Cognitive Identity Management: Synthetic Data, Risk and Trust

S. Yanushkevich[1], A. Stoica[2], P. Shmerko[1], W. Howells[3], K. Crockett[4], R. Guest[3]
[1]*Biometric Technologies Laboratory, Department of Electrical and Computer Engineering, University of Calgary, Canada,* Web: http://www.ucalgary.ca/btlab, E-mail: syanshk@ucalgary.ca; peter.shmerko@ucalgary.ca
[2]*California Institute of Technology, California, Pasadena, USA, E-mail: adrian.stoica@jpl.nasa.gov*
[3]*School of Engineering and Digital Arts, University of Kent, U.K.,* E-mail: W.G.J.Howells@kent.ac.uk, R.M.Guest@kent.ac.uk
[4]*Department of Computing and Mathematics, Manchester Metropolitan University, U.K.,* E-mail: K.Crockett@mmu.ac.uk

***Abstract*—Synthetic, or artificial data is used in security applications such as protection of sensitive information, prediction of rare events, and training neural networks. Risk and trust are assessed specifically for a given kind of synthetic data and particular application. In this paper, we consider a more complicated scenario, – biometric-enabled cognitive cognitive biometric-enabled identity management, in which multiple kinds of synthetic data are used in addition to authentic data. For example, authentic biometric traits can be used to train the intelligent tools to identify humans, while synthetic, algorithmically generated data can be used to expand the training set or to model extreme situations. This paper is dedicated to understanding the potential impact of synthetic data on the cognitive checkpoint performance, and risk and trust prediction.**

Keywords: *Synthetic data, cognitive identity management, risk, trust, bias, computational intelligence*

I. Introduction

In Artificial Intelligence (AI) system design and development, *synthetic* data often replaces *authentic* data, or is used together with the latter. Synthetic data is generated from a population model, and used to test datasets, to validate mathematical models and to train machine learning algorithms. One problem is *'How well the synthetic data replicates the authentic data?'* [31], [32], [51]. In this paper, we formulate it as follows *'How risky this replacement?'* and *'Can we trust this synthetic 'life' attributes?'* Partial answers can be found in [5]. Our work advances this further and focuses on a specific application, – the cognitive biometric-enabled identity management.

A cognitive security checkpoint for identity management is a complex dynamic system [23], [52], [53]. Various performance projections of cognitive checkpoint include security measures, resistance to cyber attacks, public acceptability, depth of embedding in social infrastructure, type of biometric traits, links to forensics [29], intelligent models [25], [26], [28], as well as privacy, risk and trust (R&T) assessments [53], identity disclosure risks [2]. In this paper, we consider the AI performance evaluation in terms of Risk and Trust (R&T) under projections onto synthetic data. We distinguish the following kinds of synthetic data used in cognitive checkpoint:

1) *Synthetic biometrics* such as face [49], handprints [32], speech [39], signatures [15], iris [7], [27] for testing biometric algorithms [33] and modeling critical scenarios such as biometric attacks [27];
2) *Other synthetic data*, e.g. for the sensors that detect concealed (illicit) items; it is usually radar illumination to detect knives, pistols, grenades) [22], [43];
3) *AI decision assistance* such as avatar-like human-machine interfaces [52]; this concept is analogue to the engineered life form concept [1], [5].

Impact of synthetic data on performance and privacy in complex dynamical systems such as security checkpoint is a challenging problem. Various aspects of this problem were studied in [7], [32], [54], [55]. Privacy issues of synthetic data were discussed in [5]. However, effects of synthetic data involved in operational and decision-making processes in complex biometric-enable systems is an open problem. Our work contributed to this area.

This paper is organized as follows. In next Section II, we provide more detailed motivation of our study as analysis of potential impact of synthetic data on a checkpoint performance and formulate the problem. In Section III, an approach and contribution are explained. Definitions of the relevant concepts are given in Section IV. Our approach is introduced in Sections V through VI and explained using an experiment (VII). Section VIII concludes the paper.

II. Motivation and problem formulation

The goal of this paper is to develop an approach to exploration of synthetic data with respect to performance of cognitive checkpoints. Synthetic data are essential for modeling and operating of cognitive checkpoints. Specifically: 1) synthetic biometric traits are used in modeling and training various subsystems and scenarios; 2) Synthetic radar signals are used for modeling and training detectors of concealed items; and 3) Embodied AI decision assistants perform actions on behalf of an operator. These applications of synthetic data involve various kind of errors. A common property of these errors is that they are difficult to 'undo'. For example, mis-identification of a person of interest, mis-detection of concealed item, mis-detection of an attack, or other extreme scenarios may have catastrophic consequences.

This problem requires R&T assessment at all levels of identity management that utilizes synthetic data. Trust contributes to synthetic data acceptance, while risk contributes to its rejection. For example, to operate effectively on the human's behalf, embodied intelligent assistants might need confidential or sensitive information of the users such as financial details and personal contact information [9]. Acceptance of synthetic data in modeling is determined by the combination of both R&T factors [14], [47], [56]. The contributing factors include belief, confidence, experience, certainty, reliability, availability, competence, credibility, completeness, and cooperation [8]. Perception of R&T can be established quite independently, and together they determine the intelligent tools success.

There are three main reasons why the synthetic data are used in modeling and development of cognitive checkpoint:

Reason I: Authentic data, or large volumes of authentic data including only partially available scenarios, e.g. cyber attacks and biometric traits for training of deep learning tools;

Reason II: Critical (boundary) scenarios when it is impossible to obtain authentic data, e.g. rare events/scenarios such as impersonation, plastic surgery facial changes, and ageing process of biometric traits; and

Reason III: Privacy issues. There are two aspects: 1) AI decision assistants acquire, analyze, and accumulate privacy-sensitive data to make decision on identity and estimate R&T [5]; and 2) personal sensitive data can be replaced by synthetic data [41].

III. Contribution

This paper contributes to solving the two important challenges in identity management based on a cognitive security checkpoint model. The following research questions and approaches to their solution are detailed in this paper:

−How to distinguish the attributes of authentic vs. synthetic data?

Our Approach: The key instrument for this is the proposed taxonomy of the operational landscape.

−How to incorporate the R&T related to synthetic data into identity management process?

Our Approach: Given operations with R&T (causality detection, propagation, etc.) and causal network (e.g. Bayesian network), we propose to use Conditional Probability Tables (CPTs) as the carriers for the synthetic data R&T.

IV. Background

A cognitive security checkpoint for identity management is a complex dynamic system with the following elements of a cognitive system [23], [52], [53]:

Perception-cycle that enables information gain about the state of identified person,

Memory distributed across the entire system (personal data are collected in physical and virtual world),

Attention driven by memory to prioritize the allocation of available resources, and

Intelligence driven by perception, memories, and attention; its function is to enable the control and decision-making mechanism to identify intelligent choices. These cognitive elements are distributed in the form of a multi-state perception-action cycle semi-automated model [52].

In addition, a cognitive checkpoint is a privacy-sensitive model [9]. Note that a cognitive checkpoint is the most advanced form of the biometric-enable systems.

Definition 1: Risk is a measure of the extent to which an entity is threatened by a potential circumstance or event, and typically is a function of: (i) the adverse impact, or magnitude of harm, that would arise if the circumstance or event occurs; and (ii) the likelihood of occurrence [34]. Formally, `Risk=` F`(Impact, Probability)`. A protocol for writing a risk statement is the `Condition-If-Then` construct. Given events A and B, the part `Condition-If` of the risk statement is formally defined as $0 < P(A|B) = \alpha < 1$, where α is the probability risk A occurs given the conditioning event B (the root cause event) has occurred.

Definition 2: Trust is the willingness of the trustor (evaluator) to take risk based on a subjective belief that a trustee (evaluatee) will exhibit reliable behaviour to maximize the trustor's interest under uncertainty (e.g. ambiguity due to conflicting evidence and/or ignorance caused by complete lack of evidence) of a given situation based on the cognitive assessment [8], [34].

For example, the R&T identity management process includes the trusting behaviour, trusting intention, and trusting belief.

Definition 3: Trustworthiness is the degree to which an information system can be expected to preserve the confidentiality, integrity, and availability of the information being processed, stored, or transmitted by the system across the full range of threats [34].

For example, a trusted biometric sample acquisition system should satisfy a set of requirements such as resistance to: 1) fake biometric target presentation, 2) communication attack, and 3) acquisition system tampering. This is the bases for the trustworthy identity management in security checkpoints [55].

In our approach, R&T and trustworthiness are measured in terms of probabilities. These general notions of R&T should be specified within the concept of identity management at a cognitive checkpoint.

The typical goals of security checkpoint modeling are the following assessments:

- Performance under uncertainty and rare events/scenarios;
- R&T of the decision (correct identification of a given individual or his/her behavioral patterns), and
- Trust of the human operator in the decision supplied by the machine intelligence.

These R&T assessments are based on multiple criteria such as reliability of sources, credibility of information, sensor precision, recognition algorithm performance etc. These assessments vary among systems. Specifically, R&T assessment in a multi-state model is (a) distributed over states, (b) represented in causal relations, available for (c) propagation,

(d) adjustment, (e) fusion, and (f) prediction. The mechanism enabling these operations is known as probabilistic inference called machine reasoning [36], [37]. This mechanism with respect to cognitive checkpoint is explained in [53]. We will deploy this approach in this paper in Section VII.

V. Synthetic data exploration

Definition 4: Synthetic data is algorithmically generated data in order to model specific needs or conditions that are not available, e.g. privacy sensitive data, rare scenarios/events such as attacks, rare biometric traits, and decision-making of embodied intelligent assistance, as well as a large value of data for training and testing tools [5], [54].

Biometric data is the information extracted from biometric samples, or biometric traits.

Definition 5: Synthetic biometric traits are a class of algorithmically generated biometric characteristics (e.g., face and facial expressions, fingerprints, palmprints, iris, voice, and gait) used as source for constructing a biometric profile for purposes of identity management [19], [33], [51].

Synthetic data quality is evaluated using likelihood between the authentic and synthetic data, in terms of likelihood metrics of differences between authentic and synthetic features. An example of a such metric includes *utility* as a measure of worth, satisfaction, or preference of an outcome [41]):

$$\left\{ \begin{array}{c} \textbf{Authentic} \\ \textbf{Data} \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{c} \textbf{Utility} \\ \textbf{Measures} \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{c} \textbf{Synthetic} \\ \textbf{Data} \end{array} \right\}$$

It should be noted that several related works on general systematic understanding of synthetic data and synthetic (artificial, or engineered) 'life'. For example, in [1], synthetic 'life' forms are identified with various aspects of robotic learning using perception-action cycles. A more general vision of synthetic 'life' is provided in [48].

Separating the authentic biometric traits from synthetic ones is crucial in a security system development because this addresses the R&T of the identity management process [20].

Synthetic data are used in complex systems in various forms, at various levels of decision-making, and at various phases of system life cycle. In most cases,

1) Synthetic data are mixed with authentic data using computational intelligence operations such as R&T assessment, causal analysis, and reasoning.
2) It is impossible to separate the authentic data from synthetic data for purposes of analysis,
3) It is possible to observe an indirect impact of synthetic data on the outcome of the system. In our study, we explore this avenue in various modeling dimensions, including R&T estimation.

A. Synthetic data landscape

Given a multi-state cognitive checkpoint [53], different kinds of synthetic data are needed for purpose of modeling and development. For example,

- synthetic attacks help develop attack detectors and tools for their impact mitigation,
- synthetic facial expressions are needed for training deception detectors, and
- synthetic biometric watchlist helps improve security measures.
- detectors of concealed items are trained using a synthetic data such as illicit items (e.g. guns, knifes, grenades).

B. Synthetic data attributes

Synthetic, or artificial data are characterized by the following key attributes: 1) hierarchy, 2) life cycle, 3) relations to synthetic life, and 4) legality. Details are given in Table I. For example, synthetic biometric traits and synthetic radar illumination of concealed objects are characterized by a life cycle but they have the simplest relations with synthetic life attributes.

C. Extension of modeling dimensions

We propose the following three-step extension of modeling dimensions of the cognitive checkpoint through exploration of synthetic data:

Extension of modeling dimensions

```
Step 1: Specify the operational landscape.
Step 2: Define cognitive hierarchy of
    synthetic data.
Step 3: Provide the taxonomical view of the
    multi-state cognitive checkpoint.
```

Step I: Taxonomy of the operational landscape: The goal is to develop a taxonomical view on data used in a checkpoint. The operational landscape is identified and decomposed into three parts: 1) authentic, 2) synthetic, and 3) semi-synthetic attributes (Fig. 1):

- Authentic, synthetic and semi-synthetic data can be secured using encryption or cancellable techniques [17];
- Authentic biometric traits (evidence) are acquired from human, and synthetic biometric traits (synthetic evidence) embodied in complex system;
- Decision-making based on authentic (original) data (e.g. human biometrics) or synthetic data (e.g. embodied biometrics);
- Reasoning or judgment based on authentic or/and synthetic data (biometrics, evidence);
- Jurisdiction based on authentic or/and synthetic life forms.

Step II: Cognitive hierarchy of synthetic data: Synthetic data used in the checkpoint is divided as follows:

- *Primary synthetic forms* that are biometrics or non-biometrics generated using computational intelligence tools such as Generative Adversarial Networks (GANs) [10]; and
- *Advanced synthetic forms* such as embodied AI decision assistant [52].

TABLE I
KEY ATTRIBUTES OF SYNTHETIC DATA USED IN COGNITIVE CHECKPOINTS.

Attribute	Comments
Synthetic data hierarchy	Refers to synthetic data organization such as synthetic biometrics (face and facial expressions, fingerprints, signatures, etc.), synthetic attack, and rare event/scenarios.
Synthetic data life cycle	Systematic process that represents the main stages of synthetic data development such as 1) problem definition, 2) requirement formulation, 3) model formalization, 4) testing, and 5) implementation.
Relations to synthetic life	This is category of AI that mimics living systems or processes [48]. For example, the embodied intelligent assistant mimics humans in conversation, detection of emotion state, and deception features [51], [53].
Legality	In some cases, synthetic data may have a privacy problems such as a non authorized use of synthetic biometric traits in physical access systems, wrong trained embodied intelligent assistant [5].

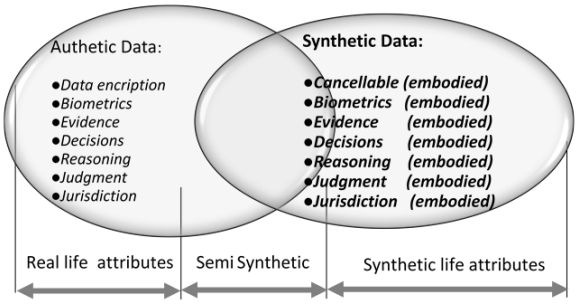

Fig. 1. Taxonomy of the operational landscape of cognitive checkpoint: real life attributes (left plane), synthetic 'life' attributes (right plane), and their intersection as semi-synthetic 'life' attributes.

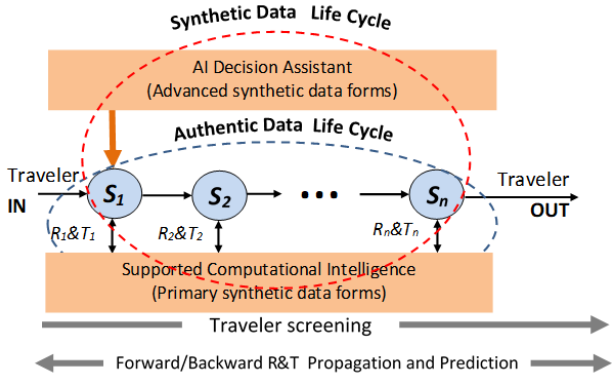

Fig. 2. Synthetic data hierarchy at the technology-independent model of a multi-state security checkpoint. Synthetic data primary forms (the operational landscape support) and advanced forms (embodied intelligent assistants).

The R&T privacy impact of these synthetic data forms is different and requires a different detection procedure. The focus of this paper is an impact of primary forms of synthetic data. Privacy impact and control of advanced forms of synthetic data produced by embodied AI decision assistant have been studied in [52] using a so called *Conflict Resolver* that is the *R&T impact detector* of decisions produced by the AI assistant. The trust dynamic of human and AI assistant interactions can be formalized, for example, using an approach [24]. Trust level variation is a probabilistic function of human experience $E(n) = 1-K(n)$, where $K(n)$ is human response on warning as the probability of a miss (ignore warning), or a false alarm (mistrust response).

Application of this synthetic data hierarchy to the security checkpoint model is illustrated in Fig. 2. The R&T of a subject (traveler and related identity attributes) is assessed using various mechanisms such as forward R&T propagation (cause and effect process), and backward R&T propagation (from cause to effect) through the system states. R&T states are adjusted using their causal relationships. This is the core principle of traveler's risk mitigation. Given the R&T scores and the screening resources at each state, the R&T fusion results in a final decision. Data life cycle includes primary and advanced forms. They impact the authentic data life cycle.

Step III: Taxonomical view of the multi-state cognitive checkpoint: We consider a multi-state screening in the dynamic cognitive system that:

1) Monitors the traveler data throughout the process of e-ID checking, face recognition, and continuously assess the R&T using various sources such as behavioral biometrics, watchlist, AI decision assistant results, etc.,
2) Updates its states based on the intelligence gathered via human-machine interactions (AI decision assistant), results of the biometric traits recognition based on machine learning, results of the concealed object detection (by adjusting radar illumination, in particular), and others.

Fig. 3 shows an example of the aforementioned functions in the context of the cognitive identity management for travelers crossing the borders. The traveler's identity management process is implemented in three states, S_1 (ID validation), S_2 (Traveler authentication), and S_3 (Concealed object detection). Each state contains several sub-states. There are two types of dependency relationships that exist between states S_i and sub-states: *intra-iteration* dependency (sub-states in the same loop), and *cross-iteration* (previous states) dependency. Each state S_i and sub-state is a part of the 'Layered Security Strat-

egy', a contemporary security doctrine [44]. Each state S_i and sub-state generates R&T assessments for further processing and inference using operations such as propagation, causal analysis, reasoning, etc.

Because R&T are measured as probability events, they can be *fused* or combined, and *propagated*. These two operations are the core of two strategies for manage the R&T: *Forward propagation* reflects R&T assessment process from causes to effect, and *Backward propagation* reflects R&T assessment process from the effect to the causes. Forward and backward R&T propagation process provides the systematic evaluation on traveler R&T caused by various threats, hazards, and concerns, and cost-effective measures for lowering risk to an acceptable level. Note that risks propagation can create a ripple effect generating further risks across the network with an amplified impact such as snowball effect.

Synthetic data is required at various computational intelligence operations and processes. For example, the sub-state $S_m^{(1)}$ of state S_m is defined under *learnt* ID source reliability using authentic data from previous experience, while the data of potential attacks is synthesized. This enables assessment of the R&T of a such rare events (attacks). Transition between states S and $S^{'}$ with action α is denoted as $S \stackrel{\alpha}{\rightarrow} S^{'}$. A screening trace $\delta =< \alpha_1, \alpha_2, \ldots, \alpha_k >$ is a sequence of actions such that there exists an execution $S_0 \stackrel{\alpha_1}{\Rightarrow} S_1, \ldots, S_{k-1} \stackrel{\alpha_k}{\Rightarrow} S_k$.

Note that each state operates as a cognitive agent that makes a decision regarding the user R&T based on the specific resources such as previous experience (statistics) and observed information.

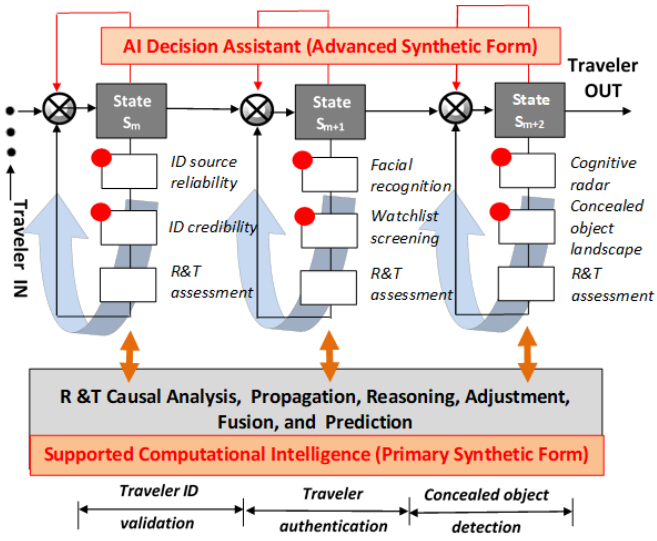

Fig. 3. Taxonomical view of the multi-state cognitive identity management process. Synthetic data are distributed in various forms and decision levels. R&T of synthetic data are assessed in complex causal relationships with authentic data. R&T are propagated from the first state (input) to the last state (output), and at each state their R&T status is causal analyzed, adjusted, fused, and predicted. Each state is represented by a perception-action cycle of sub-states. Red circle marks the sub-states where synthetic data are used.

VI. Impact Assessment of Synthetic Data

This section introduces the final phase of our approach. It was shown in previous sections that 1) Probabilistic nature of R&T assessments; and 2) There is a mechanism exists in a multi-state identity management that operate with R&T assessments such as causal analysis, propagation, prediction, reasoning, adjustment, and fusion (Fig. 3); Our idea is to explore this mechanism for the assessment of synthetic data impact on performance of a checkpoint. For this, R&T assessments of sources of synthetic data (red circles in Fig. 3) should be integrated in general checkpoint model developed in [52], [53].

A. Bridging models: supply chain and identity profiling

Multi-state representation of identity profiling (Fig. 3) is a framework for bridging this model with the advanced supply chain models.

Supply chain R&T is an event-oriented concept in which R&T strongly relates to the probability and consequence of a potentially harmful event. The aim of modeling is to understand the dynamic behavior of the supply chain facing R&T of supply chain disruptions.

In supply chain R&T management, standard R&T model is used that consists the following attributes: R&T event, event driver, probability of R&T event, impact (R&T), impact driver, impact probability, and total loss [8], [56].

B. The R&T theoretical framework

The R&T theoretical framework of the multi-state identity management process consists of three carriers: 1) advances in supply chain R&T management, 2) emergent needs in identity R&T management, and 3) theoretical fundamentals of bridging these approaches. This view is represented in Table II. Table II (a) represents the research landscapes that we intend to bridge: the left panel represents 'Advances in Supply Chain R&T Management' and the left one 'Emergence Needs of Identity R&T Management'; (b) surveys the advances in "Supply Chain R&T Management'; and (c) provides 'Theoretical fundamentals for bridging the approaches'. This bridging approach enables: 1) separation of cause-effect paths, 2) propagating R&T through these paths, and 3) predict R&T using appropriate networks. The following similarities enables this bridging: (a) The multi-echelon supply chain is similar to the multi-state identity management process; (b) Performance evaluation is based on probabilistic R&T notions, and (c) R&T management assumes the R&T causal-effect path discovery over a set of probabilistic operations. The main mechanism of R&T assessment in both models is the separation of causal paths. That is, carrier of R&T, including various kinds of synthetic data R&T, is the *causal-effect R&T path*.

VII. Motivational Experiment

The motivational experiment aims at highlighting 1) Practical details of detecting the synthetic data impact on identity management process, and 2) Needs of breakthrough solutions for meta-detection and meta-recognition.

TABLE II
BRIDGING ACHIEVEMENTS IN SUPPLY CHAIN R&T MANAGEMENT AND EMERGENCE NEEDS OF IDENTITY R&T MANAGEMENT.

Advances in Supply Chain R&T Management	**Emergence Needs of Identity R&T Management**
1. *R&T formalization*, *assessment*, *adjustment*, and *reasoning*, e.g. [3], [18], [35], [45]	The core of an identity management is the ability to form an intelligent conclusion or judgment under uncertainty.
2. *R&T propagation*, e.g. [3], [6], [18], [35], [45], [56]	In multi-state security checkpoint, the R&T assessment should be propagated (forward-backward) through other states, and adjusted.
3. *R&T prediction*, e.g. [56]	R&T should be predicted
4. *R&T causal analysis*, e.g. [14]	The 'cause-effect' paradigm (e.g. Bayesian causality analysis [36], [42]) plays the crucial role in identity management.
Theoretical fundamentals of bridging R&T managements and extension	
• Probabilistic reasoning, e.g. Bayesian networks [36], [37]; • Causal analysis, e.g. Granger causality [21]; • Probabilistic fusion, e.g. copula [16];	• Extrema value theory, and discovery detection, e.g. [4]; • Meta analysis and recognition, e.g. [40];

The common real-world scenario of the ID management is chosen for this purpose: Given an e-ID, assess the ID information credibility, that is `<Credibility>` ≡ `<Trustworthiness>` + `<Expertise>`.

A. Inference engine for identity management scenario

Assessment of ID information credibility is represented in Fig. 4 in the form of Bayesian network and corresponded Conditional Probability Tables (CPTs) where:

Node *ID source reliability'* ($R \in \{r_1, r_2, r_3\}$) denotes the three level (r_1 ='high', r_2 ='medium', r_3 ='low') reliability of the e-passport/ID authentication, which depends on many risk factors such as country of issue, number of defense levels in the document, life cycle history, type of the chip, type of biometric modality, type of encryption, and the type of RFID mechanism.

Node *'Valid ID'*, or *'Trusted ID'* ($V \in \{v_1, v_2\}$) denotes whether the e-passport ID should pass the validation procedure (valid v_1) or not (invalid v_2). The 'valid' or 'invalid' state reflects the true state of the e-passport using factors such as watermarks, holograms, ultra violet threads, micro text, and optical variable ink.

Node *'ID validation'* ($S \in \{s_1, s_2, s_3, s_4\}$) denotes the outcome of the authentication of the e-passport. The scan is subject to various unwanted effects such as the individual's mistakes in using the scanning device, scanner errors, as well as hidden reasons related to errors in the use of the database, conflicts of comparisons, and communication errors or delays. These effects are encoded in the form of the number of attempts at scanning the individual document; three attempts are allowed (s_1, s_2, s_3), after which the individual is directed to manual control (s_4).

Node *'ID credibility'* ($C \in \{c_1, c_2, c_3\}$) describes the three level (c_1 ='high', c_2 ='medium', c_3 ='low') credibility of the outcome of the validation process. If the credibility of the validation process is known a priori, it can be used to compute posterior beliefs related to the validity of the individual document (node V).

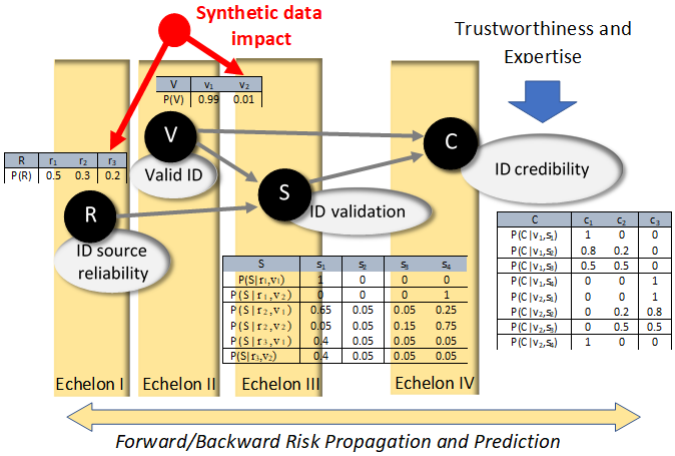

Fig. 4. Assessment of ID credibility (trustworthiness and expertise) using IV-echelon (state) identity management scenario and its implementation as 4-node Bayesian network. Synthetic data impact is incorporated using CPTs for nodes R and V.

As an example, consider the following particular scenario: `IF the reliability of the ID source is known to be 'low' and the resulting credibility to be 'high':` $R = r_3$, `and` $C = c_1$, `THEN what is the posterior probability that the ID is valid:` $P(V = v_1 | R = r_3, C = c_1)$. This scenario models a situation of conflict where an unreliable source produces a credible outcome. The final result is $P(V = v_1 | R = r_3, C = c_1) \approx 0.989$. It is very likely that the ID was valid. That is, trustworthiness of statement 'the ID was valid' making over expert knowledge (incorporated in algorithms) is 98.9%.

B. Synthetic data traits

Let us assume that in training algorithms for validation of ID (node V) and identification of ID source reliability (node R), synthetic data was used to represent rare events such as false ID, multiple ID of the same person, and features of intentional data alteration in the chip (e.g. biometric traits and text data) as well as a false life cycle history [50]. Probabilities of these threats are represented in Conditional Probabilities Tables (CPTs) for nodes V and R: $P(V = v_2) = 0.01$ and $P(R = r_3) = 0.2 \equiv \texttt{low}$ (Fig. 5). There is always risk that the validation algorithm makes a mistake should the real rare event occur. For example, features of forgery e-ID are not detected, or these features can be mistakenly detected in valid ID. The goal is to assess these risks caused by usage of synthetic data.

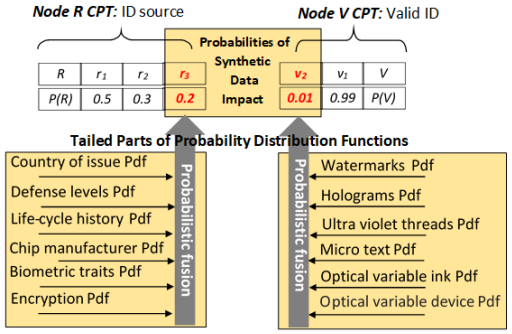

Fig. 5. Assessment of primary synthetic data impact at CPTs as fused Pdf(s) of R&T factors. Left plane corresponds the impact $P(R = r_3) = 0.2$ ('ID source reliability') that represents Pdf(s) of rare events/scenarios models such as country of issue, number of defense levels, life cycle history, chip manufacturer, etc. Right plane describes impact $P(V = v_2) = 0.01$ ('Valid ID') that represents Pdf(s) of rare events/scenarios models such as watermarks, holograms, ultra violet threads, micro text, etc. Rare events/scenarios models are represented by tails of corresponded Pdf(s), that are fused in order to obtain summaries of synthetic data impacts.

C. Algorithm

The goal of this motivational experiment is to assess the impact probabilities shown in Fig. 5. The underlying assumption is that these probabilities addresses the worst-case scenarios contain valuable information about synthetic data impact. In this section, we introduce the algorithm for assessment of synthetic data impact at identity profiling.

It is well understood that the frequency of object occurrence in identity management process follows a long-tail distribution. For example, people with true IDs and expired IDs are much more common than people with false IDs and multiple IDs. This problem addresses the novelty detection (known also as anomaly detection, or one-class classification), – the task of recognizing that test data differ in some respect from the data that are available during training [38]. Theoretical framework is extreme value theory, – a branch of statistics analyzing the distribution of data of abnormally high or low values. Tailed probability distributions have been used, for example, in study cyber-risks such as ID theft [30].

Specifically, Fig. 5 represents the framework of the algorithm for computing the values of CPTs that addresses rare events/scenarios. This is a part of Bayesian causal network (Fig. 4). The algorithmic description is provided below; Pdf denotes the probability distribution function.

Standard copula based technique is used for the Pdf fusion [16]. This algorithm detects synthetic data impact at the CPT level such as $P(R = r_3)$ and $P(V = v_2)$. However, an extension is needed for evaluating the impact at the reasoning level of Bayesian network, that is $P(C \in \{c_1, c_2, c_3\})$ in our case.

Algorithm for assessment of synthetic data impact on identity profiling

Input: The Pdf(s) of R&T factors for node R and V (synthetic data)
Output: The Pdf(s) of synthetic data impacts for CPTs for node R and V
Step 1: Fuse the Pdf(s) of R&T factors for node R; result is the CPT value $P(R = r_3)$
Step 2: Fuse the Pdf(s) of R&T factors for node V; result is the CPT value $P(V = v_2)$

VIII. Summary and Conclusion

Synthetic data is an integrated part of complex biometric-enable systems. Their impact on system performance can be unpredictable, for example, a neural network that was trained using synthetic data of unknown quality can produce erroneous results [5]. In existing studies [19], [51], synthetic data were not certified qualitatively, and their use was quite casual. In the real world multi-state identity management, various kinds of synthetic data (including non-certified) are employed through the system.

The key results reported in this paper suggest that:

– To distinguish the attributes of authentic versus synthetic data, one shall use the proposed taxonomy of the operational landscape.

– To evaluate the impact of synthetic data on identity management process, one shall conduct the R&T analysis of synthetic data in terms of impact on privacy. We have shown that this is translated to 1) the statistical inference problem, 2) the belief propagation model (causal network), and 3) the novelty detection technique.

The future work in this direction will include study of various biases, e.g. biases influencing face recognition [12], and their impact on the R&T process. Such biases can be introduced by the models used to generate synthetic data.

Acknowledgments

This Project was partially supported by the European Union's Horizon 2020 research [11], [13]; Natural Sciences and Engineering Research Council of Canada (NSERC) through grant "Biometric-enabled Identity management and Risk Assessment for Smart Cities", and the Department of National Defence's Innovation for Defence Excellence and Security (IDEaS) program, Canada. The authors acknowledge Eur Ing *Phil Phillips*, CEng, for useful suggestions.

REFERENCES

[1] A. Adamatzky and M. Komosinski (Eds). *Artificial Life Models in Hardware*, New York, Springer, 2009.

[2] A. Andreou, O. Goga, and P. Loiseau, Identity vs. Attribute Disclosure Risks for Users with Multiple Social Profiles, *Proc. IEEE/ACM Int. Conf. Adv. Soc. Net. Anal. and Mining*, 2017, pp. 163–170.

[3] C. Baudrit, D. Dubois, and D. Guyonnet, Joint Propagation and Exploitation of Probabilistic and Possibilistic Information in Risk Assessment, *IEEE Trans. Fuzzy Sys.,* vol. 14, no. 5, 2006, pp. 593–608.

[4] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers, *Statistics of Extremes: Theory and Applications*, New York, Wiley, 2004.

[5] S. M. Bellovin, P. K. Dutta, and N. Reitinger, Privacy and Synthetic Datasets, *Stanford Tech. Law Rev.*, vol. 22, no. 1, 2019, pp. 1–52.

[6] A. Bueno-Solano, and M. G. Cedillo-Campos. Dynamic Impact on Global Supply Chains Performance of Disruptions Propagation Produced by Terrorist Acts. *Transp. Res. Part E: Logistics and Transp.*, Review 61, 2014, pp. 1–12.

[7] L. Cardoso, A. Barbosa, F. Silva, *et al.*, Iris Biometrics: Synthesis of Degraded Ocular Images, *IEEE Trans. Inf. Forensics and Security*, vol. 8, no. 7, 2013, pp. 1115–1125.

[8] J.-H. Cho, K. Chan, and S. Adali, A Survey on Trust Modeling, *ACM Comp. Surv.*, vol. 48, no. 2, Article 28, 2015.

[9] G. G. Clavell, Protect rights at automated borders *Nature,* vol. 543, Issue 7643, March, 2017.

[10] A. Creswell, T. White, V. Dumoulin, *et al.*, Generative Adversarial Networks: An overview, *IEEE Sig. Proc. Mag.*, Jan. 2018, pp. 53–65.

[11] H2020 Intelligent SMART Border Control (iBorderCtrl), https://www.iborderctrl.eu/

[12] A. Das, A. Dantcheva, and F. Bremond, Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach, *Proc. Eur. Conf. Comp. Vision,* 2019, pp. 573–585.

[13] European Union: Technical Study on Smart Borders, *EU, European Commission*, B-1049, Brussels, 2014.

[14] N. Feng, H. Wang, and M. Li, A security risk analysis model for information systems: Causal relationships of risk factors and vulnerability propagation analysis, *Inf. Sci.*, vol. 256, 2014, pp. 57–73.

[15] M. Ferrer, M. Diaz-Cabrera, and A. Morales, Static Signature Synthesis: A Neuromotor Inspired Approach for Biometrics, *IEEE Trans. Pattern Anal. and Mach. Intel.*, vol. 37, no. 3, 2015, pp. 667–680.

[16] E. W. Frees and E. A. Valdez, Understanding relationships using copulas, *North American Actuarial J.,* vol. 2, no.1, 1998, pp. 1–25.

[17] Q. Gao and C. Zhang, Constructing cancellable template with synthetic minutiae, *IET Biometrics*, vol. 6, Iss. 6, 2017, pp. 448–456.

[18] M. D. Garvey, S. Carnovale, and S. Yeniyurt, Analytical framework for supply network risk propagation: A Bayesian network approach, *European J. Operational Res.,* vol. 243, 2015, pp. 618–627.

[19] M. Gomez-Barrero, J. Galbally, Reversing the Irreversible: A Survey on Inverse Biometrics, *Comp.& Security*, vol. 90, March 2020.

[20] C. Gottschlich and S. Huckemann, Separating the authentic from the synthetic: minutiae histograms as fingerprints of fingerprints, *IET Biometrics*, vol. 3, Iss. 4, 2014, pp. 291–301.

[21] C.W. Granger, Investigating causal relations by econometric models and cross-spectral methods, *J. Econ. Soc.*, 1969, pp. 424–438.

[22] S. W. Harmer, N. Bowring, D. Andrews, *et al.,* A Review of Non-imaging Stand-Off Concealed Threat Detection with Millimeter-Wave Radar, *IEEE Microwave Magazine*, Jan./Feb., 2012, pp. 160–167.

[23] S. Haykin, *Cognitive Dynamic Systems (Perception-Action Cycle, Radar, and Radio)*, New York: Cambridge University Press, 2012.

[24] W.-L. Hu, K. Akash, T. Reid, N. Jain, Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions, *IEEE Trans. Human-Machine Syst.*, vol.49, no. 6, 2019, pp. 485–497.

[25] IATA (International Air Transport Association): Checkpoint of the future. Executive summary. 4th Proof. 2014.

[26] IATA (International Air Transport Association): Automated Border Control. Implementation Guide, 2015.

[27] N. Kohli, D. Yadav, M. Vatsa, *et al.*, Synthetic Iris Presentation Attack using iDCGAN, arXiv:1710.10565v1 [cs.CV] 29 Oct. 2017.

[28] R. D. Labati, A. Genovese, E. Munoz, *et al.,* Biometric Recognition in Automated Border Control: A Survey, *ACM Comp. Surv.*, vol.49, no.2, 2016, pp. A1-A39.

[29] K. Lai, S. Yanushkevich, V. Shmerko, and S. Eastwood, Bridging the Gap Between Forensics and Biometric-Enabled Watchlists for e-Borders, *IEEE Comput. Intell. Mag.*, vol. 12, no. 1, 2017, pp. 17–28.

[30] T. Maillart and D. Sornette, Heavy-tailed distribution of cyber-risks, *Eur. Phys. J. B,* vol 75, 2010, pp. 357–364.

[31] N. Mayer, E. Ilg, P. Fischer, *et al.,* What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation? *Int. J. Comp. Vis.*, vol. 126, 2018, pp. 942–960.

[32] A. Morales, R. Cappelli, M. A. Ferrer, and D. Maltoni, Synthesis and Evaluation of High Resolution Hand-Prints, *IEEE Trans. Information Forensics and Security*, vol. 9, no. 11, 2014, pp. 1922–1932.

[33] T. M. Murphy, *et al.* Use of synthetic data to test biometric algorithms, *J. Elect. Imag.*, vol. 25, no. 4, 2016, pp. 1–11.

[34] National Institute of Standards (NIST), Security and Privacy Controls for Information Systems and Organizations, NIST Special Publication 800-53, Revision 5, 2017.

[35] R. Ojha, A. Ghadge, M. K. Tiwari, and U. S. Bititci, Bayesian network modelling for supply chain risk propagation, *Int. J. Production Research,* vol. 56, no. 17, 2018, pp. pp. 5795–5819

[36] J. Pearl, The Seven Tools of Causal Inference, with Reflections on Machine Learning, *Com. ACM,* vol. 62, no. 3, 2019, pp. 54–60.

[37] J. Pearl, Causal inference in statistics: an overview, *Stat Surv.*, vol.3, 2009, pp. 96–146.

[38] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, A review of novelty detection, *Signal Processing*, vol. 99, 2014, pp. 215–249.

[39] J. Sanchez, I. Saratxaga, I. Hernaez, *et al.*, Toward a Universal Synthetic Speech Spoofing Detection Using Phase Information, *IEEE Trans. Inf. Forensics and Security*, vol. 10, no. 4, 2015, pp. 810–820.

[40] W. J. Scheirer, A. Rocha, R. J. Micheals, *et al.,* Meta-Recognition: The Theory and Practice of Recognition Score Analysis, *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 33, no. 8, 2011, pp. 1689–1695.

[41] J. Snoke, G. M. Raab, B. Nowok, *et al.,* General and specific utility measures for synthetic data, *J. of the Royal Statist. Soc., Series A*, vol. 181, Part 3, 2018, pp. 663–688.

[42] P. Spirtes and K. Zhang, Causal discovery and inference: concepts and recent methodological advances, *Appl. Inform.,* 2016, 3, issue 3.

[43] T. Truong, and S. Yanushkevich, Generative Adversarial Network for Radar Signal Generation, *Int. Joint Conf. Neural Networks (IJCNN)*, Budapest, 2019, pp.1–6.

[44] Transportation Security Administration, *Layers of Security,* 2013. Available at: http://www.tsa.gov/about-tsa/layerssecurity

[45] P. Victor, C. Cornelis, M. de Cock, *Trust Networks for Recommender Systems,* Atlantis Press, 2011.

[46] Y. Wang and M. P. Singh, Evidence-Based Trust: A Mathematical Model Geared for Multiagent Systems, *ACM Trans. Autonomous and Adaptive Sys.*, vol. 5, no. 4, Article 14, 2010.

[47] M. Whittaker, *et al.*, AI Now Report, New York University, 2018, https://ainowinstitute.org/AI_Now_2018_Report.pdf

[48] R. V. Yampolskiy, On the origin of synthetic life: attribution of output to a particular algorithm, *The Royal Swedish Academy of Sciences, Phys. Scr,* vol. 92, 2017, pp. 1–10.

[49] R. V. Yampolskiy, B. Klare, and A. K. Jain, Face Recognition in the Virtual World: Recognizing Avatar Faces, *Proc. 11th Int. Conf. Mach. Learning and Appl.*, 2012, pp. 40–45.

[50] B. Yang, C. Busch, J. Bringer, *et al,* Towards Standardizing Trusted Evidence of Identity, *Proc. ACM workshop on Digital identity management*, Berlin, Germany, 2013, pp. 63–72.

[51] S. Yanushkevich, A. Stoica, and V. Shmerko, Developmental Tools - Synthetic Biometrics, *IEEE Comp. Intel. Mag.*, vol. 2, no. 2, 2007, pp. 60–69.

[52] S. Yanushkevich, K. Sundberg, N. Twyman, *et al.,* Cognitive checkpoint: Emerging technologies for biometric-enabled watchlist screening, *Comp. and Security*, vol. 85, 2019, pp. 372–385.

[53] S. Yanushkevich, W. Howells, K. Crockett, *et al.,* Cognitive Identity Management: Risks, Trust and Decisions using Heterogeneous Sources, *Proc. IEEE Int. Conf. Cognitive Mach. Intel.*, Los Angeles, 2019.

[54] S. Yanushkevich, N. Reitinger, A. Stoica, *et al.,* Inverse Biometrics: Privacy, Risks, and Trust, In: *Encyclopedia of Cryptography*, Security and Privacy, S. Jajodia, *et al.,* Eds., Springer, 2020.

[55] S. Yanushkevich, *Fundamentals of Biometric System Design*, Lecture notes, University of Calgary, Canada, 2020, http://www.ucalgary.ca/btlab

[56] R. Zhang and Y. Mao, Trust Prediction via Belief Propagation, *ACM Trans. Inf. Sys.*, vol. 32, no. 3, Article 15, 2014.