for Large-Scale Applications. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 403–427. https://doi.org/10.1007/978-3-319-96334-1_21

[14] Samuel Larner. 2019. Formulaic Sequences as a Potential Marker of Deception: A Preliminary Investigation. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 327–346. https://doi.org/10.1007/978-3-319-96334-1_17

[15] Bowen Li, Xiaojuan Qi, Philip Torr, and Thomas Lukasiewicz. 2020. Lightweight Generative Adversarial Networks for Text-Guided Image Manipulation. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*. 12. https://proceedings.neurips.cc//paper/2020/hash/fae0b27c451c728867a567e8c1bb4e53-Abstract.html

[16] Shang Li and Xiaodong Wang. 2014. Quickest Attack Detection in Multi-AgentReputation Systems. *IEEE Journal of Selected Topics in Signal Processing* 8, 4 (2014), 653–666. https://doi.org/10.1109/JSTSP.2014.2309943

[17] Yi-Ling Liu and Alessio Lomuscio. 2020. MRobust: A Method for Robustness against Adversarial Attacks on Deep Neural Networks. In *2020 International Joint Conference on Neural Networks*. IEEE, 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207354

[18] Eric Mac Giolla and Pär Anders Granhag. 2019. True and False Intentions: A Science of Lies About the Future. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 385–401. https://doi.org/10.1007/978-3-319-96334-1_20

[19] David M. Markowitz and Jeffrey T. Hancock. 2019. Deception and Language: The Contextual Organization of Language and Deception (COLD) Framework. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 193–212. https://doi.org/10.1007/978-3-319-96334-1_10

[20] Fabio Pierazzi, Feargus Pendlebury, Jacopo Cortellazzi, and Lorenzo Cavallaro. 2020. Intriguing Properties of Adversarial ML Attacks in the Problem Space. In *2020 IEEE Symposium on Security and Privacy*. IEEE, 1332–1349. https://doi.org/10.1109/SP40000.2020.00073

[21] John H. Powers. 2019. Discursive Dimensions of Deceptive Communication: A Framework for Practical Analysis. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 167–191. https://doi.org/10.1007/978-3-319-96334-1_9

[22] Ilia Shumailov, Xitong Gao, Yiren Zhao, Robert Mullins, Ross Anderson, and Cheng-Zhong Xu. 2019. Sitatapatra: Blocking the Transfer of Adversarial Samples. , 9 pages. arXiv:1901.08121 [cs.LG] https://arxiv.org/abs/1901.08121

[23] R. Weylin Sternglanz, Wendy L. Morris, Marley Morrow, and Joshua Braverman. 2019. A Review of Meta-Analyses About Deception Detection. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 303–326. https://doi.org/10.1007/978-3-319-96334-1_16

[24] Chris N. H. Street, Jaume Masip, and Megan Kenny. 2019. Understanding Lie Detection Biases with the Adaptive Lie Detector Theory (ALIED): A Boundedly Rational Approach. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 227–247. https://doi.org/10.1007/978-3-319-96334-1_12

[25] Jiaxi Tang, Hongyi Wen, and Ke Wang. 2020. Revisiting Adversarially Learned Injection Attacks Against Recommender Systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*. ACM, 318–327. https://doi.org/10.1145/3383313.3412243

[26] Carlos E. Tejada, Raf Ramakers, Sebastian Boring, and Daniel Ashbrook. 2020. AirTouch: 3D-printed Touch-Sensitive Objects Using Pneumatic Sensing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–10. https://doi.org/10.1145/3313831.3376136

[27] Sophie Van der Zee, Ronald Poppe, Paul J. Taylor, and Ross Anderson. 2019. To Freeze or not to Freeze: A Culture-Sensitive Motion Capture Approach to Detecting Deceit. *PLOS ONE* 14, 4 (2019), 1–18. https://doi.org/10.1371/journal.pone.0215000

[28] Emma J. Williams and Kate Muir. 2019. A Model of Trust Manipulation: Exploiting Communication Mechanisms and Authenticity Cues to Deceive. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 249–265. https://doi.org/10.1007/978-3-319-96334-1_13

[29] John Paul Wilson and Nicholas O. Rule. 2019. The Many Faces of Trustworthiness: Accuracy and Inaccuracy in Predicting Deception from Facial Appearance. In *The Palgrave Handbook of Deceptive Communication*. Palgrave Macmillan, 429–442. https://doi.org/10.1007/978-3-319-96334-1_22

[30] Tuo Yu and Klara Nahrstedt. 2019. ShoesHacker: Indoor Corridor Map and User Location Leakage through Force Sensors in Smart Shoes. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3, Article 120 (2019), 29 pages. https://doi.org/10.1145/3351278