



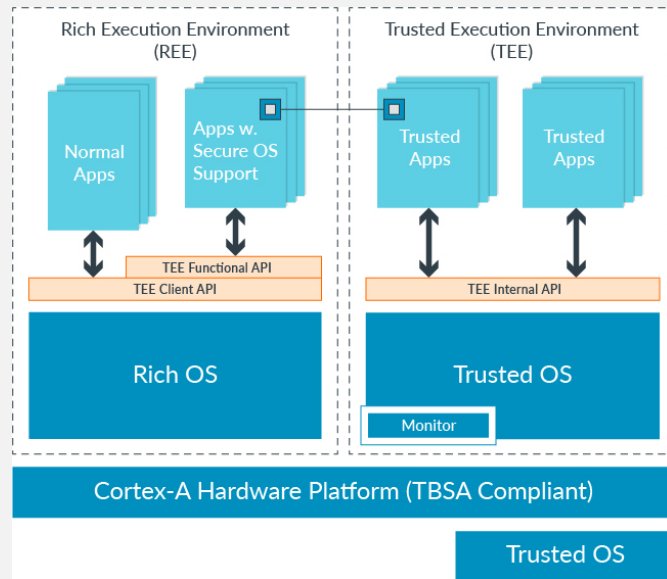
Trustworthy AI – Practical Collaborative Engineering

Results of industry consultation
TAIBOM: a foundation for trustable AI

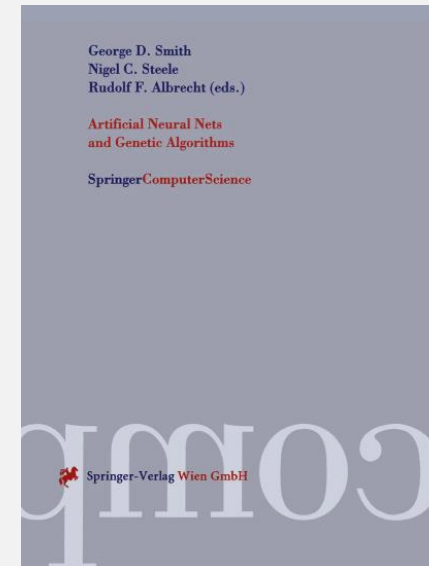
Dr Nicholas Allott
nick@nqminds.com

Trust

AI



Hardware trust foundation for modern computing



George D. Smith
Nigel C. Steele
Rudolf F. Albrecht (eds.)

Artificial Neural Nets
and Genetic Algorithms

Springer Computer Science



Connectionism and Symbolism in Symbiosis

N. Allott, P. Fawcarter and P. Hales
Computing Department, The Nottingham Trent University, Burton St,
Nottingham, NG1 4BU, England
Email: nma@ntu.ac.uk

Abstract

In this paper we examine a previously published algorithm which addresses the problem of network growth by implementing a clustering algorithm to operate on time dependent data. The computational constraints of the problem forced the development of an architecture, which in retrospect can be analysed in terms of a computational and symbolic module operating systematically. Here we attempt to identify the computational constraints that necessitate the use of the architecture, and any further merits it has. Further, we analyse the nature of the interaction between the two modules and highlight the manner in which the behaviour the symbolic module correlates with what is known of human problem solving behaviour.

1 Introduction

This paper both develops and outlines the computational merits of a symbolic symbolic and connectionist network used within previously published clustering algorithm [1]. Within this algorithm a connectionist network was used to embody the relationship between discrete observable elements, for example letters of the alphabet. In the simplest case the relationship modelled in the relative statistical distribution, or context, of the letters. As such the network can be shown to be very similar to *N*-gram analysis [3, 7, 10] or the transition network of a Markov model [6]. However it is superior in that the scope of analysis to be considered does not have to be specified globally, but is dynamically and locally determined for each node.

An algorithm was required to produce these connectionist trees from empirical data. It was the original intention that the algorithm be developed within the context of the connectionist paradigm. By this we mean capable of being implemented in a parallel manner such that, the functions used to compute the working parameters for each node have access only to those items to which the node is ar-

chitecturally linked (such as the back propagation algorithm or simple Hebbian learning). However the fact that (a) the algorithm attempts to grow the network (b) time dependent data was being handled, made this design goal difficult to satisfy. In the next section we attempt to formalise the source of the difficulty.

2 Formal Definition of Problem

The network can be characterised as an *n*-tuple $\langle P, N, L, \alpha, \Sigma \rangle$. Where
 P is the set of primitive nodes,
 N is the set of all nodes, initially $N = P$,
 L is the set of links between nodes, initially $L = \emptyset$,
 and each element of L is a 3-tuple $\langle n, c, n' \rangle$, parent, child, strength, where $n = 1$,
 α is the activation function, $\alpha : N \times L \times \Sigma^* \rightarrow \{0, 1\}$
 Σ is the set of possible primitive evidence.

Derived from this we have:
 Σ^* the set of sentences possible from Σ ,
 $P(N)$ the power set of all nodes,
 A the set of abstract nodes, defined $A = N \cap P^*$,
 α is the set of children of a node, and can be defined in terms of L and N .

To simplify the problem, in the initial case the strength of all links is assumed to be 1, and the activation function is boolean returning $\{0, 1\}$.

If $S \in \Sigma^*$ then $\sigma(S) \in \Sigma$ and is the first element of S . It follows for simple sequence analysis (such as the text string discussed above) where there is a 1:1 mapping between P and Σ , the activation function for a node n , where $n \in P$,

$$\alpha(n, l, S) = \begin{cases} 1, & \sigma(S) = n \\ 0, & \sigma(S) \neq n \end{cases}$$

And for node n where $n \in A$, the activation is some function $f()$ of the activation of the children of n , $\alpha(n, l, S) = f(\sigma(n), S, l)$.

Connectionism and Symbolism in Symbiosis

Known unknowns

What is: Trustworthy AI?

How to build Trustworthy AI?

Practical Collaborative Engineering

Consultation



Your invitation to the Engineering Trustworthy AI Workshop

**Caledonian Club
London**

REGISTER HERE:

[TECHWORKS.ORG.UK/EVENT/ENGINEERING-TRUSTWORTHY-AI-WORKSHOP](https://techworks.org.uk/event/engineering-trustworthy-ai-workshop)



ON TUESDAY, 18TH JULY 2023



12:00 - 16:30 BST

Questions descoped

- What is AI?
- Can the outputs of an AI system be explained? Is there transparency in the decision process?
- How do we ensure fairness and prevent bias?
- Does UK industry have access to necessary capabilities (skills etc)?
- Does the UK have the infrastructure to train large AI models?
- Compared to traditional/contemporary systems, how complex are AI systems?

<https://www.techworks.org.uk/wp-content/uploads/2024/01/Engineering-Trustworthy-AI.pdf>

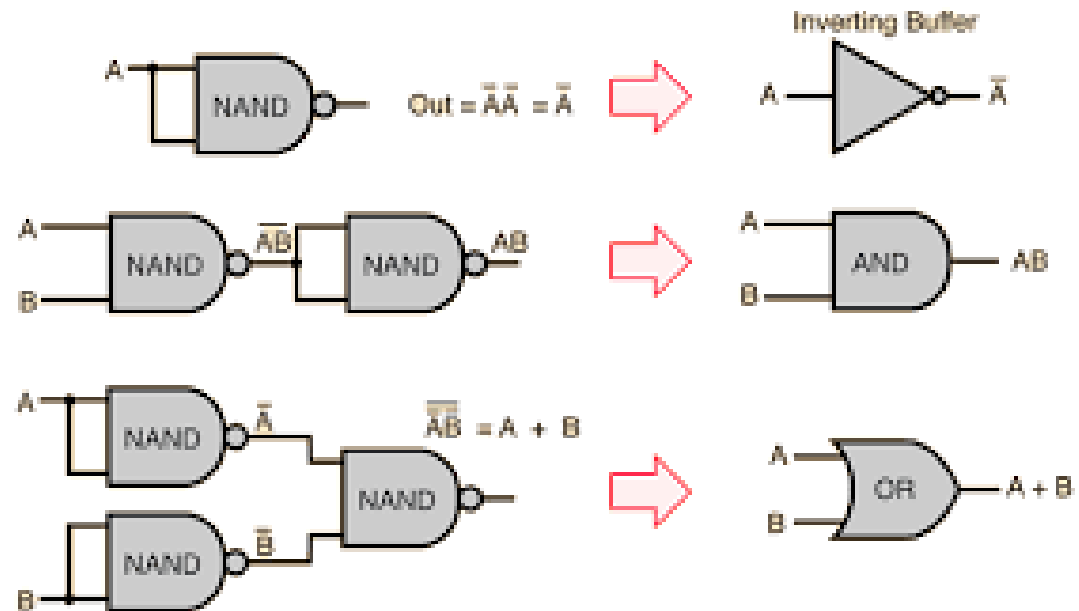
Questions in scope

- How do you know if an AI system's behaviour will be fundamentally changed?
- How can we encourage best practice amongst AI system developers and AI users?
- How do we confidently test an AI system?
- Who is liable if/when damage is done by an AI system?
- Is there a notion of consent to train?
- Where did your AI system come from?
- Trustworthiness is dependent on the hardware we train and run the AI systems on
- Trustworthiness is dependent on the software we train and run the AI systems on
- What is the system being used for?
- Are assertions of trustworthiness inherently subjective?
- How do we ensure supply of good quality and requisite quantity of data for the UK industry?
- How can we ensure our AI system fails in a controlled/defined manner?

Framing the Problem

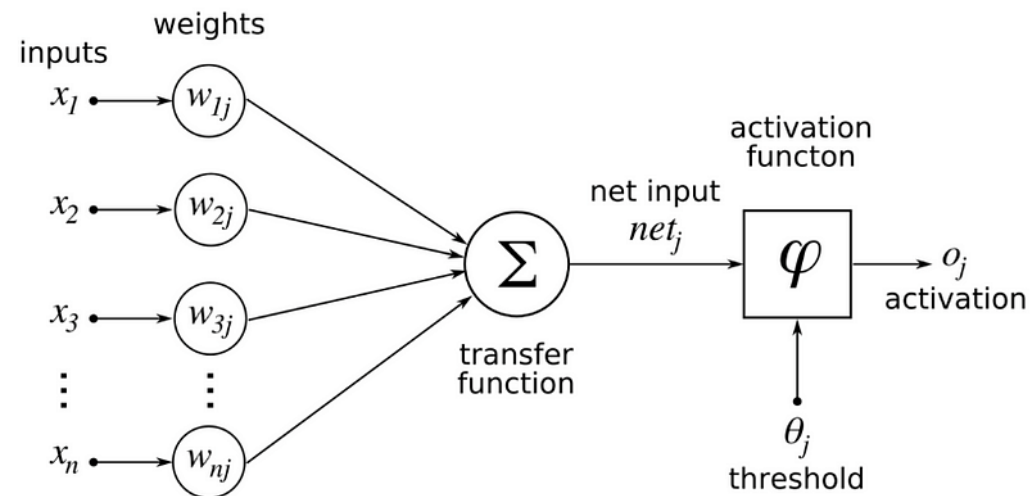
Logical foundation

Most modern computing



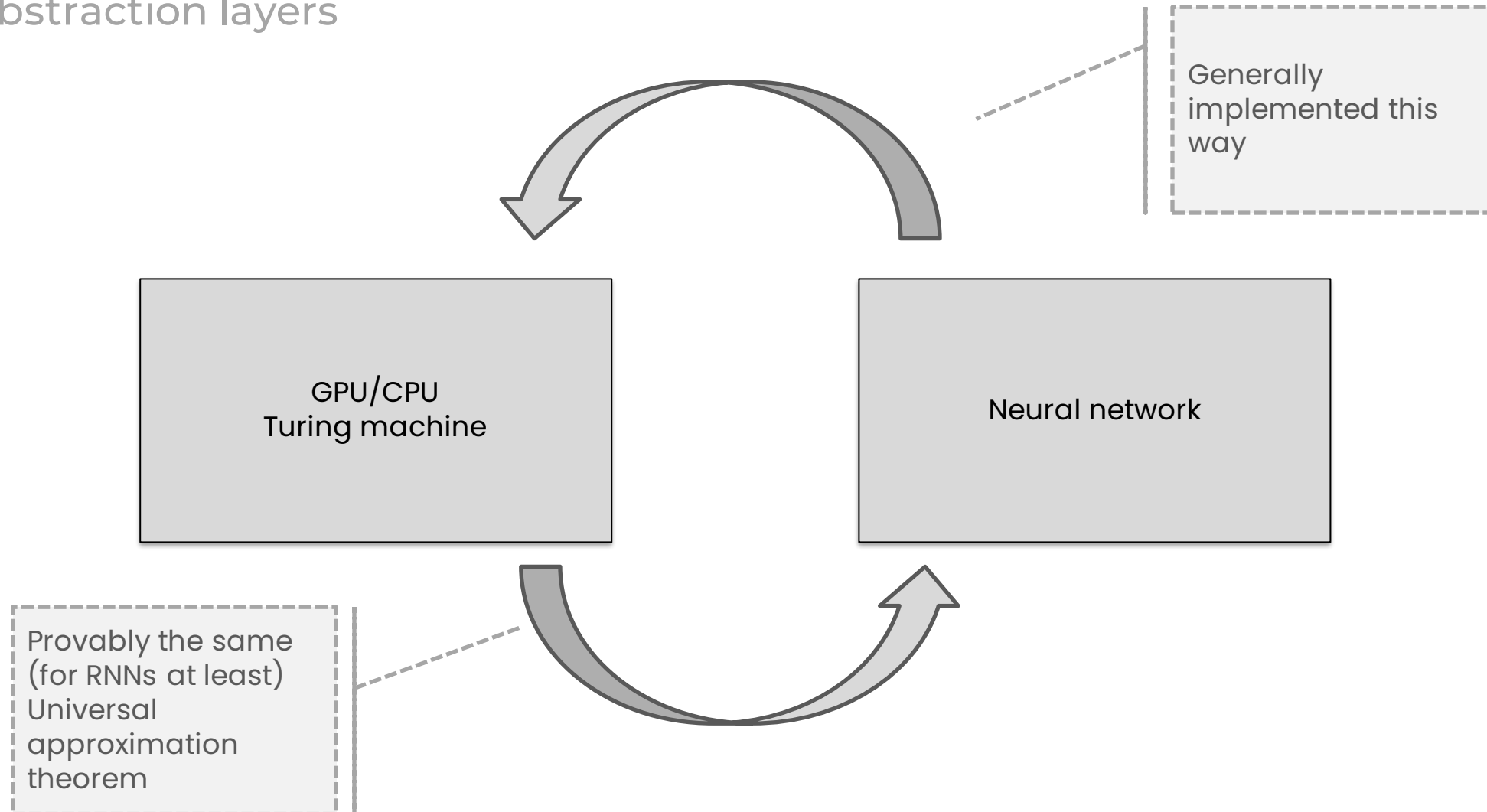
Neural foundation

Most modern AI



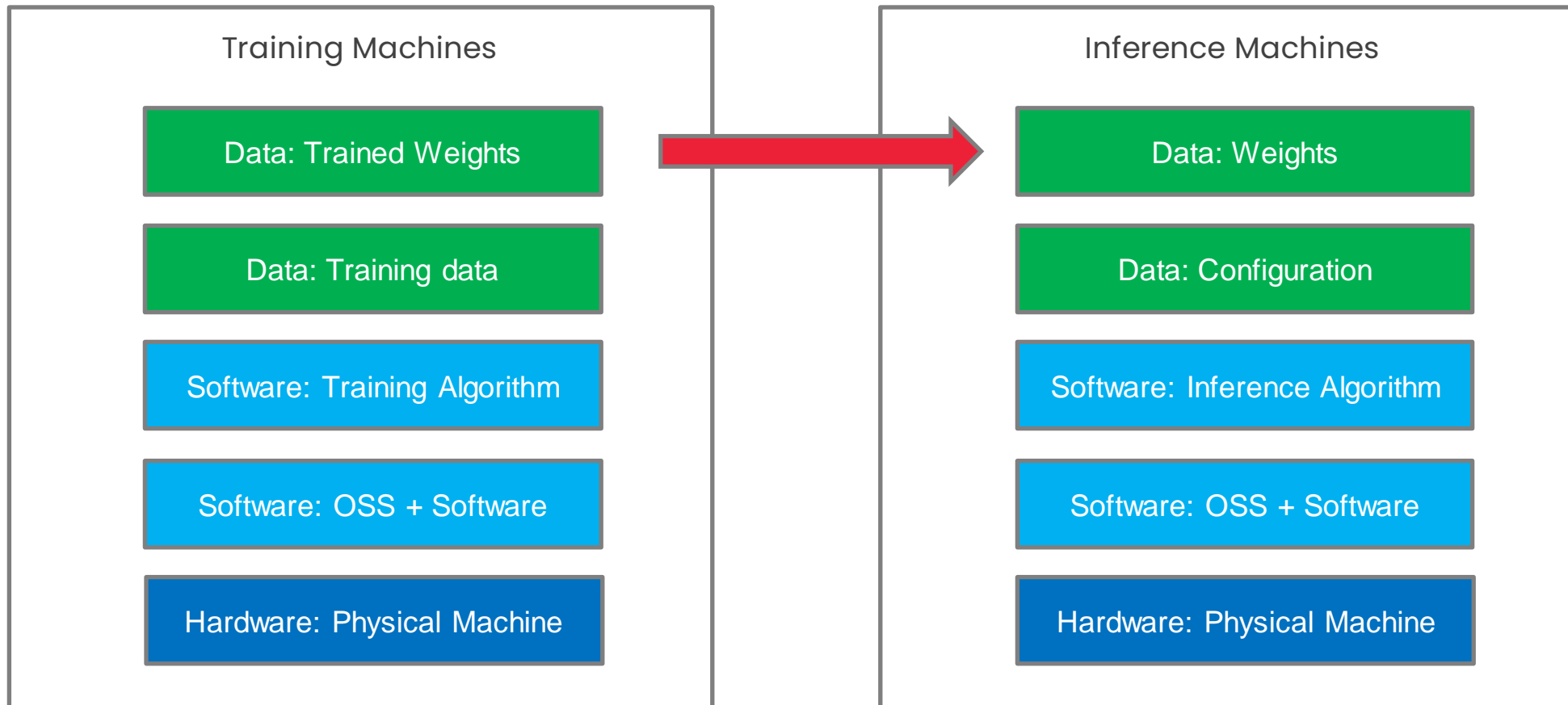
Emulation – Substitution

Abstraction layers



Simplified NN lifecycle

Most modern computing



Dimensionality

Which bits do we need to trust

Training data: 13 Trillion Tokens

OS (Linux kernel) 27 million lines of code

Training weights: 1.76 Trillion Parameters

GPH hardware (Nvidia H100): 80 Billion transistors

- All of it!!!
- Emergent/complex behaviours.
- Across the lifecycle (training v inference)
- And dynamic
- A single bit/weight can change behaviour.

Trust

Trust

Is Subjective

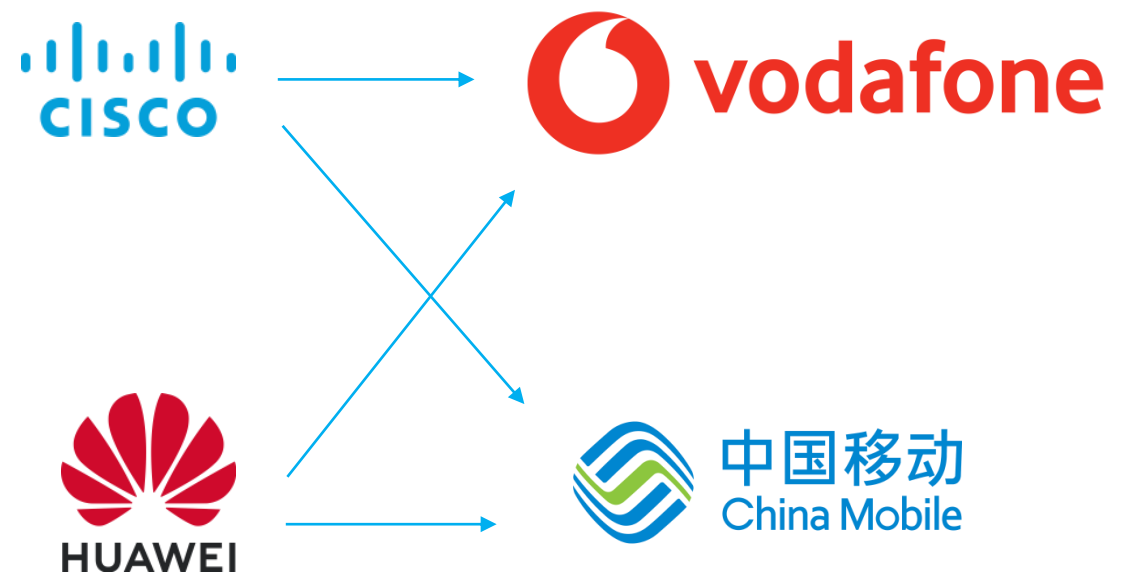
Not linguistically intransitive

Trust (A)

Trust (A,B)

Trust (A,B,t)

Trust (A,B,t,X)



The Government has asked Vodafone and other mobile operators to remove Huawei kit from our 5G networks by 2027, and to stop buying any new Huawei 5G kit from the end of this year. So what does this mean for customers?

Trust

Is Transitive
(mathematically)

$\text{Trust}(A,B), \text{Trust}(B,C) \rightarrow \text{Trust}(A,C)$

Training
data

Training
machines

Algorithm

Validation
Data

Deployment
machine



TAIBOM

Trusted AI Bill of Materials

Scope

Addressing the Challenge

- How do we delineate the AI system we are measuring?
- How do we define and ideally measure trustworthiness?

Scope

Addressing the Challenge

TAIBOM (Trusted AI Bill of Materials) directly addresses this challenge by providing

- A method of defining the immutable properties of a complete but complex AI system; defining a stable AI system
- A method of making and evaluating both objective and subjective claims about the trustworthy attributes of a stable AI system and its constituent parts.

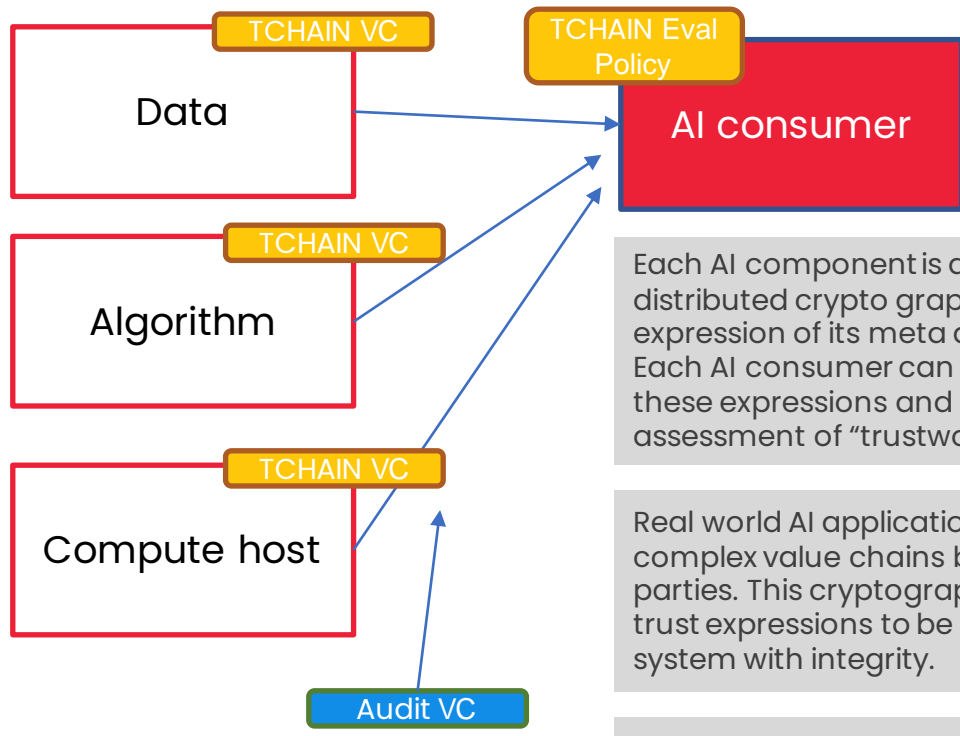
TAIBOM will produce two essential deliverables

- TAIBOM standard: a formal industry standard for making trustworthy assertions over AI building blocks and complete, stable AI systems. (collaborative)
- TAIBOM software suite: a suite of commercial tools for making and publishing compliant AI building blocks on a fully distributed basis (proprietary)

These will enable the AI ecosystem to buy and sell with confidence, and provide a potential route for building trustworthy AI marketplaces.

TAIBOM

Trustworthiness evaluation polices are subjective. An algorithm suitable for one application is not for another. A validation data set might be representative for one application – not for another. E.g. (real world example) genomic model trained on US health records, exhibits major anomalies (bias) when applied to Taiwanese data. The system works is trustworthy in US, not Taiwan

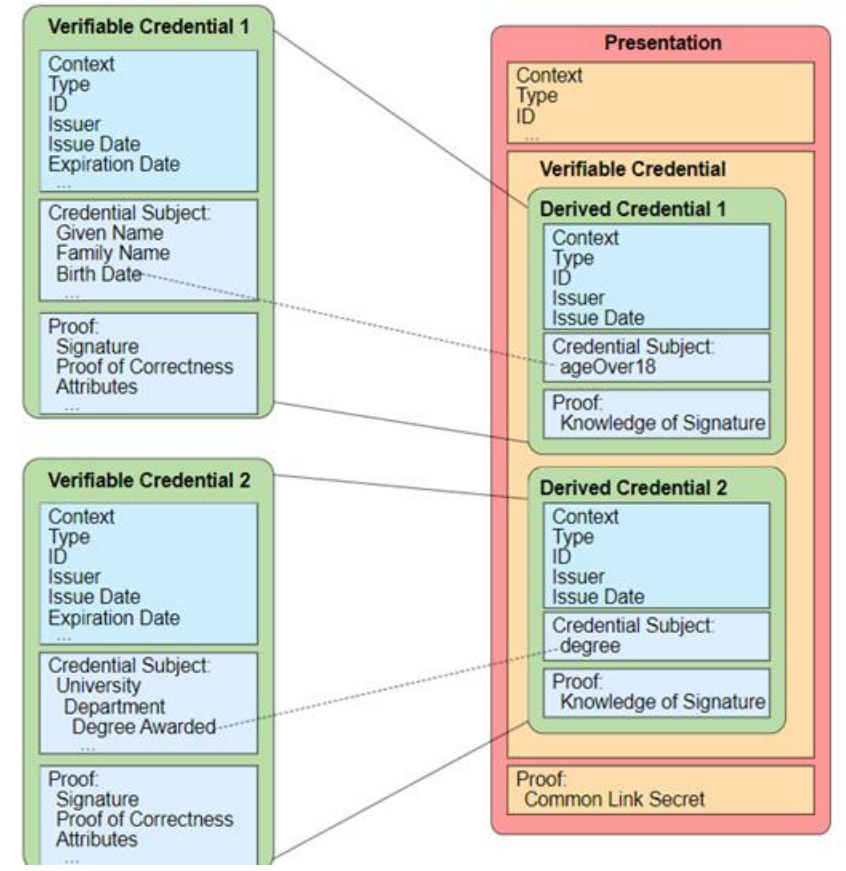



Each AI component is accompanied with a distributed crypto graphically secure expression of its meta data and provenance. Each AI consumer can confidently consume these expressions and come up with its assessment of "trustworthiness"

Real world AI applications, are embodied in complex value chains between multiple parties. This cryptographic method allows the trust expressions to be passed across the system with integrity.

The trustworthiness of individuals element can be audited by third parties. People or AI systems. These auditors may comment on "fairness", "security", "bias" (etc) Any other quality on which trust assessments are made

W3C VCs allows limited disclosure. Parties can exchange "trusted assurances" without over disclosure.



Verifiable Credentials Data Model v1.1 
<https://www.w3.org/TR/vc-data-model/>

Similar work from CISA

CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY



AMERICA'S CYBER DEFENSE AGENCY

Search

Topics ▾ Spotlight Resources & Tools ▾ News & Events ▾ Careers ▾ About ▾

Home / News & Events / News


BLOG

Software Must Be Secure by Design, and Artificial Intelligence Is No Exception

Released: August 18, 2023

By Christine Lai, AI Security Lead and Dr. Jonathan Spring, Senior Technical Advisor

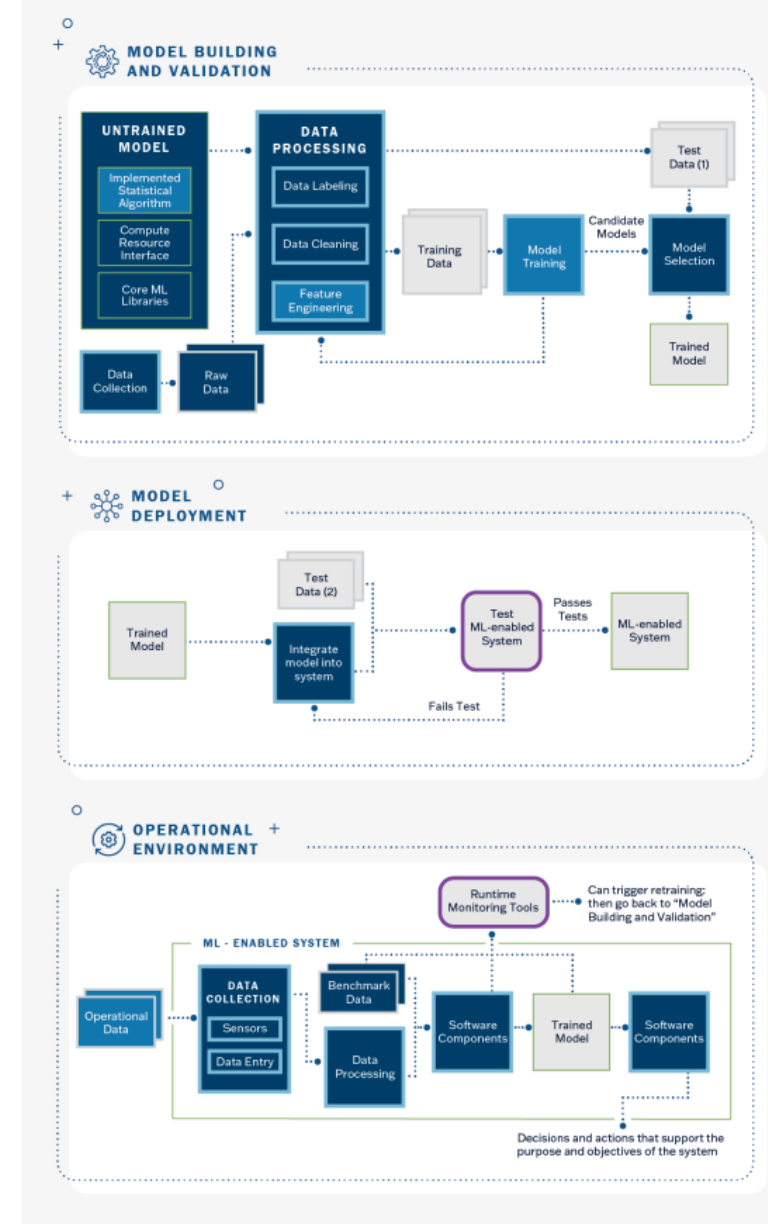
<https://www.cisa.gov/news-events/news/software-must-be-secure-design-and-artificial-intelligence-no-exception>



Allan Friedman, PhD
 Leading CISA's efforts to coordinate SBOM initiatives inside and outside the USG and around the world.

[+ Follow](#)

[View full profile](#)



KEY

- Cybersecurity processes should apply with few adaptations
- Cybersecurity processes require major adaptations to apply
- Some processes apply as-is and some require major adaptations
- Human software development process
- Artifact intended as data
- Software artifact
- Quality testing process

Features

TAIBOM – Basic capabilities



Labelling/Versioning

Every aspect of a complex AI system needs labelling and versioning. (data, code and physical systems). Ideally there should be a method of attesting to the version. There can be various trust models to implement this

Dependencies

A complex AI system has dependencies that need describing to fully understand provenance. TAIBOM will provide an interoperable method of describing these dependencies

Attestation

Any actor (author or third party) can provide descriptors for each component of the system as a whole. (e.g. a training content review, as SBOM validation, a system integrity check, a fairness assessment).

TAIBOM provides both a mechanism of making these attestations, but also a framework for the dynamic and subjective evaluation, of combinations of these attestations.

TAIBOM Use Cases

TAIBOM – How used



AI Store / Distribution

Distribution of systems (paid or internally managed) needs workflow and compliance. If this is distributed, you need a sophisticated signing mechanism. The deep provenance TAIBOM provides enhances this considerably. It also underpin security evaluation and revocation

WAC experience informs this.

TAIBOM – How used



AI system inventory

Any actor using an ensemble of AI systems will benefit from automated process that can create dynamic inventory of working systems. If this system can determine provenance, dependencies and third party attestations, then this process is more effective. There are many use cases for this

- license validation
- security assessment
- data flow validation etc.

TAIBOM – How used



Operational chaining

The result of any real-time AI system will produce a result. This result can be annotated with meta data describing the full provenance, dependencies author and third party attestations relating to the production of this result.

This is useful for real-time analysis and monitoring and essential for complex AI systems of many parts

<https://contentcredentials.org/> addresses similar issues for a very narrow use case and scope

Community building

TAIBOM - Trusted AI Bill of Materials

Register interest: <https://www.techworks.org.uk/ai>

Questions: nick@nqminds.com

Press Release: