The Alan Turing Institute

Technical Briefing

Facets of Trustworthiness in Digital Identity Systems

Authors

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

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

Dr Nagananda Kyatsandra Gurukumar, Research Associate, The Alan Turing Institute

The authors would like to thank the members of the project's International Advisory Board for their advice and comments which have informed the development of this work.

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 Turing 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 Turing also liaises with public bodies and is supported by collaborations with major organisations.

The Alan Turing Institute

British Library

96 Euston Road

London

NW1 2DB

Table of Contents

Purpose 4
Executive Summary 4
1. Background 6
1.1 Why trustworthy digital ID systems? 7
1.2 Trustworthiness principles and trust modelling 7
1.3 Trusted digital identity systems 8
2. Trustworthiness Facets, Attributes and Features 11
2.1 Pillars (Facets) 11
2.2 Attributes 12
2.3 Features and Mechanisms 13
3. CONCLUSION 27
Bibliography 28
Annex A: Further reading 38
ANNEX B: ABREVIATIONS 40
ANNEX C: Referenced Standards & Regulations 41

Purpose

This Technical Briefing presents the first iteration of a new framework to be developed as a resource for and in consultation with governments, humanitarian organisations and industry stakeholders that are advancing digital identity systems. It marks the beginning of a process of consultation and is part of the Trustworthy Digital Infrastructure for Identity Systems project. The framework is to be developed under a creative commons license in response to growing use of digital identity in modern society, and particularly the influence of governments in the advancement of digital identity programmes.

Executive Summary

This work presents the different pillars of trustworthiness for digital identity management systems. We specifically consider the *trustworthiness* of these systems, rather than whether or not such systems are *trusted.* The latter is an important area that informs our work but is not the subject of this work. Whereas trusted can be defined as a *belief* in the integrity, ability or character of an entity, trustworthiness of an entity regards the extent to which it is *deserving* of trust.

The emphasis in this work is placed on the various features and mechanisms for each of the dimensions that can be used to capture the trustworthiness assurance levels (TAL) of identity-related functions. This presentation outlines the preliminary assurance features and mechanisms in each of the pillars *of security, privacy, ethics, robustness, reliability and resiliency* in digital identity systems (often referred to as electronic identity systems and denoted by EIDS, as in here).

This is the preliminary step in identifying and elaborating the core principles of a trustworthiness framework that can be utilised for establishing the different assurance levels of an EIDS in terms of its system components, information system flows, physical and logical processes, and information systems.

These pillars can be used to assess the trustworthiness of EIDS in terms of their functions, processes and specific environmental and regulatory conditions. To establish the trustworthiness of an EIDS, we identify the requirements in each pillar, and the mechanisms and features that may provide scales to measure EIDS' trustworthiness. This approach can provide a holistic understanding of the current level of perceived vs achieved trustworthiness for any given EIDS and assist existing efforts to expand assurance efforts across all six pillars.

Through the clear identification and quantification of its features, our approach can be used as a means to obtain the degree of confidence that the EIDS satisfies certain trust requirements while supporting its mission and functions in a secure and resilient manner that considers usability at all levels. By addressing these pillars, mechanisms and features, we provide the initial base for the future development of TAL to enable perceived trustworthiness estimation and assessment of EIDS systems.

Introduction

Understanding the different trust and security requirements in Electronic Identity Management Systems EID(M)S is an essential step towards their appropriate, secure and independent use and functions. The ability to ensure trustworthy operations in users' identity lifecycle management is often more than the narrow scope of security requirements and regulatory compliance. This is because technology in ID management is evolving faster than regulatory and security requirements in that space. Thus, the ability to design and develop appropriate risk-based approaches requires trust-related requirements embedded in the process.

Existing digital ID assurance frameworks and technical security standards often overlook the importance of trust requirements and facets when developing assurance levels for the certification or accreditation of EIDS systems. The emphasis is usually placed on establishing linkages between privacy and security controls with a sole purpose to enforce security and privacy requirements in different stages within the ID management lifecycle. This seems to promote closer collaboration between various entities controlling the security and privacy aspects (especially in federated EIDS deployments). There are different ways to classify and categorise the privacy dimensions. In our work, we define communication, and processing as the main privacy dimensions to be examined. This is after an exhaustive search in the published literature and discussion with our stakeholders. We acknowledge the fidelity of the data exchanged and shared within and across different EIDS as an essential aspect for all the privacy dimensions. Indeed, data fidelity is an integral component in the trustworthiness of EIDS. The set of mechanisms that influence these systems' decision-making relates to the credibility of data processing, rendering the latter an essential factor to their success in terms of broader acceptance and adoption. The cost of poor quality and untrusted data leading to erroneous decisions has a significant financial, security and productivity impact with potentially dramatic effects on the trustworthiness of these systems. Elements such as resilience and reliability of these systems are also investigated separately regarding the assurance they offer in EIDS components and information systems and services governing their functions.

We argue that considering both resilience and reliability, alongside other components, are essential to understand how entities can achieve trustworthy EIDS systems. This will become more important with the advent of Self-Sovereign Identity (SSI) where users will have full access to the management of their ID and technology that can potentially automate access to services and execution of ID-related processes. Also, still, there is no framework that ensures trusted operations in SSI with existing efforts mainly focussing on security ramifications only. This briefing presents the key pillars of trustworthiness in an attempt to sketch a framework that can be used to assess the capability of an EIDS system to carry out trusted operations against a family of properties defined for that purpose.

1. Background

From an information security perspective, the concept of trust is often linked to the ability of a system to behave in secure and predictable manner under different operational conditions. This ability is also linked to the satisfaction of specific security requirements that allow the components of an EIDS system for example, to resist errors and disruption while ensuring its secure operations often described as *security capability*.

NIST defines trust as relative to the security capability of the system. This is due to the fact that trust at a system level can be measured (subjectively) by the complex interactions amongst different trusted entities within a system. However, these interactions between secure and trusted systems might not result in trustworthy operations because the combination of mutually reinforcing security controls are only a small part of the confidence that users can build and preserve for these systems.

While the *trust* can be considered to be the *confidence* or *belief* one entity holds about the integrity, ability or character of another, trustworthiness of an entity regards the extent to which an entity is *deserving* of trust. Developing for trustworthiness represents a step change in current influences, advancing new opportunities to reduce the reliance on an assumption of trust in a system and its underlying organisations with verifiable arguments around whether claims made warrant being trusted.

Trustworthy EIDS systems are expected to operate in a certain level of acceptable risk despite environmental disruption and users place their trust on every aspect of these systems including its security-related attributes. In that respect, we approach the concept of trustworthiness in EIDS systems not only from their existing security capabilities and assurance levels associated with it, but also consider additional pillars that are influenced by these complex interactions (such as physical, technical, operational and so forth).

The process of verifying an identity attribute for authentication purposes is at the core of EIDS systems and often combines organisational, technical and legal functions to collect and manage identity information. Therefore ensuring the operation of the technology used to execute these processes follows the EIDS's mission and objectives, is an essential factor that influence owners' and users' perception of trust on them. These systems often amalgamate different technologies that co-exist in the EIDS ecosystem (such as service access and provision; authentication and identity mapping) and there needs to be evidence that these technologies and systems work and function as intended.

These systems are often subject to different policies and procedures. The way these policies are enforced and monitored can be subject to internal and external scrutiny in the form of regulatory compliance audits. However, without achieving a certain degree of "openness" of these systems and their operations and empowering ID users to monitor mishandling, existing security ramifications alone will not be enough to increase trust in the use and operation of these systems.

1.1 Why trustworthy digital ID systems?

Digital identity management is essential to expand digital economic activities and is considered a fundamental requirement for ID verification and proofing, financial transactions and online activities. The success of these systems is linked to the degree of trust that users, organisations and governments place on their services and operations. The end users are willing to take certain risks related to the use of the EIDS systems (trustee). Therefore, the users have a certain level of expectations from these systems that often impacts on the way they perceive the trustworthiness of EIDS.

At the system level the users should also be provided arguments as to why they should trust the service provider and the internal processes and procedures that govern ID management in these systems. In that respect, there should be a process to assess trustworthiness for users and stakeholders to be able to build trust with the system components and reduce the risks associated with EIDS use. Based on that we therefore see the concept of trustworthiness being a core component in the EIDS design and assurance. Figure 1 illustrates how trustworthiness can be decomposed to its different facets (pillars) and the different features and mechanisms related to EIDS systems for each of these facets.

1.2 Trustworthiness principles & trust modelling

Trust forms the basis for decision-making and provides the motivation for maintaining long-term relationships based on cooperation and collaboration. It is also well recognised that trust is important for promoting quick responses to a crisis, reducing transaction costs, avoiding harmful conflict and enabling cooperative behaviour. While the notion of trust has a storied history and is intuitively easy to comprehend, it has not been formally defined. There exists widespread literature on trust-based mechanisms and trust metrics. However, very little is known on how to model trustworthiness and how to quantify this with sufficient detail and context-based adequateness. Modelling and quantifying trust should account for system parameters including communication protocols, information sharing, social interactions and cognitive principles.

Trusted systems' design is typically limited to a particular application domain. For instance, in sociology, trust is considered to be the subjective probability that another party will perform an action that will not hurt one's interests under uncertainty and ignorance. While, in philosophy, trust refers to the risky action deriving from personal, moral relationships between two entities. In the realm of economics, trust is defined as a trustor's motivations for cooperation in risky situations. In psychology, trust is treated as a cognitive construct an individual learns from social experience such as positive or negative consequences of trusting behaviour. In international relations, trust is described as the belief that the other party is trustworthy with the willingness to reciprocate cooperation. In organisational management, trust is defined as the willingness of the trustor to take risk and be vulnerable based on the ability, integrity, and benevolence of the trustee. In automation, trust is defined as the attitude that one agent will achieve another agent's goal in a situation where imperfect knowledge is given with uncertainty and vulnerability. In computing and networking, the concept of trust is a subjective belief about whether another entity will exhibit behaviour reliably in a particular context with potential risks.

Several factors affect an entity's assessment of trust. Commonly investigated factors (See Annex B) include risk, faith, fear, feeling, valence, power, delegation, control, credit,

cooperation, altruism, reciprocation, adoption, social and relational capital, norms, regulations, laws and contracts.

With the proliferation of machine learning technologies, there is greater demand than ever before on building Artificial Intelligence (AI)-based systems which take into account the aforementioned notions of trust and trustworthiness. These technologies promise to revolutionise the identity and access management systems and mitigate against data breaches and additional risks in identity management (such as credential transfer, control access). They can also support the flexible and adaptable implementation of compliance rules for new security laws that can be a burden. However, expectedly, system designers are faced with a staggering challenge to incorporate each one of the aforementioned parameters into next-generation AI-systems. Recent progress in trustworthy AI is centred around designing systems that offer fairness, explainability, auditability and safety. These four dimensions are distilled from policy framework qualities which include privacy, accountability, safety and security, transparency, nondiscrimination, human control of technology, professional ethics and promotion of human values.

Trustworthy AI-systems based on fairness, explainability, auditability and safety are widely analysed from two different perspectives - data-centric and model-centric. The data-centric stage consists of data collection, data preparation and feature engineering. The model-centric stage comprises training, testing and inference (See Annex B).

1.3 Trusted digital identity systems

Once trust has been modelled, the subsequent step is to validate underpinning metrics. A good measurement system provides an objective indicator of the trustworthiness of a system's ability to meet the desired requirements. Early machine learning systems measured trustworthiness in terms of accuracy and precision. However, with the growing need of AI-based systems, measurements and metrics will go beyond the traditional metrics of security and privacy, and will include parameters such as reliability, resilience, robustness and ethics. Furthermore, the relationship and correlation between these six dimensions will aspect the choice of measurement systems and the associated metrics.

Trusted digital Identity systems must always ensure a consistent set of verifiable attributes for each digital identity. They should also provide an immutable and verifiable link between individuals and their digital identity. This often entails access to 3rd party resources and services such as database systems, social records and credit card information. Users should also be able to use their trusted digital identity to access multiple security-sensitive services such as online banking and mobile money. Thus a trustworthy ID must be able to assert its trustworthiness in capturing, verifying and digitising ID-related information. With the automation and intelligence introduced in these processes, identity verification and capturing becomes a seamless process for subscribers. Smart digital identity verification is gradually replacing older procedures due to the added value to be had in identity governance. Examples include, but are not limited to, enhancing security in authentication processes, better user experience, consistent legal and regulatory compliance and reduced costs. The process of combining data security, user experience and insight management for different stakeholders (customers, employees and suppliers) using AI/ML in Identity and access management systems seems to promote efficient management of users' digital identity.

There is widespread work on how to measure attributes of trust from complex, composite networks consisting of communication, information, social, and cognitive domains. (See Annex B)

In defining a Trustworthy Digital Identity Systems Framework (TDISF), we will identify the measurable elements behind the relevant facets that support the establishment of trustworthiness across the system as a whole. These facets are complex to achieve within system design and it is understood that they can have potential to be conflicting in their objectives. We capture the complexities of each facet and consider the tensions to determine the measures and pseudo measures that can be used to facilitate representation of trustworthiness within the context of particular use cases.

Figure 1 : Trustworthiness Facets, Attributes and Features

2. Trustworthiness Facets, Attributes and Features

This section describes the key components of our trustworthiness framework (See Fig. 1). We decompose each of these components to articulate better their relationship to the six (6) facets identified. This conceptual framework analyses the facets of trustworthiness that can influence trust in EIDS. By addressing these facets, we can better measure the impact of perceived trust in these systems and assess their trustworthiness assurance. Trust towards the system is important for all stakeholders to engage with its services; thus, capturing and assessing EIDS trustworthiness can lead to deliberate decisions whether to take risks involved with EIDS use. Each of the pillars identified is broken down to their respective attributes.

These attributes have been selected based on their relevance and significance on EIDS software features and processes during the whole ID management lifecycle. These attributes are intended to directly impact trust and influence the trustworthiness assessment of these systems holistically. We argue that for an EIDS to be trustworthy, the facets of security, ethics, privacy, robustness, reliability and resiliency must be considered in research. Moreover, the attributes for each of these facets (see Sec 3.2) must be defined and described within the context of EIDS operations. For the reason of distinction, we discuss them under the term *trustworthiness attributes.* We treat these attributes (for each of the facets) as important factors that can impact the overall perception of trustworthiness for an EIDS. Also, we focus on these attributes as they may have an impact not only on the performance, usability and accuracy of EIDS but also their ability to certify and accredit against formal specifications (that might have an impact on trustworthiness). We then proceed and decompose each of these attributes to their respective features and mechanisms. These features are key tactical and operational means to address the facets.

2.1 Pillars (Facets)

We identify and define the following trustworthiness pillars (facets) as part of our framework:

Security: The protection of data, information and systems against unauthorised access or modification whether in storage, processing, or transit and against denial of service to authorised entities.

Privacy: Ensure that personal and sensitive information transmitted, processed and shared is treated privately, in adherence to legal and regulatory restrictions governing its use.

Robustness: The ability of the system to continue functioning in the presence of internal and external challenges without fundamental or drastic changes to its original operations or state[1].

Ethics: Ensure transparent, responsible and auditable operations throughout the whole lifecycle of data and information management in systems whilst enabling user empowerment into this process.

[1] There is no globally agreed definition of robustness, and the situation is further blurred by its relationship to resilience and stability (Michael Glodek, 2006)

Reliability: The ability of the system to perform in a consistent and expected way during a period of time in which it adheres to its performance specifications adequately.

Resiliency: The ability of the system to adjust to internal and external conditions by adapting its operations to ensure the continuation of expected service under these new conditions.

2.2 Attributes

For each of the facets (pillars) in our framework, we define their corresponding properties as follows:

Security

- ***Confidentiality:*** The appropriate means and practices to controlling access to data, information and processes to prevent unauthorised access and disclosure.
- ***Integrity:*** The appropriate means and practices to ensure the correctness and accuracy of data and information in use, transit and store.
- ***Availability:*** The appropriate means and practices to ensure that Information systems, services and platforms are always available and provide timely access to resources to authorised users.

Privacy

- ***Communication:*** the expectation that measures are in place to ensure that individuals' information is communicated and or shared in a secure and private manner and only be accessible by authorised and intended recipients.
- ***Processing:*** the expectation that measures are in place to ensure that individuals' information is processed[2] in a secure and private manner and only for the purposes it has been collected.

Robustness

- ***Variability Control:*** the ability of a system to demonstrate appropriate and standardised handling of unexpected changes in the expected outcomes of a system's operation.
- ***Maintainability:*** The degree at which the system and its sub-systems and components are capable of being retained in or restored to serviceable operation (Anil Mital, 2014).

Ethics

- ***Transparency[3]:*** the ability of the system to enable visibility to users, and auditability against its internal and external operations to ensure predictability of its outputs.

[2] By "processing" we define the process of information collected, recorded, altered, stored, retrieved and deleted by manual or automated means.

[3] One may argue that this is not an ethical principal per se but it could be an ethically enabling or impairing factor (Turilli, 2009)

- ***Fairness:*** the ability of the system to demonstrate that it performs actions and is governed by processes that are morally right and equitable for all its users.
- ***Explainability:*** The ability of a system to demonstrate comprehensibility to assist users with understanding why particular outcomes have been delivered *(Schneider, 2020)*.

Reliability

- ***Repeatable:*** the ability of the system, and users, to execute repetitive tasks and actions to govern the identity life cycle.
- ***Consistent:*** the ability of the system to handle and reconcile errors which can led to inconsistencies in the identity life cycle management.

Resiliency[4]

- ***Recover:*** the ability of the system to implement and deploy appropriate mechanisms and functions to maintain plans for resilience and to restore any capabilities or services that were impaired due to an incident.
- ***Detect:*** the ability of the system to implement and deploy appropriate mechanisms and functions to discover and identify the occurrence of an incident
- ***Protect:*** Develop and implement appropriate safeguards to ensure the delivery of critical services.

2.3 Features and Mechanisms

In this section, we identify the key features for each of the facets' attributes defined in section. 3.2. We tabulate these features for better articulation. These features will be used in our future research activities to derive appropriate TAL requirements for each of the facets that will enable a systematic trustworthiness assessment of existing and emerging EIDS systems.

For each of the facets' attributes we identify the following features and mechanisms (Abbreviations identified in Annex C):

[4] Following from the NIST CSF (Technology, 2018)

Table 1 Trustworthiness features and mechanisms - Security

SECURITY	
Confidentiality, Integrity, Availability	
Features and Mechanisms	**Description**
Access Control (AC) Methods	• Evidence on deployment and use of relevant and suitable AC mechanisms; • Segmented access to resources; • Grant access rights management processes; • deployment of DAC, RBAC, ABAC, MAC; • Physical and logical segmentation;
Security Assessment and Auditing	• Evidence on systematic security assessment (for example, penetration testing, breach assessment); • Use of standardised methodologies and tool (such as OSSTMM, OWASP, PTES and ISSAF); • Identification and prioritisation of threats; • Assess current security performance; • External and internal security audits; • Definition of periodic and regular assessments; vulnerability assessment; • Design control policies and procedures. • Review standard operating procedures and policies; configure management; • AC assessment; • Security policy sanity checks;
Authentication-Authorisation-Accounting (AAA)	• Identify all applications and entities in the environment; Principle of Least Privilege; • Principle of Separation of Duties; • AC capability assessment; • User authentication and authorisation processes and procedures; • Assigned authorisation levels that define access; • Network or applications way of identifying a user; • Maturity level of identification processes; • Level of enforcing policies; • Attributes to describe authorised actions; • Clear set of restrictions on access (including geographical, network-based, route-based and the like); • Accounting measures in place; • Managing of accounting records;

	• Access to standardised authentication methods (including RADIUS, TACACS+ and Kerberos among others)
SSO, MFA	• Restrictions on multiple authentications in single sessions; • Evidence on several different factors to verify a person's Identity; (including smartcard, FIDO token, OTP, mobile devices, biometrics); • Location-based authentication using GPS, IP address, or Integrated Windows Authentication (IWA); • Evidence on secure storage and processing of various SSO credentials; • Criteria for MFA verification;
Identity Governance and Intelligence	• Ensure connectivity to all sources of Identity; • Purpose-built repository to share Identity; • Evidence industry standards for protocols, • Data sources to exchange identity and account information; • Evidence and audit of all connectors used to manage identity and account information; • Audit and administer access to different classes of applications (group membership); • Using examples of connectors such as: LDAP, JDBC, CSV, REST, and SCIM; clear evidence of connecting people to accounts (correlation); • Evidence of compliance with standards when processing data in unstructured data repositories (such as Onedrive and Sharepoint); • Evidence of technical access control facility (entitlement catalog); • A defined set of LCM Human Resources (HR) states;
Evidence of Layered-Security	• Evidence of protecting Digital ID components and digital assets with several layers of security; • Evidence of Layered security controls (administrative, technical, physical);
Defence-in-depth	• Evidence of a multifaceted strategic plan for monitoring, alerting and emergency response for the EIDS; • Rapid notification and response procedures in place;

Policy enforcement and protection of PII/SPII	• Data Loss Prevention (DLP) technologies in place (including cloud access security brokers (CASBs) technical security controls such as NGA firewalls and gateways); • Use of Privacy Enhancing Technologies;
Regulatory Compliance (DID assurance)	• Evidence of Internal and external audits; • Proactively address compliance requirements; compliance with regulatory standards (such as ISO27001, ISO/IEC 24760, ISO/IEC 29115, ISO/IEC 29146, GDPR, DPA, PCI-DSS and HIPAA); • Assess the ability to aggregate and correlate identity data; • Conduct baseline access certification; • Provide proof of compliance; • Perform closed-loop audit on all changes;
Cryptographic protection	• Evidence of appropriate cryptography policies; • Use of strong cryptographic algorithms for symmetric and asymmetric cryptographic operations; • Use of approved algorithms with appropriate key lengths; evidence of practical recommendations for deployment in EIDS; • Ensure non-repudiation (such as the use of digitally signed transactions);
Maturity level of security policies in place	• Attributes Requirement Analysis conducted; • Clear set of requirements for the information security management system governing identity; • Evidence of appropriate security policy management capability models used; • Monitoring and alerting on policy changes; • Business impact from security changes; • Evidence on security policy automation/optimisation based on emerging technologies and developments in identity requirements; • Evidence of assessing gaps between logical and physical infrastructure;

SSO or federated access Mngt support	• Separation of Duty (SoD) policies; • Evidence of business rules captured in a governance platform; • Account policies management; understanding and cataloguing entitlements;
Vulnerability Management	• Evidence of classifying, prioritising, remediating, and mitigating vulnerabilities; use of vulnerability Management tools; • Formatting checklists and test procedures; • Measuring vulnerability impact; • Share information obtained;
Risk Response	• Evidence of security risk management; • Plans for risk responses; • Evidence of risk management planning; • Compliance with relevant standards for cybersecurity risk management (including ISO 27001L2013, NIST CSF, ITU-T X1208 (01/2014), ISO/IEC 15408:2009, ISO/IEC 17030:2003 and NIST SP 800-122 among others)
Systems & Communications Protection	• System and communications protection policy and procedures; • Enforcing requirements; • Implementation and testing of communication protection policy; • Evidence of good practice and adherence to directives and standards (from bodies such as NIST and ISO)

Table 2 Trustworthiness features and mechanisms - Privacy

PRIVACY	
Communication, Processing	
Features and Mechanisms	**Description**
Collection and Data minimisation	• Evidence of the amount of personal data that needs to be collected; • Specify reasons for the collection and use of data;
User Consent for attributes' collection, processing, re-use and release	• Evidence that user consent has been obtained; automated processing of personal data should be documented; • Provide information to users about the data processing; • Facilitate human intervention or challenge a decision with regards to data processing; • Regular checks that data is treated as intended;
Limited attribute retention	• Appropriate retention policy in place; • Demonstrate controlled access to personal attributes; • Restrict the information that can be retrieved by 3rd party service providers; • Support conditional anonymity;
Remote Identity Proofing and Non-Face-to-face Onboarding	• Ensure controls are in place to facilitate remote identity proofing and enrolment in a way that reduces privacy and fraud risks;
Use Limitation	• Clear and standard data use limitation (DUL) statements;

ID attributes collected fit for scope and purpose	• Ensure all ID attributes collected are fit for use; • Proof that ID attributes have been collected for a specific purpose; • Proof that ID attributed have been used for the purpose they have been collected; evidence of measures and safeguards against ID attributes' abuse'; • Audit-trail record of ID attributes' management lifecycle;
Troubleshooting identity proofing	• Measures in place to validate sufficient information about users; • Demonstrate steps in identity verification troubleshooting; verify information in a privacy-preserved manner;
Privacy Impact Assessment (PIA)	• Conduct periodical PIA; • Maintain an up-to-date privacy risk register; • Clear sets of criteria for each of the processing operation that requires PIA; • Evidence that DPIA is continuously reviewed and regularly re-assessed;
Privacy Risk Mitigation Plans	• Appropriate and documented privacy risk mitigation steps;
Well defined privacy models and policies	• Documented privacy policies and requirements; • Privacy goals analysis;
Secondary use	• Clear evidence of information flows; • Plans to achieve the safety for information shared; • Evidence of multilateral information flow control technically enforced;

Privacy Standard(s)	• Compliance with relevant privacy standards (including, but not limited to, GDPR, ISO/IEC 27701, Regulation on Privacy and Electronic Communications; Cybersecurity Act; ISO/IEC 29101:2013, ISO/IEC 27550 and ISO/IEC 27550)

Table 3 Trustworthiness features and mechanisms - Robustness

ROBUSTNESS	
Variability Control, Maintainability	
Features and Mechanisms	**Description**
Expected outcomes from unexpected inputs	• Ensure the system can predict affective responses to unexpected outcomes;
In time provisioning and de-provisioning processes	• Presence and management of an enterprise-class provisioning engine; • Documented processes for delivering access to applications and data; evidence on how the system interacts with legacy provisioning systems (such as identity provisioning integration patterns); • Evidence of ongoing assignment and de-assignment lifecycle(s); • Provisioning engine fail-over, retry and recovery scenarios present; • Metrics available on execution and process flows; built-in tracking; • Monitoring, and root-cause analysis capabilities;
Tolerate process variability in operating and environmental conditions	• Monitor environment instability; manage equipment malfunctions and / or errors; • Evidence of standardisation on the way that ID attributes are collected and processed; • Monitor changes associated with data storage and processing;
Evidence of maintainability Requirements	• Ongoing monitoring and maintenance of ID governance policies;

Table 4 Trustworthiness features and mechanisms - Ethics

ETHICS	
Transparency, Fairness, Explainability	
Features and Mechanisms	**Description**
Data and process provenance	• Extensive delegation and scoping capabilities defined; • Approval workflows clearly defined and communicated; • Evidence of a flexible request catalog model; • Documentation on testing; • Demonstrate contextual allocation of data; • Applicable data handling rules; • Deletion routines implemented; • Detailed register of data processing tools;
Evidence of user empowerment to monitor use and potential misuse	• Full visibility into what is being requested by whom; • Tracking and management data for reporting; • Evidence of communication of improper reporting or access awarded; • Easy access to own personal data;
"Openess" of systems and algorithms managing the EIDS	• Ensure recognisable algorithms used in EIDS; • Evidence that users have been given knowledge and control over the processing of their data; • Analytical parameters to eliminate “black-box” situations; • Human-readable policies;

Audit-trail on how the Identity was used	• Clear identification of all log sources in the system; appropriate log storage and disposal policies; • Evidence of enterprise-level log security mechanisms; • Transparent Log Management infrastructure components; • Clear audit trail on how ID is managed in the system; • Monitor log rotation;
Decisions are made in an appropriate manner based on users' consent	• Evidence of legal and regulatory compliance for all automated individual decision-making and profiling;
Evidence of measures taken to assure inclusivity and accessibility for all who have a right of access	• effort to assess and reduce barriers to access • considerations and provision of alternative access where needed • governance of its application • clear assessment of the impact of non-engagement on basic user rights • documented processes to facilitate inclusive enrolment • Clear assessment of the ease of access for different user groups • Clear criteria for exclusivity and reasoned argument (such as societal or health protections) as to why this is appropriate

Table 5 Trustworthiness features and mechanisms - Resiliency

RESILIENCY	
Recover, Detect, Prevent	
Features and Mechanisms	**Description**
EIDS Internet face protection	• Update programs and systems regularly; • Patch management; • Network security controls catalogue;
Internal EIDS security processes enforced	• Measure the health of controls in place (evidence on cyber due diligence); • Monitor 3rd party vendors; • Identify additional resilience measures; • Benchmark performance against different systems;
Back up and Disaster Recovery Plans	• A comprehensive BCP and DR strategy in place; outputs of BIA; • Determine acceptable downtime for each critical function; Clear scope of recovery plans; • Evidence of testing the continuity plans; • Review and improvements;
Cyber resilience strategy in place	• Evidence of plans to ensure business and system delivery; document safe-to-fail capabilities of systems and processes; • Demonstrate multi-layer protections of systems and sub-systems; evidence of a resilience architecture in place;
Demonstrate balance between preventive and detective controls	• Separation of duty rules implemented; • Monitor new access provisioning or self-service access request; • Periodic policy evaluation for detective controls; access reviews; • Reporting and analysis, and inventory variance assessment;

Reaction to Security incidents against the EIDS	• Mitigation plans to re-establish operational efficiency; evidence of well-defined incident response plans; • develop and Document IR Policies; • Define Communication Guidelines; • Use of threat intelligence feeds; • Evidence of post incident handling processes;
ID recovery	• Fully automated password recovery and reset capabilities;

Table 6 Trustworthiness features and mechanisms - Reliability

RELIABILITY	
Repeatable, Consistent	
Features and Mechanisms	**Description**
Streamlined user contact points and processes	• Usability considerations specific to the pre-enrolment/enrolment sessions; • Clear enrolment session procedures and guidelines; • Required identity evidence for registration; • Notification and confirmation procedures;
	• Checklist of actions and requirements to ensure successful enrolment; • Appropriate safeguards to detect discrepancies in the identity evidence; • Set user expectations regarding the outcomes;
EIDS sufficient Assurance Levels against Fraud risks	• Ensure the independence and reliability of the EIDS against fraud by the accuracy of the results it produces; • Benchmark against FATF Guidance on Digital Identity; • Check compliance with existing national and international ID assurance frameworks and standards (see UK Identity Assurance Programme)
Government approved audits	• Provide evidence of Government approved audits as part of the system's ID assurance;
Evidence of assurance assessment	• Provide evidence of internal/external assurance assessments; • Define the IAF against which the EIDS has been assessed;

Handling unexpected termination and unexpected actions	• Protocols and procedures for error handling in systems and processes; • Configure error handling methods;

3. CONCLUSION

Understanding of the assurance levels of EIDS necessitates a decomposition of the technology, architecture and governance of these systems. However, existing applications of risk-based approaches to understand assurance levels and assess whether the EIDS is reliable, should also consider additional facets when it comes to trustworthiness. This report decomposes the different facets that constitute trustworthiness with emphasis placed on EIDS operations. We provide a set of attributes for each of the facets and a comprehensive list of associated features. The latter will be used as the basis in our forthcoming research activities to derive the different trustworthiness assurance levels (TAL) for each of the facets. We anticipate this framework to act as a tool for the perceived trustworthiness estimation and assessment in EIDS. We currently investigate steps for the verification of the relationship between the trustworthiness features and its addressed trustworthiness facets for EIDS. This will enable us to measure a feature's relative impact on trustworthiness and systematically derive the EIDS trustworthiness goals and requirements to measure their assurance for each of the facets.

Bibliography

Anil Mital, A. D. (2014). 8 - Designing for Maintenance. In A. D. Anil Mital, Product Development (Second Edition) (pp. 203-268). Elsevier.

Kubach, M. S. (2020). Self-sovereign and Decentralised Identity as the future of identity management? Open Identity Summit, (pp. 35-47). Bonn.

Michael Glodek, S. L. (2006). Process Robustness - A PQRI White Paper. ISPE.org.

NIST. (2020). NIST Special Publication 800-53: Security and Privacy Controls for Information Systems and Organisations. NIST.

P. C. Bartolomeu, E. V. (2019). Self-Sovereign Identity: Use-cases, Technologies, and Challenges for Industrial IoT. EEE International Conference on Emerging Technologies and Factory Automation (ETFA) (p. EEE International Conference on Emerging Technologies and Factory Automation (ETFA)). Zaragosa: IEEE.

Schneider, L. C. (2020). Explainability as a non-functional requirement: challenges and recommendations. Requirements Engineering, pages493–514.

Technology, N. I. (2018). Framework for Improving. NIST.

Turilli, M. F. (2009). The ethics of information transparency. Ethics in Information Technology, 105–112.

Y. Yamamoto. (1990). A morality based on trust: Some reflections on japanese morality. Philosophy East and West 40, 4 (October 1990), 451–469. Understanding Japanese Values.

D. Gambetta. (1988). Can we trust trust? In Trust: Making and Breaking Cooperative Relations, D. Gambetta (Ed.). *Basil Blackwell*, New York, USA, 213–237.

B. Lahno. (1999). Olli lagerspetz: Trust. The tacit demand. *Ethical Theory and Moral Practice* 2, 4 (1999), 433–435.

H. S. James. (2002). The trust paradox: A survey of economic inquiries into the nature of trust and trustworthiness. *Journal of Economic Behavior and Organization* 47, 3 (March 2002), 291–307.

J. B. Rotter. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologis*t 35, 1 (Jan. 1980), 1–7.

A. H. Kydd. (2005). A blind spot of philosophy. *Trust and Mistrust in International Relations*. Princeton University Press.

R. C. Mayer, J. H. Davis, and F. D. Schoorman. (1995). An integrative model of organizational trust. *Academy of Management Review* 20, 3 (1995), 709–734.

J. H. Cho, A. Swami, and I. R. Chen. (2011). A survey of trust management in mobile ad hoc networks. *IEEE Communications Surveys and Tutorials* 13, 4 (2011), 562–583.

S. Adah. (2013). Modeling Trust Context in Networks. *Springer Briefs in Computer Science*.

J. Li, R. Li, and J. Kato. (2008). Future trust management framework for mobile Ad Hoc networks. *IEEE CommunicationsMagazine* 46, 4 (April 2008), 108–114. Security in Mobile Ad Hoc and Sensor Networks.

N. Luhmann. (1979). Trust and Power. *John Wiley & Sons Inc*.

C. Castelfranchi and R. Falcone. (2010). Trust Theory: A Socio-Cognitive and Computational Model, Michael Wooldridge (Ed.). *Series in Agent Technology*. Wiley.

R. S. Lazarus, J. R. Averill, and E. M. Opton. (1970). Towards a cognitive theory of emotion. In Feelings and Emotions, M. B. Arnold (Ed.). *Academic Press*, New York, 207–232. The Loyola Symposium.

C. Castelfranchi. (2009). A non-reductionist approach to trust. *In Computing with Social Trust*, J. Golbeck (Ed.). Springer, London Limited, Human-Computer Interaction Series.

J. R. Dunn and M. E. Schweitzer. (2005). Feeling and believing: The influence of emotion on trust. *Journal of Personality and Social Psychology* 88, 5 (2005), 736–748.

H. Farrell. (2009). Distrust. In Trust, Distrust, and Power, R. Hardin (Ed.). *Russell Sage Foundation*, New York, 84–105.

T. J. Norman and C. Reed. (2010). A logic of delegation. *Artificial Intelligence* 174, 1 (Jan. 2010), 51–71.

C. Castelfranchi and R. Falcone. (2000). Trust and control: A dialectic link. *Applied Artificial Intelligence Journal:Special Issue on Trust in Agent*, Part I 14, 8 (2000), 799–823

A. Jøsang, R. Hayward, and S. Pope. (2006). Trust network analysis with subjective logic. *In Proceedings of the Australasian Computer Science Conference (ACSC'06)*. Vol. 48, 85–94.

R. Axelrod. (1981). The evolution of cooperation. *Science* 211 (1981), 1390–1396.

R. Trivers. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology* 46, 1 (March 1971), 35–57.

A. H. Harcourt. (1991). Help, cooperation and trust in animals. *In Cooperation and Prosocial Behaviour*, R. Hinde and J. Groebel (Eds.). Cambridge University Press, 15–26.

R. Hardin. (2002). Trustworthiness. Trust and Trustworthiness. Russell Sage Foundation, New York, 28–53.

Roy, M.C., Dewit, O. and Aubert, B.A. (2001). The impact of interface usability on trust in web retailers. Internet Research.

R. D. Putnam. (2000). Bowling Alone. Simon and Schuster, New York.

C. Castelfranchi. (1995). Social commitment: From individual intentions to groups and organizations. *In Proceedings of the 1st International Conference on Multi-Agent Systems (ICMAS'95).* AAAI-MIT Press, San Francisco, California, USA, 41–49.

Avrim L Blum and Pat Langley. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97, 1-2 (1997), 245–271.

Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* 5, 2 (2017), 120–134.

Thomas G Dietterich and Eun Bae Kong. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical Report. Technical report, Department of Computer Science, Oregon State University.

Rehab Duwairi and Mahmoud El-Orfali. (2014). A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. *Journal of Information Science* 40, 4 (2014), 501–513.

Bradley Efron. (1994). Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* 89, 426 (1994), 463–475.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. (2019). A comparative study of fairness-enhancing interventions in machine learning. *In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM*, 329–338.

Carlos Vladimiro González Zelaya. (2019). Towards Explaining the Effects of Data Preprocessing on Machine Learning. 2019 *IEEE 35th International Conference on Data Engineering (ICDE)* (2019).

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 93.

Himanshu Gupta, Sameep Mehta, Sandeep Hans, Bapi Chatterjee, Pranay Lohia, and C Rajmohan. (2017). Provenance in context of Hadoop as a Service (HaaS)-State of the Art and Research Directions. *In 2017 IEEE 10th International Conference on Cloud Computing (CLOUD)*, 680–683.

Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (2011), 141–154.

Soumendra Nath Lahiri. (2013). Resampling methods for dependent data. Springer Science & Business Media.

Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research* 18, 1 (2017), 559–563.

Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM*, 502–510.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. (2008). Discrimination-aware data mining. *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 560–568.

Burr Settles. (2009). Active learning literature survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences

Alper Kursat Uysal and Serkan Gunal. (2014). The impact of preprocessing on text classification. *Information Processing and Management* 50, 1 (2014), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. (2017). Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410 (2017).

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. (2017). On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267 (2017).

Weilin Xu, David Evans, and Yanjun Qi. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017).

Glenn Fung, Sathyakama Sandilya, and R Bharat Rao. (2005). Rule extraction from linear support vector machines. *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 32–40.

Sara Hajian, Francesco Bonchi, and Carlos Castillo. (2016). Algorithmic bias: From discrimination discovery to fairnessaware data mining. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2125–2126.

Sara Hajian, Josep Domingo-Ferrer, and Oriol Farràs. (2014). Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery* 28, 5-6 (2014), 1158–1188.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. (2012). Fairness-aware classifier with prejudice remover regularizer. *In Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. *In Proceedings of the Conference on Fairness, Accountability, and Transparency*. 349–358.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, 1721–1730.

Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. (2014). A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery* 28, 5-6 (2014), 1503–1529.

Sanjay Krishnan and Eugene Wu. (2017). Palm: Machine learning explanations for iterative debugging. *In Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. 1–6.

Yin Lou, Rich Caruana, and Johannes Gehrke. (2012). Intelligible models for classification and regression. *In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 150–158.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. (2013). Accurate intelligible models with pairwise interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD.* 623.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. *In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM*, 1322–1333.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. (2014). Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. *In 23rd USENIX Security Symposium*. 17–32.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. (2016). Stealing machine learning models via prediction apis. *In 25th USENIX Security Symposium*. pp. 601–618.

Selvaraj, A., and Sundararajan, S. (2017). Evidence-based trust evaluation system for cloud services using fuzzy logic. International Journal of Fuzzy Systems, 19(2), 329–337.

Rafique, N., Khan, M. A., Saqib, N. A., Bashir, F., Beard, C., and Li, Z. (2016). Black hole prevention in vanets using trust management and fuzzy logic analyzer. International Journal of Computer Science and Information Security, 14(9), 1226.

Nagy, M., Vargas-Vera, M., and Motta, E. (2008). Multi agent trust for belief combination on the semantic web. In The 4th international conference on intelligent computer communication and processing (ICCP) , pp. 261–264 .

Lesani, M., and Bagheri, S. (2006). Fuzzy trust inference in trust graphs and its application in semantic web social networks. In World automation congress (WAC) , pp. 1–6.

Chen, H., Yu, S., Shang, J., Wang, C., and Ye, Z. (2009). Comparison with several fuzzy trust methods for p2p-based system. In International conference on information technology and computer science (ITCS) , Vol. 2, pp. 188–191 .

Liao, H., Wang, Q., and & Li, G. (2009). A fuzzy logic-based trust model in grid. IEEE International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC), 1, 608–614.

Luo, J., Liu, X., Zhang, Y., Ye, D., and Xu, Z. (2008). Fuzzy trust recommendation based on collaborative filtering for mobile ad-hoc networks. In The 33rd IEEE conference on local computer networks (LCN), pp. 305–311.

Manchala, D. W. (1998). Trust metrics, models and protocols for electronic commerce transactions. In Proceedings of the 18th international conference on distributed computing systems , pp. 312–321 .

Nefti, S., Meziane, F., and Kasiran, K. (2005) A fuzzy trust model for e-commerce. In Proceedings of the 7th IEEE international conference on E-commerce technology (CEC) , pp. 401–404 .

Jøsang, A. (2016). Bayesian reputation systems. In Subjective logic (pp. 289–302). New York: Springer

Jiang, J., Han, G., Wang, F., Shu, L., & Guizani, M. (2015). An efficient distributed trust model for wireless sensor networks. IEEE Transactions on Parallel and Distributed Systems, 26(5), 1228–1237.

Filali, F. Z., & Yagoubi, B. (2015). Global trust: A trust model for cloud service selection. International Journal of Computer Network and Information Security, 7(5), 41.

Alhadad, N., Busnel, Y., Serrano-Alvarado, P., & Lamarre, P. (2014). Trust evaluation of a system for an activity with subjective logic. In International conference on trust, privacy and security in digital business, pp. 48–59. New York: Springer

Ahmadi, M., Gharib, M., Ghassemi, F., & Movaghar, A. (2015). Probabilistic key pre-distribution for heterogeneous mobile ad hoc networks using subjective logic. In The 29th IEEE international conference on advanced information networking and applications (AINA), pp. 185–192.

Cerutti, F., Kaplan, L. M., Norman, T. J., Oren, N., & Toniolo, A. (2015). Subjective logic operators in trust assessment: An empirical study. Information Systems Frontiers, 17(4), 743–762.

Liu, G., Yang, Q., Wang, H., Lin, X., & Wittie, M. P. (2014). Assessment of multi-hop interpersonal trust in social networks by three-valued subjective logic. In Proceedings of IEEE international conference on computer communications (INFOCOM), pp. 1698–1706.

Jøsang, A. (1999). An algebra for assessing trust in certification chains. The Network and Distributed System Security Symposium (NDSS), 99, 80–89.

Jøsang, A. (2001). A logic for uncertain probabilities. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 9(03), 279–311.

Lioma, C., Larsen, B., Schütze, H., & Ingwersen, P. (2010). A subjective logic formalisation of the principle of polyrepresentation for information needs. In Proceedings of the 3rd symposium on information interaction in context (pp. 125–134).

Oren, N., Norman, T. J., & Preece, A. (2007). Subjective logic and arguing with evidence. Artificial Intelligence, 171(10–15), 838–854.

Deepa, R., and Swamynathan, S. (2014). A trust model for directory-based service discovery in mobile ad hoc networks. In International conference on security in computer networks and distributed systems (pp. 115–126). New York: Springer

Wang, J., and Sun, H. (2007). Inverse problem in DSMT and its applications in trust management. In The 1st international symposium on data, privacy, and E-commerce (ISDPE) pp . 424–428.

Wang, K., and Wu, M. (2007) A trust approach for node cooperation in manet. In International conference on mobile Ad-Hoc and sensor networks (pp. 481–491). New York: Springer.

Zhang, W., Zhu, S., Tang, J., and Xiong, N. (2017). A novel trust management scheme based on Dempster–Shafer evidence theory for malicious nodes detection in wireless sensor networks. The Journal of Supercomputing, 74, 1–23.

Nguyen, V., and Huynh, V. (2016). Integrating with social network to enhance recommender system based-on Dempster–Shafer theory. In International conference on computational social networks (pp. 170–181). New York: Springer.

Esposito, C., Castiglione, A., and Palmieri, F. (2018). Information theoretic-based detection and removal of slander and/or false-praise attacks for robust trust management with Dempster–Shafer combination of linguistic fuzzy terms. Concurrency and Computation: Practice and Experience

Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. (2000). Reputation systems. Communications of the ACM, 43(12), 45–48.

Abdul-Rahman, A., and Hailes, S. (2000). Supporting trust in virtual communities. In Proceedings of the 33rd Hawaii international conference on system sciences Vol. 6, pp. 6007–6016.

Jonker, C. M., and Treur, J. (1999). Formal analysis of models for the dynamics of trust based on experiences.In European workshop on modelling autonomous agents in a multi-agent world, pp. 221–231. New York: Springer.

Jonker, C. M., Schalken, J., Theeuwes, J., and Treur, J. (2004). Human experiments in trust dynamics. In International conference on trust management (pp. 206–220). New York: Springer.

Buchegger, S., and Le Boudec, J.-Y. (2004). A robust reputation system for peer-to-peer and mobile adhoc networks. In The 2nd workshop on economics of peer-to-peer systems (P2PEcon), pp. 1–6.

Pirzada, A. A. and McDonald, C. (2004). Establishing trust in pure Ad-hoc networks. In Proceedings of the 27th Australasian conference on computer science (Vol. 26, pp. 47–54). Australian Computer Society, Inc.

Sabater, J., and Sierra, C. (2001). REGRET: reputation in gregarious societies. In Proceedings of the 5th international conference on autonomous agents pp. 194–195.

Wang, Y., and Varadharajan, V. (2005) Two-phase peer evaluation in P2P E-commerce environments. In Proceedings of IEEE international conference on e-technology, e-Commerce and e-Service, pp. 654–657.

Azzedin F., and Maheswaran, M. (2002). Evolving and managing trust in grid computing systems. In IEEE Canadian conference on electrical and computer engineering, Vol. 3, pp. 1424–1429.

Hung, K., Lui, K., and Kwok, Y. (2007). A trust-based geographical routing scheme in sensor networks. In IEEE wireless communications and networking conference, pp. 3123–3127.

Song, W., and Phoha, V. (2004). Neural network-based reputation model in a distributed system. In Proceedings of IEEE international conference on e-commerce technology (CEC) pp. 321–324.

Baohua, H., Heping, H., and Zhengding, L. (2005). Identifying local trust value with neural network in P2P environment. In The first IEEE and IFIP international conference in central Asia on internet, pp.1–5.

Songsiri, S. (2006). MTrust: a reputation-based trust model for a mobile agent system. In International conference on autonomic and trusted computing, pp. 374–385. New York: Springer.

Wang, Y., Cahill, V., Gray, E., Harris, C., and Liao, L. (2006). Bayesian network based trust management. In International conference on autonomic and trusted computing, pp. 246–257. New York: Springer.

Momani, M., Challa, S., and Alhmouz, R. (2008). BNWSN: Bayesian network trust model for wireless sensor networks. In Mosharaka international conference on communications, computers and applications, pp. 110–115.

Nguyen, C. T., Camp, O., and Loiseau, S. (2007). A bayesian network based trust model for improving collaboration in mobile ad hoc networks. In IEEE international conference on research, innovation and vision for the future, pp. 144–151.

Michiardi, P., and Molva, R. (2002). Core: a collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks. In Advanced communications and multimedia security pp. 107–121. New York: Springer.

Xiong, L., and Liu, L. (2003). A reputation-based trust model for peer-to-peer e-commerce communities. In IEEE international conference on e-commerce, pp. 275–284.

Jiang, T., and Baras, J. S. (2004) Ant-based adaptive trust evidence distribution in MANET. In Proceedings of the 24th international conference on distributed computing systems workshop, p. 588–593.

Wang, W., Zeng, G., and Yuan, L. (2006). Ant-based reputation evidence distribution in P2P networks. In The 5th international conference grid and cooperative computing, pp. 129–132.

Mármol, F. G., and Pérez, G. M. (2011). Providing trust in wireless sensor networks using a bioinspired technique. Telecommunication Systems, 46(2), pp. 163–180.

Marmol, F. G., Perez, G. M., and Skarmeta, A. (2009). TACS, a trust model for P2P networks. Wireless Personal Communications, 51(1), pp. 153–164.

Santos, N., Rodrigues, R., Gummadi, K. P., and Saroiu, S. (2012) Policy-sealed data: A new abstraction for building trusted cloud services. In USENIX security symposium, pp. 175–188.

Neuman, B. C., and Ts'o, T. (1994). Kerberos: An authentication service for computer networks. IEEE Communications Magazine, 32(9), 33–38.

Winslett, M., Yu, T., Seamons, K. E., Hess, A., Jacobson, J., Jarvis, R., et al. (2002). Negotiating trust in the web. IEEE Internet Computing, 6(6), 30–37.

Li, N., Winsborough, W. H., and Mitchell, J. C. (2003). Distributed credential chain discovery in trust management. Journal of Computer Security, 11(1), 35–86.

Nejdl, W., Olmedilla, D., and Winslett, M. (2004). Peertrust: Automated trust negotiation for peers on the semantic web. In Workshop on secure data management, pp. 118–132. New York: Springer.

Bonatti, P., and Olmedilla, D. (2005). Driving and monitoring provisional trust negotiation with metapolicies. In The 6th IEEE international workshop on policies for distributed systems and networks, pp. 14–23.

Winsborough, W. H., Seamons, K. E., and Jones, V. E. (2000). Automated trust negotiation. In Proceedings of DARPA information survivability conference and exposition, Vol. 1, pp. 88–102.

Becker, M. Y., and Sewell, P. (2004). Cassandra: Distributed access control policies with tunable expressiveness. In Proceedings of the 5th IEEE international workshop on policies for distributed systems and networks, pp. 159–168.

Olmedilla, D. (2007). Security and privacy on the semantic web. In Security, privacy, and trust in modern data management, pp. 399–415. New York: Springer.

Lee, M. K., and Turban, E. (2001). A trust model for consumer internet shopping. International Journal of Electronic Commerce, 6(1), pp. 75–91.

Theodorakopoulos, G., and Baras, J. S. (2006). On trust models and trust evaluation metrics for ad hoc networks. IEEE Journal on Selected Areas in Communications, 24(2), pp. 318–328.

Bao, F., Chen, R., Chang, M., and Cho, J.-H. (2012). Hierarchical trust management for wireless sensor networks and its applications to trust-based routing and intrusion detection. IEEE Transactions on Network and Service Management, 9(2), pp.169–183.

Boukerche, A., and Ren, Y. (2008). A trust-based security system for ubiquitous and pervasive computing environments. Computer Communications, 31(18), pp. 4343–4351.

Kamvar, S. D., Schlosser, M. T., and Garcia-Molina, H. (2003) The eigentrust algorithm for reputation management in P2P networks. In Proceedings of the 12th international conference on World Wide Web, pp. 640–651.

Xiong, L., and Liu, L. (2004). Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. IEEE Transactions on Knowledge and Data Engineering, 16(7), pp. 843–857.

Regan, K., and Cohen, R. (2005). A model of indirect reputation assessment for adaptive buying agents in electronic markets. Proceedings of the business agents and semantic web, pp. 41–51.

Zacharia, G., and Maes, P. (2000). Trust management through reputation mechanisms. Applied Artificial Intelligence, 14(9), pp. 881–907.

Su, X., Zhang, M., Mu, Y., and Sim, K. M. (2010). PBTrust: A priority-based trust model for service selection in general service-oriented environments. In IEEE/IFIP 8th international conference on embedded and ubiquitous computing, pp. 841–848.

Annex A: Further reading

Assessment of trust

Several factors affect an entity's assessment of trust. Commonly investigated factors include risk (Luhmann, 1979), faith (Castelfranchi and Falcone, 2010), fear (Lazarus et al, 1970), feeling (Castelfranchi, 2009), valence (Dunn and Schweitzer, 2005), power (Farrell, 2009), delegation (Norman and Reed, 2010), control (Castelfranchi and Falcone, 2000), credit (Jøsang et al, 2006), cooperation (Axelrod, 1981), altruism (Trivers, 1971), reciprocation (Harcourt, 1991), adoption (Hardin, 2002), usability (Roy et al, 2001), social

and relational capital (Putnam, 2000), norms-regulations-laws-contracts (Castelfranchi, 1995).

Trustworthiness of AI

Issues related to trustworthiness of AI in data-centric stage have been investigated.in (Blum and Langley, 1997), (d'Alessandro et al, 2017), (Dietterich and Kong, 1995), (Duwairi and El-Orfali, 2014), (Efron, 1994), (Friedler et al, 2019), (Zelaya, 2019), (Guidotti et al, 2018), (Gupta et al, 2017), (Huysmans et al, 2017), (Lahiri, 2013), (Lemaître et al, 2017), (Luong et all, 2011), (Pedreshi et al, 2008), (Settles, 2009), (Uysal and Gunal, 2014), (Feinman et al, 2017), (Metzen et al, 2017), (Xu et al, 2017). Trustworthiness in model-centric stage of AI have been reported, for example, in (Fung et al, 2005), (Hajian et al, 2016), (Hajian et al, 2014), (Kamishima et al, 2012), (Madras et al, 2019), (Caruana et al, 2015), (Henelius et al, 2014), (Krishnan and Wu, 2017), (Lou et al, 2012), (Lou et al, 2013), (Fredrikson et al, 2015), (Fredrikson et al, 2014), (Tramèr et al, 2016).

Measuring trust and trustworthiness in machine learning

Examples of the main formal techniques for measuring trust and trustworthiness in machine learning-based systems include fuzzy logic (Selvaraj and Sundarajan, 2017)- (Nefti et al, 2005), subjective logic (Josang, 2016) - (Oren et al, 2007), Dempster–Shafer theory (Deepa and Swamynathan, 2014) - (Esposito et al, 2018), ratings (Resnick et al, 2000) - (Buchegger and Le Boudec, 2004), weighting (Pirzada and McDonald, 2004) - (Hung et al, 2007), neural network (Song and Phoha, 2004), (Baohua et al, 2005), Bayesian networks (Songsiri, 2006) - (Nguyen et al, 2007), game theory (Michiardi and Molva, 2002), (Xiong and Liu, 2003), swarm intelligence (Jiang and Baras, 2004) - (Marmol et al, 2009), credential and policy (Santos et al, 2012) - (Olmedilla, 2007), and others (Lee and Turban, 2001) - (Su et al, 2010).

ANNEX B: ABREVIATIONS

EIDS	Electronic Identity management systems
OSSTMM	The Open Source Security Testing Methodology
OWASP	Open Web Application Security Project
NIST	National Institute of Standards and Technology
PTES	Penetration testing execution standard
ISSAF	Information Systems Security Assessment Framework
CSF	Cyber Security Framework
MAC	Mandatory Access Control
RBAC	Role-Based Access Control
ABAC	Attribute-Based Access Control
DAC	Discretionary Access Control
SSO	Single sign-on
MFA	Multi-factor Authentication
RADIUS	Remote Authentication Dial-In User Service
TACACS+	Terminal Access Controller Access-Control System Plus
SCIM	System for Cross-domain Identity Management
LCM	Lifecycle management
HR	Human-resources
BCP	Business Continuity Planning
DR	Disaster Recovery
PIA	Privacy Impact Assessment
BIA	Business Impact Analysis
CDD	Customer Due diligence
FATF	Financial Action Task Force
IAF	Identity Assurance Framework
TAL	Trustworthiness Assurance Levels

ANNEX C: Referenced Standards & Regulations

GDPR	General Data Protection Regulation
ISO27001L2013	Information Security Management System
ISO 27001	Information Security Management
ISO/IEC 27701	Security techniques -Privacy Information Management
ISO/IEC 29101:2013	Information technology -security techniques – privacy architecture framework
ISO/IEC 27550	Information technology - security techniques – privacy engineering for system lifecycle processes
ISO/IEC 24760	IT Security and Privacy — A framework for identity management — Part 1: Terminology and concepts
ISO/IEC 29115	Information technology — security techniques — Entity authentication assurance framework
ISO/IEC 29146	Information technology — Security techniques — A framework for access management
ISO/IEC 24760	IT Security and Privacy — A framework for identity management — Part 1: Terminology and concepts
ISO/IEC 17030:2003	Conformity assessment — General requirements for third-party marks of conformity
NIST SP 800-122	Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)
NIST CSF	Cyber Security Framework
ITU-T X1208 (01/2014)	A cybersecurity indicator of risk to enhance confidence and security in the use of telecommunication/information and communication technologies
Regulation on Privacy and Electronic Communications	UK law that implements the EU's ePrivacy Directive (Directive 2002/58/EC)
Cybersecurity Act	European Union Act on Cybersecurity
DPA	United Kingdom Data Protection Act
PCI-DSS	Payment Card Industry Data Security Standard
HIPAA (US)	Health Insurance Portability and Accountability Act