The Changing Nature of Threats for Facial Recognition Systems
Authors

This White Paper summarises findings documented within the three technical briefings that accompany this paper: Attacks Against Facial Recognition Systems; Threat Models for Face Recognition Systems; Benchmarking Biometrics for Identity Systems in Developing Countries, authored by:

Professor Carsten Maple, Turing Fellow, Project Principal Investigator, and Professor of Cyber Systems Engineering with Institute partner University of Warwick

Dr Gregory Epiphaniou, Associate Professor in Security Engineering, University of Warwick

Dr Roberto Leyva, Research Associate, University of Warwick

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation [INV-001309]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript.

The Institute is named in honour of Alan Turing, whose pioneering work in theoretical and applied mathematics, engineering and computing is considered to have laid the foundations for modern-day data science and artificial intelligence. It was established in 2015 by five founding universities and became the United Kingdom’s (UK) National Institute for Data Science and Artificial Intelligence. Today, the Institute brings together academics from 13 of the UK’s leading universities and hosts visiting fellows and researchers from many international centres of academic excellence. The Institute also liaises with public bodies and is supported by collaborations with major organisations.
# Table of Contents

Purpose.................................................................................................................................................. 4
Introduction ............................................................................................................................................. 5
Executive Summary ................................................................................................................................. 6
Facial Recognition & Developing Economies .......................................................................................... 7
A New Architecture .................................................................................................................................. 8
The Evolving Threat Landscape .............................................................................................................. 9
Taxonomy for Modelling Threats ............................................................................................................ 11
  1.1 Computing Resources ...................................................................................................................... 11
  1.2 Biometric Samples ............................................................................................................................. 12
  1.3 Models .............................................................................................................................................. 12
Architectural Context .............................................................................................................................. 13
Looking Ahead ......................................................................................................................................... 15
Purpose

The **Trustworthy Digital Infrastructure for Identity Systems** project is working to advance good technical choices as the digitisation of economies transforms how people are identified, included, and supported in society. The project team is producing tools and guidance for the assessment and development of systems according to six facets of Trustworthiness: security, privacy, ethics, resilience, robustness, and reliability. Their mission incorporates an objective to advance understanding of requirements particular to lower income countries as many embrace digital and biometric technologies to establish modern foundational identity systems and in some cases support citizens that lack any form of legal documentation. The body of work summarised in this paper is offered to support trustworthy design, maintenance and governance of identity and other systems using facial recognition technologies. A prerequisite understanding or technical systems knowledge is not required of the reader.
Introduction

Facial recognition technologies are becoming a pillar of global commerce and security, embraced for their efficacy and availability in advancing modern ways of transacting, and facilitating automated and remote access to goods and services. Alongside other biometrics such as fingerprint and iris scans, these technologies are becoming integral to the pursuit of economic, humanitarian, and societal support, deployed by major industries, and governments alike as they make their way into foundational national identity systems alongside the systems that facilitate access to banking services, border crossings, online commerce, facilities access, and much more. The proliferation of these technologies has also fuelled questions around whether they are ethical, given acknowledgement of algorithmic bias, functional advances within the technology such as automated emotion detection and analysis, and their growing role as a gateway to resource and entitlement.

For organisations, biometrics are providing a cost-effective alternative to the issuance of identity or membership cards, while facial recognition specifically provides a non-contact opportunity to facilitate remote access to their resources and services, a rising priority in the wake of COVID-19. Unlike a card or document, biometric credentials cannot be easily altered or routinely lost, and they are universal presenting significant opportunity to recognise and serve people who currently lack or find themselves separated from formal identity documents. Turing researchers working within the Trustworthy Digital Infrastructure for Identity Systems project are examining the proliferation of biometric data and what its use must teach us about emerging risks. Given the growing scope of use, cyber security risks and their potential impact are developing at pace. Unlike other forms of identity credentials that people rely upon to access services and resources, compromised biometrics cannot be revoked: If someone’s biometric is exploited there is a significant risk that they lose the ability to use it. Delving into developments with facial recognition technologies, this paper highlights the Turing’s examination of how new vulnerabilities are developing as processes and systems evolve to incorporate many organisations and be augmented by machine-learning technologies. Emerging areas of vulnerability are increasingly exploited as attackers develop a sophisticated understanding of the systems and appreciate their growing value.

Part of an ongoing mission to advance understanding of how risks are evolving with the development of national foundational digital identity systems, and explore opportunities for their governance, researchers at the Alan Turing Institute, propose a new reference architecture and taxonomy of systems security risk to support the development of trustworthy facial recognition systems. The work draws from a comprehensive analysis of current trends in exploits, which among its many observations, revealed significant vulnerability in the common tendency to trust the initial reference data collected as people registered for services. It complements the Institute’s recently published guide: Understanding bias in facial recognition technologies, which outlines the historical and technical development influences that have led to algorithmic bias in facial recognition technologies. While some of the threats and system vulnerabilities discussed in this paper may facilitate risks linked to algorithmic bias, the focus is on elevating understanding of the opportunity for, and the impact of the compromise of the systems, models, and data within them.
**Executive Summary**

Biometrics refer to the uniquely identifying biological information of individuals. A biometric system examines and stores this unique information and facilitates comparison with these features when the individual claims access to a resource, service, or facility. As the most common and familiar biometric feature, the face is widely used in many scenarios. Face recognition systems compare the difference between an input image, and an image or attribute extracted from a stored image. Generally, facial recognition systems consist of three stages: face detection, feature extraction, and face recognition. The face detection checks whether there is a face; feature extraction is the next step to extract the features of the face, such as patterns or landmarks. The face recognition step is provision of a decision for verification or identification tasks.

Analyses of the current threat landscape reveal that attacks on these systems are evolving with the convenience of unsupervised registration using personal mobile phones or devices, and the online availability of registrants’ images on social media and other sites that ease impersonation. Levels of developing technical sophistication are also leading to the proliferation of fake imagery and attacks aiming to perturb the processes and machine learning models driving systems’ operations.

The design of a trustworthy facial recognition system relies on identifying how the system can be targeted to compromise its performance and security, and the privacy of the people relying on it. Offering a new Authentication Reference Architecture (ARA), researchers working at The Alan Turing Institute present a model for analysing threats present at each stage of the authentication process. Departing from current convention, the registration process is recognised as part of the authentication architecture to enhance the scope of systems security risk modelling. Following the recently proposed European Union Agency for Cybersecurity (ENISA) Risk Management and Information Security Management Systems RM/ISMS framework, the team conducted a comprehensive strategic review of the threat landscape to elaborate the architecture and devise a taxonomy for identifying threats within its context. The review revealed vulnerability in processes for the collection, storage, development of templates, and interpretation of the reference data stored in the registration components of a system which can be exploited to undermine the authentication process to various aims.

In all, thirteen types of attack have been identified and mapped to component vulnerability in the architecture offering a current view of the most prominent groups of the reported attacks, their nature, and the threats they represent. Such an approach facilitates end-to-end visibility for risk modelling within the context of the desirable qualitative characteristics of the services offered by an organisation, such as the availability of the service, assurance of reliable and fair access, preservation of data confidentiality and integrity, and other requirements.
Facial Recognition & Developing Economies

An identity system is a critical element of modern commerce and societal governance around the world. Many lower income and developing economies have lacked a modern identity system and many support significant elements of their population that lack any type of official documentation, such as a birth certificate.

Disadvantages range from barriers to individual access to public services or opportunities that could be facilitated by the ability to open a bank account to a nation’s capacity overall to thrive in increasingly global, digitally enabled marketplaces. Given this, legal identity is considered core to development, recognised by the UN within its sustainable development goals, while digital forms of identity are presenting significant opportunities for many of these countries to leap forward. Digital biometric identification technologies using facial fingerprint, or iris recognition are rapidly advancing to help developing countries improve the efficiency of documenting individual identity and authenticating claims of identity when accessing services. They are also advancing with use by humanitarian initiatives, including the UN’s World Food programme and UNHCR, and Red Cross and Red Crescent Societies.

Biometric systems are being embraced as cost-effective solutions for identity and service management, more efficient than traditional document-based identity systems, while the global impact of the COVID-19 pandemic accelerated emphasis on remote capabilities.

Within its technical briefing Benchmarking Biometrics for Identity Systems in Developing Countries, the Institute’s project team note that the African and Middle Eastern biometrics market is forecast to grow at an annual rate of 21%, with fingerprint and face recognition being the two most popular biometrics used. Researchers analysed the relative costs, ease of use, and availability of biometric modalities, alongside other characteristics offering some explanation for the deference to facial recognition and fingerprints. The briefing also highlights that many biometric systems and solutions available originated within developed economies for different aims, designed for law enforcement or border control, rather than broad service delivery.

Facial recognition has matured to be used across a wide variety of use cases making them more cost accessible. They are also developed to varying levels of sophistication and robustness and are increasingly targeted with their expanding use. Design and risk analysis requires careful consideration of the context in which the system is being deployed. India’s ambitious Aadhaar program is the largest national digital identity system, that employs facial recognition alongside other modes of biometric authentication, with The Unique Identification Authority of India (UIDAI) collecting iris scans of both eyes, ten fingerprints, and a digital face image from each enrollee. Many people, especially the elderly and manual labourers, have worn fingerprints and or damaged eyes. The methods combined together ensure accessibility for these individuals.

In assessing the appropriateness of biometrics for identity systems in developing countries, the project team examine a number of parameters, with many of these considerations varying across different economic settings, such as permanence, whether the biometric would be affected by age or use; collectability, how easily it can be acquired; acceptability, whether it would be accepted by country’s public for widespread use; the cost of developing the system; and it’s stability across variances in environments, such as...
access to power and connectivity. Their analysis, which provides an evaluation of the applicability of different biometric modalities in identity systems concludes that the face recognition system is suitable for most developing countries: the face recognition system is convenient for users to understand and cooperate; there is no physical contact. Cameras as sensors are economical to purchase and maintain, while markets for these are mature, making the hardware and software components less costly than some other biometric systems. However, performance is low as facial features change over a lifetime, expression, illumination, pose, and occlusion affects the recognition performance while the captured faces are sensitive to privacy abuse. The system should be combined with human inspection or other biometric systems in high-security circumstances.

In assessing the changing nature of threats to facial recognition systems, the project team delve into specific systems vulnerabilities, adding to the repository of guidance and resources available on the design of systems for foundational identity programmes, including the World Bank’s ID4D toolkits, and the ID4Africa knowledgebase, both from organisations that advise the project through its International Advisory Board.

A New Architecture

The World Bank Group describes four stages in the process of establishing a person’s identity and then using it in later transactions referring to it as the Identity Lifecycle: registration, issuance, use, and management. The Identity Lifecycle emphasises that systems are designed to support a process that starts when a person first registers and their identity is created; continues with authentication of that identity and is maintained with updates to their attributes and credentials over time, ending when an identity record is retired, after death, or for other reason.

There is a need to ensure that systems and processes are secure, and that the data, templates, and algorithms that manage the data retain their integrity, throughout this lifecycle. Proper design and management of the systems, including facial recognition technologies, relies on an up to date understanding of what can be targeted by an attacker and to what aim. While there are many frameworks for assessing this, they place particular importance on the operational presentation stage when the system is being used to authenticate a claim of access. This perhaps reflects the fact that the most prevalent attacks are presentation attacks aimed at impersonating a user (see The Evolving Threat Landscape below).

After conducting a comprehensive review of the most prominent groups of the reported attacks on facial recognition systems, the project team identified threats to systems and processes that extend beyond this operations stage. Biometric reference modification threats, for example, can be invoked at both the registration and authentication stages to alter databases with the goal of impersonating or inflicting rejection of a user. Training data for a system’s machine learning model, which is the core of a facial recognition system, can be attacked to modify the algorithm in charge of deciding whether the processed sample matches an existing, opening the door for poor quality images to fool the system. It can also be the case that the attacker adds to reference databases or alters raw images and feature templates without having to compromise operational systems. A detailed overview of how such attacks threaten facial recognition systems is documented in the technical briefing Threat Models for Face Recognition Systems.

Analysis of the threats led the Institute’s project team to elaborate a reference architecture incorporating the registration stage as an essential component of the authentication process (see figure below). The authentication process is invoked in two stages of the lifecycle, first
to verify or confirm an identity before a user credential is issued, and then to verify as credentials are used. The new architecture increases visibility of system vulnerabilities, advancing understanding of the threats to the system components, the disparate registration and authentication processes, and their interactions.

The Evolving Threat Landscape

Generally, facial recognition systems consist of three stages: face detection, feature extraction, and face recognition. The face detection checks whether there is a face; feature extraction is the next step to extract the features of the face, such as patterns or landmarks. The face recognition step is provision of a decision for verification or identification tasks.

Systems vary, including the levels of sophistication developed within the machine learning algorithms considered to be the core of recognition functionality as different use cases prioritise different requirements. Basic elements will include reference data, which can be the initial raw data collected or templates created from this data, machine learning models and scoring parameters for comparing users to reference data, devices and sensors that capture the images, communications signals, and storage, all of which are becoming targets of attack. Attacks can target the whole system or a discrete element depending on the goal. An attack may, for example, seek simply to reduce scoring thresholds driving the comparison model to increase tolerance for incorrect images.

Further, vulnerability to attack may not necessarily be a system design concern alone: Hacktivists, who seek to protest through disruption rather than commit fraud or theft through impersonation, enable the fraudsters when they release details of their exploits or registrants as part of their protest.

The Institute’s project team conducted a comprehensive review of records from 2012 – 2021 to provide a state-of-the-art review of how attacks are developing toward facial recognition systems. A full technical briefing containing their review Attacks Against Facial Recognition Systems: A State of the Art Review groups them into five categories based on their nature.
and target area of vulnerability as illustrated in the figure below, and outlines the levels of sophistication in terms of systems and technical knowledge needed to perpetrate them. The work highlights that some of the most severe threats do not require a high-level of technical sophistication on the part of the attacker. Advances in high-quality printing and video technology can be employed to fool systems. Video or computing motion techniques can also be used to defeat liveness testing, where a system is designed to confirm a user is alive to authenticate access, illustrating the arms race that is developing with advances in facial recognition technology.

The project team observed that the most persistent threats come from presentation attacks. This is in part evolving with the convenience of unsupervised registration using personal mobile phones and devices and the growing online availability of registrants’ images on social media and other sites. Presentation attacks on facial sensors have garnered broad interest given the ease at which a face artefact can be generated, and the low level of sophistication needed: it does not require knowledge about the operational details of the biometric system. A recently discovered attack, for example, requires placing a sticker on the top of the head where the system considers the patch as part of the face. Morphing, which produces fake images from genuine ones, can include that of an impostor and can generate a morphed image of high quality that exhibits similar visual characteristics to the registrants.

It is also possible to train a machine-learning model to approve subjects based on carefully selected patches that contain the key landmarks in the model, rather than the entire face. Model threats aimed at the system’s features processing are the second most prevalent attacks and are showing an exponential trend as they become more feasible with the ever-growing need for pre-processed data or unsupervised reference collection. Machine learning models and the templates they generate can be targeted in many ways and methods of attack can be combined resulting in the corruption of a system’s training data, acceptance of low similarity scores, or the creation of adversarial biometric samples that are used in the recognition stage. Template attacks target the processed data from the input biometric

---

**Attacks by proportion groups from 2018–2021**

- **Template:** intermediate sophistication, requires an understanding of the features functionality. 3.50%
- **Hardware:** advanced sophistication, requires an understanding of internal circuits’ functionality and security properties during their operation. 3.60%
- **Transmission/Storage:** advanced sophistication, involves knowledge of the system components, users, and experience to identify weaknesses. 8.50%
- **Presentation:** low sophistication, although artistic skill may be deployed. 29.60%
- **Model:** intermediate sophistication. 54.80%
samples to be stored and used for authentication purposes to undermine the model's operation. Basic techniques include adding slight perturbations and data injection to either the registration data or the authentication samples. A recently observed trend uses Generative Adversarial Network (GAN) models to create new data instances that resemble training data. It is also possible to reveal secured or even deleted original reference data (despite the fact, that it is not actively used during the authentication process) with a sophisticated membership inference attack whereby training data can be reconstructed by observing its output. Once identities are revealed, an attacker can produce presentation attacks later.

Although less frequent, hardware threats pose a serious risk when considered in the context of the growing use of distributed computing for mobile devices. Security flaws in the device provide an open door to the spread of malware (software of varied levels of sophistication intentionally designed to cause disruption or gain unauthorised access to a system), a vulnerability that can lead to the override of the image capture process or hijacking of the system access given to the device. Hardware threats can compromise imaging sensors or bypass them for unseen sensors to send fabricated/stored images. Hardware is also particularly weak against side-channel attacks whereby the device’s access to the system is overridden, and an attacker takes control of the device to produce biometric samples from non-trusted sources.

Transmission/Storage threats are present during information processing, particularly when raw, rather than tokenised data is processed, with the potential to impact all data, including the models, raw images, scoring parameters and more. Man-in-the-middle attacks, for example, can potentially replace the end-user with communication established with an impersonator: if the attacker manages to gain administrator-level credentials, the whole system is compromised.

Taxonomy for Modelling Threats

Following the recently proposed European Union Agency for Cybersecurity (ENISA) Risk Management and Information Security Management Systems RM/ISMS framework (which has at its core the requirement to identify threats) the teams’ analysis of the reported attacks allowed them to elaborate a taxonomy of threats across three areas: the Computing Resources, which is the infrastructure used to process the data and all operations; Biometric Samples, which concerns the input data given to the system, and Models which are the core of the authentication process. Each of these high-level threat areas were further broken down into types for each area, reflecting their nature, whether they profile or facilitate access, are specific or indiscriminate, disrupt or violate, commit fraud, breach confidentiality and the like, as detailed in the figure and following section.

1.1 Computing Resources

Profiling: we can split this type into four main groups covering threats that concern reconnaissance in which the victim is identified; fraud where there is a personal or economic gain from the user or service providers; sabotage which implies destroying the system; and eavesdropping where the attacker obtains important information from all the system actors.

Access: where threats concern gaining access to remote sources via the authentication process, for example server logins or physical access to equipment or facilities. These threats can impact: availability where systems may not provide access to the desired resources; confidentiality where the user data is exposed to the public; and integrity where the user data is subject of manipulation.
1.2 Biometric Samples

*Influence:* three main areas are generative attacks, where biometric samples are fabricated using machine learning models; spoofing attacks where authentic samples are used, and intrusive attacks where tampered samples are used somehow for impersonation.

*Specificity:* two types of threats are those aimed at one subject and indiscriminate ones, impersonation attacks that are not aimed at a particular subject.

1.3 Models

*Influence:* this group comprises causative threats where adversaries can change the distribution of training data inducing parameter changes of learning models when retraining, resulting in a significant decrease of the performance of classifiers in subsequent classification tasks; and exploratory threats that aim at causing misclassification with respect to adversarial samples or to uncover sensitive information from training data and learning models.

*Violation:* we can categorise these into integrity threats that aim to increase the false negatives of existing classifiers when classifying harmful samples; availability threats that will cause an increase of the false positives of classifiers concerning benign samples; and privacy violation threats that aim at obtaining sensitive and confidential information from training data and learning models.

*Specificity:* this group includes targeted threats aiming to reduce the performance of classifiers on one particular sample or one specific group of samples, and indiscriminate threats that cause the classifier to fail indiscriminately on a broad range of samples.
Architectural Context

Proposing a new Authentication Reference Architecture, the project team elaborates on the established architecture published by the US National Institute of Standards and Technology (NIST) to incorporate the registration stage. The figure below illustrates how this offers visibility to address vulnerabilities that exist during the registration stages and the generation of biometric samples. This architecture decomposes and allows for analyses of the registration and authentication processes separately to advance understanding of the potential threats presented by different attacks.

In all, 13 types of threat have been identified and mapped to how they may exploit component vulnerabilities across the architecture:

1. Presentation: impersonation takes place using varied techniques.
2. Device Override: an attacker overrides the capture devices.
3. Sample Modification: to impersonate and/or bypass authentication during signal processing or quality assessment of the sample image.
4. Signal Processing Override: compromise of the feature extraction process or sample modelling.
5. Probe Modification: sending intercepted or tampered features templates.
6. Comparator Override: compromise of the algorithm in charge of deciding whether the processed sample matches an existing user.

7. Database Override: compromise of the raw images, models, and feature templates.

8. Biometric Reference Modification: incorrect or compromised samples sent to the database during transmission.

9. Score Modification: modification of comparison criteria to alter outcomes or performance.

10. Decision Engine Override: modifying the scoring criteria to lower security levels and weaken decisions.

11. Decision Modification: directly modifying or bypassing the comparison decision.

12. Assessment Engine Override: the quality assessment of the registration stage is avoided or not performed properly.


The diagram below depicts the system’s architecture in the presence of these attacks, showing for example that some model threats to the assessment engine are present in the registration stage only. Device override threats, where the attacker may override the capture devices and samples are never actually taken from the intended device are present both in the registration and authentication stages of the architecture, as are biometric sample modification threats, and signal processing override threats. The latter, impacting the model used to interpret information from the captured image to create the feature template may be corrupted in many ways.

*System architecture in the presence of attacks*
Looking Ahead

In elaborating a new Authentication Reference Architecture for facial recognition systems, greater visibility is given to emerging areas of vulnerability as we acknowledge the significant impact of evolving risks to civil society. Mitigating threats within the registration stage, which has not generally been considered within authentication architecture, has particularly emerged as the cornerstone to mitigating the most persistent type of threat, the presentation attack. The registration stage has also been significantly implicated in the more sophisticated model and transmission threats, further indicating its importance as a consideration in the design of more resilient facial recognition.

Overall, systems designers and risk managers can work with a more comprehensive view than conventionally understood of how evolving threats are likely to exploit these systems and ultimately compromise the organisations, services, and individuals they support. This white paper and the supporting technical briefings were published to support their evaluation of developing systems requirements and that of the practitioners commissioning both functional and foundational identity systems. The Trustworthy Digital Infrastructure for Identity Systems project offer it as a reference for business and technical stakeholders encompassing a state-of-the-art overview of the threat and vulnerability landscape, and a new architecture and taxonomy for modelling threats to facial recognition systems.

Moving forward, the Trustworthy Digital Infrastructure for Identity Systems project is delving deeper into the model threats as they continue to develop exponentially, facilitated by reliance on third-party processed data. This will contribute to the Taxonomy of Threats and Architectural Context being advanced by the project team as tools to enhance capabilities in risk modelling of facial recognition systems and the trustworthiness of the systems and organisations that deploy them.