Towards Provable Provenance and Privacy-Preserving Queries in Sovereign AI, Data and Identity Architectures

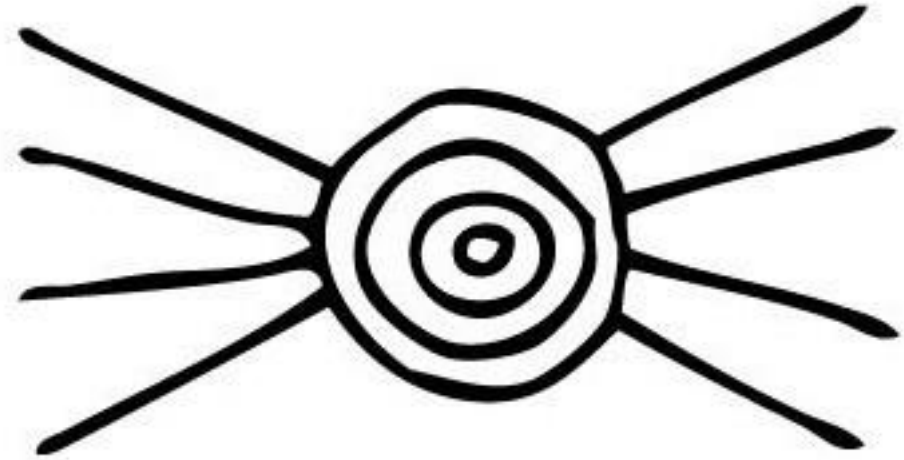
Transfer of Status Report

Jesse Macleay Wright
Jesus College, Department of Computer Science

Supervised by
Professor Sir Nigel Shadbolt FRS FREng FBCS, and
Dr Jun Zhao

Report for the degree of
DPhil in Computer Science

Hilary Term 2025



"Meeting Place"

Byron Bay (Cavenbah) has always been an important meeting place for the Arakwal, neighbouring clans and people of the Bundjalung nation. Our people, together with neighbouring tribes and clans, make up part of the wider Bundjalung Nation. This nation extends to Grafton and the mighty Clarence River in the south, up north past Tweed River to the Nerang River in southern Queensland, and out west towards the Great Dividing Range.

- *The Bundjalung of Byron Bay Aboriginal Corporation (Arakwal)*[1]

Contemporary Aboriginal society is changing at an incredible pace. Its amalgamation with western technologies and its yielding to social and cultural pressures create an immense threat to Indigenous relationships with the world ecological order. Aboriginal people are in the throes of a political struggle to have their land and rights restored. As modern society intrudes into Indigenous minds, introducing different values and directions, Aborigines can be expected to lose sight of certain principles in the process.

- *Jim Everett, Tasmanian Aboriginal leader*[2]

[1] Arakwal People of Byron Bay. *About Us*. Retrieved April 14, 2025, from http://arakwal.com.au/
[2] Grieves, V. (2009). *Aboriginal spirituality: Aboriginal philosophy – The basis of Aboriginal social and emotional wellbeing* (Discussion Paper No. 9). Cooperative Research Centre for Aboriginal Health. https://www.lowitja.org.au/wp-content/uploads/2023/05/DP9-Aboriginal-Spirituality.pdf

Table of Contents

1 Abstract

This project will advance the state of the art in the discipline of query on the Semantic Web, developing novel algorithms, abstractions and architectures to enable minimized data disclosure whilst maintaining verifiable provenance trails.

This work develops a standardised declarative query language (*data sublanguage*) for accessing graph database(s) alongside zero-knowledge verifiable provenance statements – including of data sourcing, integrity and derivations. Supported queries include "Is Jesse over 21 according to facts issued by EU or UK governments" – the verifiable response reveals only the answer: "yes." This query language is first implemented in query engine(s) which evaluate queries over a locally indexed graph database. Support is then added for queries over the union of data residing across independent and potentially malicious graph-databases; by developing algorithms and architectures which minimize data disclosure between sources when planning and executing queries. Finally, the architectures and algorithms developed for decentralised query execution are generalised to support dynamic authentication and authorisation between databases.

This work builds upon recent advancements in Privacy Enhancing Technologies (PETs), particularly Zero Knowledge Proof (ZKP), to support selectively proving properties of provenance trails; and Secure Multi-Party Computation (SMPC) to support data minimisation in planning and executing distributed queries.

This work has immediate applications, particularly for improving the privacy of EU and UK citizens using Digital Verifiable Credentials – which are being rolled out under the European Digital Identity (EUDI) and Digital Verification Schemes (DVS) respectively.

The query language preserves privacy, enabling clients – including scripts, applications and AI Agents – to precisely describe the information and verifiable provenance they require from a server. The implementation for federated graph-databases means SMPC and authorisation concerns are "hidden" from clients interfacing with numerous databases.

2 List of Abbreviations

AI	Artificial Intelligence
BBS	Boneh, Boyen and Shacham signature scheme
DSL	Domain Specific Language
C2PA	Coalition for Content Provenance and Authenticity
CBOR	Concise Binary Object Representation
CHAPI	Credentials Handling API
DCQL	Digital Credentials Query Language
DIATF	Digital Identity and Attributes Framework
DIF	Decentralised Identity Foundation
DUAB	Data (Use and Access) Bill
DVS	Digital Verification Scheme
E2EE	End to End Encryption
EAA	Electronic Attribute Attestation
eID	Electronic Identity
eIDAS	electronic Identification, Authentication, and Trust Services
EUDI	European Digital Identity
FedCM	Federated Credential Management
FIDO	Fast IDentity Online
FOSDEM	Free and Open source Software Developers' European Meeting
HCI	Human Computer Interaction
IEEE	Institute of Electrical and Electronics Engineers
IETF	The Internet Engineering Task Force
ISO	International Organization for Standardization
JSON-LD	Javascript Object Notation for Linking Data
LCM	Large Concept Model
LLM	Large Language Model
mDL	Mobile Drivers License
MPC	Multi-Party Computation
NLS	oN-Line System
OAuth	Open Authorization
OIDC	OpenID Connect
OIDC4VP	OpenID Connect for Verfiable Presentations
ODRL	Open Digital Rights Language
OWA	Open World Assumption
OWF	Open Wallet Foundation
PID	Personal Identifiable Data
QEAA	Qualified Electronic Attribute Attestation
RDF	Resource Description Framework
SD	Selective Disclosure
SD-JWT	Selective Disclosure for JSON Web Tokens
SDI	Sovereign Digital Infrastructure
SMPC	Secure Multi-Party Computation
SOLID	SOcial LInked Data
SPARQL	SPARQL Protocol and RDF Query Language
SSI	Self-Sovereign Identity

UMA	User Managed Access
VSD/VCD	Value Sensitive Design / Value Centric Design
W3C	World Wide Web Consortium
ZKP	Zero Knowledge Proof
zkSNARK	Zero-Knowledge Succinct Non-Interactive Argument of Knowledge

3 Chapter 1 – Thesis proposal

"We shape our buildings and afterwards our buildings shape us."
– Winston Churchill[3]

3.1 Research Topic – Problem Statement and Motivation

Human-Computer Interaction (HCI) emerged in the early 1980's to *empower* human capabilities and provide better means of working, communicating and thinking. Early HCI researchers including J.C.R. Licklider imagined a positive symbiosis where computers "augment human intellect by freeing it from mundane tasks" (1).

Research and industry developments have realised many parts of the early technical vision. Reflecting on Douglas Engelbart 1968 "Mother of All Demos" (2), the Web, now with 5.5 billion users Web (3) has realised the NLS (oN-Line System) computer collaboration system (4), and the early Apple Macintosh – with a fraction modern device capabilities – realised personal computing ideas such as windows, graphics, word processing and the mouse.

Yet, many contemporary systems and platforms are abjectly anti-human or misaligned with stakeholder values. Examples include: the loss of *truth* and *transparency* (5) on, and growing addictiveness of, social media (6); the loss of *privacy* in the UK as Apple was forced to disable end to end encryption (E2EE) (7); and threatened loss of *ownership* and *IP* if the UK Government proceeds in enabling AI companies to train on copyrighted data (8).

Value Sensitive Design (VSD) (9) was developed as a theoretically grounded HCI methodology that accounts for values and interests of stakeholder groups in the design process. VSD was originally developed for information system design (10), and considers the design of *systems architecture* in addition to *user interfaces* – given that choices in systems architecture reflect and enforce social values, practises and power dynamics. As Lawrence Lessig states: *"code is law"* (11).

In this paper, the term Sovereign Digital Architectures (SDA) (12) refers to the set of software architectures that support individual or group ownership of *identity*, *data*, or *compute.* These include Self Sovereign Architectures include Self-Sovereign Identity (SSI) (13) solutions which enable proof of identity without reliance on 3rd party identity providers; personal data stores – including Solid Pods (14) and some Data Wallets which enable ownership of personal data, and AI agents such as kwaai.ai which can be deployed locally and is designed to serve the interests of the user.

Sovereign Digital Architectures (SDA) (12) and Privacy Enhancing Technologies (PET) (15) have been developed to support a range of values; including *effectiveness, ownership, autonomy*, *privacy, integrity* and *transparency*, and thus can be deployed to support values

[3] Churchill, W. (1943, October 28). *House of Commons Rebuilding* [Speech transcript]. Hansard, UK Parliament. https://hansard.parliament.uk/commons/1943-10-28/debates/4388c736-7e25-4a7e-92d8-eccb751c4f56/HouseOfCommonsRebuilding

elicited as requirements in VSD. Often the functional requirements for different values cannot be simultaneously implemented – for instance complete transparency requires giving up privacy; this is termed "value conflict." This thesis reduces value conflict by enabling zero knowledge (*privacy*) proof (*integrity*) of provenance properties (*transparency*) on-demand in queries (*effectiveness, autonomy*) over a range of architectures including Sovereign Data and Identity Architectures (*ownership*).

Earlier work on this thesis also supported *effectiveness, ownership, control and power* and *autonomy* in work on data usage controls (16) alongside *empowerment* and *autonomy* through work done on topics such as personal AI, semi-autonomous web agents (17), and communication in multi-agent systems (18).

3.2 Research Questions, Goals, and Contributions

"*Computer Science is the art of abstraction.*"

– David Hyland-Wood

3.2.1 Research Question(s)

3.2.1.1 High Level Research Questions

This research intends to reduce the value conflict between *effectiveness, ownership, autonomy*, *privacy, integrity* and *transparency* in Sovereign Digital Architectures – particularly, **identity** and **data** Architectures which support Sovereign AI infrastructure. The aim is to develop standardised declarative query language (*data sublanguage*) supporting zero-knowledge verifiable provenance statements – which shall herein be referred to as a *verifiable data sublanguage*. Query-engine implementations, which provide an on-demand view of public and private knowledge on the Web will be developed. Importantly, the *verifiable data sublanguage* is to provide a layer of abstraction for interfacing with a diverse range of *identity* and *data* architecture configurations at global scale. This leads to the following research question:

Which logic profiles of verifiable data sublanguages afford computationally efficient query-engine implementations against given configurations of data and identity infrastructure?

This is broken down into the following research questions:

- **Research Question 1**: Which logical profiles can be supported by a query engine implementing a verifiable data sublanguage for a single graph database.
- **Research Question 2**: When implementing the verifiable data sublanguage of **RQ1** across a distributed set of graph databases:
    - **RQ2A:** what is the minimal set of information that can be shared (disclosed) between the graph databases in computing the result, and
    - **RQ2B:** what logical profiles of the verifiable data sublanguage from **RQ1** can be efficiently supported for given configurations of graph databases.
- **Research Question 3**: To what extent can contemporary procedurally defined authentication and authorisation protocols – such as OIDC – be replaced by:

    - services declaratively defining **what** needs to be proven to them (e.g. to provide access to a resource, or perform an operation), and
    - extending the query planning process designed for **RQ2** to account for these requirements during query evaluation.

As outlined in the abstract, **RQ1** is expected to use Zero Knowledge Proof (ZKP) techniques, **RQ2** is expected to develop federated query planning and execution engines that apply Secure Multi-Party Computation (SMPC), and in **RQ3** is expected to extend the query planning architectures from **RQ2** to dynamically support authentication and authorisation across heterogenous identity infrastructure.

3.2.1.2 Choice of Query Interface

To answer these question(s) either a new query language must be designed, or an existing one must be selected and extended. To be amenable to our research questions, this query language **must** satisfy the following properties:

- The language is **declarative** with **clear execution semantics** (19) such that the expected set of results is known, but there is room to evaluate the result in different ways – as is the case with most query languages such as SQL (20), SPARQL (21) and Cypher (22). Crucially, the semantics of the query must not be dependent on the endpoint it is executed against, as is the case with query languages such as GraphQL (23).
- The query language compatible with the Open World Assumption (OWA) (24) so that it is compatible with partial datasets and does not break when participants in the ecosystem have unexpected data or schemas.
- The query language uses globally unique identifiers (25) to support distributed data sources out of the box.
- It is possible to capture provenance, including issuer signatures, directly within the database and query result.

The SPARQL Protocol and RDF Query Language (SPARQL) supports these requirements. Specifically, the SPARQL 1.2 Query Language (26) is chosen for the initial work to answer these research questions as RDF 1.2 and SPARQL 1.2 are specifically designed to discuss reified terms (that is, statements about statements) – and thus is well suited to support describing verifiable provenance statements, including those in zero-knowledge.

Note that query interface design is still required to define the *built-in* properties that are to be used for requesting details of the verifiable provenance statements – in addition to ontology design for describing these statements.

3.2.1.3 Research Questions grounded in SPARQL 1.2

Having established SPARQL 1.2 will be used, the research question may be re-phrased as follows:

Which logical fragments of SPARQL 1.2 with novel built-ins afford computationally efficient implementations against given configurations of data and identity infrastructure?

3.2.2 Minimal use-case

Before discussing our aims, objectives, and contributions – a minimal use-case is provided to exemplify scenarios that are enabled upon answering the above research questions. Given the widespread nature of Data Wallets – as discussed in the Verifiable Credentials section of this paper – the use case is as follows:

> *A group of four friends are applying for a rental. They need to prove to their landlord that their cumulative salary is over £100k p.a. to rent a property.*
>
> *They each have a Data Wallet* (27)*, containing a Verifiable Credential* (28)*, which include their last years' worth of payslips. Three of the friends have Data Wallets that are online Solid Pods* (14)*, one friend has a Data Wallet which is an application on their phone that stores all credentials locally.*

With the state of Data Wallets today, the landlord must fetch all the friends' payslip credentials; likely using two separate applications, or authentication procedures to fetch credentials from the Solid Pods (14) (after authenticating with FedCM (29)), and phones (using a self-sovereign OIDC4VP (30) flow), directly. The landlord's application can then use this data to verify the minimum aggregate salaries.

Research Question 1 enables minimization the data sent from each friend to the landlord. This allows the landlords application ask the applicants Data Wallet "is this person's annual salary over £25k," and have a proven answer provided to the landlords application on demand.

Research Question 2 enables further this privacy preservation by allowing the landlord to ask of the three Solid Pods "is the cumulative salary across these wallets greater than £75k" whilst a separate application is still used to ask the phone user "is this person's annual salary over £25k."

Research Question 3 aims to – amongst other things – help bridge the divides of disparate authentication flows, so that a single application can *easily* be built which asks all four data wallets "is the cumulative salary across these wallets greater than £100k."

The flow diagrams of Figure 1 in the Appendix show the state of using Digital Wallets today, and how these change after each research topic. Work towards a syntax for the query interface can be found on GitHub[4].

3.2.3 Aims and Objectives

This research aims to reconcile the trade-offs between *autonomy*, *privacy*, *integrity*, and *transparency* when querying over Sovereign Data and Identity Infrastructures. Specifically, the objectives are to:

[4] https://github.com/jeswr/queryable-credentials?tab=readme-ov-file#initial-design-thoughts-for-a-queryable-api

- Understand what properties of a result provenance trails can be proven in zero knowledge with reasonable computational efficiency.
- Understand what is the minimal amount of knowledge that can be disclosed between computing nodes in producing a distributed query result with provenance trails.
- Understand the extent to which flows related to identity, authentication, and authorisation can be made to emerge from query planning processes.
- Develop standards for requesting and describing zero knowledge provenance trails in SPARQL 1.2.
- Develop algorithms and architectures to implement this query interface for a:
    - centralised service (codebase output),
    - decentralised data stores (codebase output), and
    - decentralised data stores requiring authentication and authorisation for data-access
- Integrate these architectures into a range of systems, and demonstrate application in real-world use-cases including:
    - Infrastructure
        - Semi-autonomous neurosymbolic AI agents (17), and
        - Aggregation and search services for decentralised data ecosystems such as Solid
        - To "derive credentials" within holder or orchestration service under the UK's Digital Identity and Attributes Framework (DIATF) for the Digital Verification Scheme (DVS).
    - Use cases
        - Aggregate health datasets with provenance for data-trusts,
        - Collect trusted statistics of a population without compromising the privacy of population members e.g. studying the sexual health records of a population
        - Proving a population meets a requirement e.g. to determine the average household income of a school cohort to establish benefits eligibility
        - Have AI agents apply for home loans on behalf of users, supplying required proof of credit history and employment in the process

3.2.4 Expected Contributions

This section outlines the impact and applications of answering each research question. The Timeline section provides a concrete overview of the technical contributions of this work, alongside plans for publication, standardisation and adoption.

3.2.4.1 Research Question 1

Verifiable Credentials (VCs) (31) have applications across personal credential management (e.g. Digital Driver's Licenses) (32), supply chain traceability (e.g. the UN Transparency Protocol) (33) and authorisation (e.g. proof of age online). VCs now have widespread usage, in part because the European Digital Identity Scheme (eIADS) and UK Digital Verification Scheme (DVS) mandate support for these credentials in various sectors.

W3C VCs (31) are RDF (34) native, requiring a JSON-LD encoding (35). Thus, a collection of VCs may be seen as an RDF Knowledge Graph with signed provenance on triples.

Zero Knowledge Proof (ZKP) (36) systems are becoming mature – with Zero Knowledge Virtual Machines (ZKVMs) (37) now able to prove correct execution of arbitrary RISC-V instruction sets (38). However, in VC standards ZKP usage is limited to Selective Disclosure (SD) to prove that a subset of facts is true.

In answering this RQ, the authors expect to leverage recent ZKP advances and develop a SPARQL 1.2 interface with built-ins for proving provenance properties in zero-knowledge. As VCs are RDF native, this interface can be immediately used by Digital Wallets to support asking, "Does the owner of this Digital Wallet earn over £25k p.a. based on bank statements issued by trusted UK banks." With current SD mechanisms the query would be more invasive "show all signed transactions adding money to the users bank account." Further explanation of this feature gap are given in Wright's FOSDEM presentation (39), and further use cases can be found on GitHub[5].

LLMs (40) are commonly grounded using Retrieval Augmented Generation over data from Graph Databases (GraphRAG) (41) – often with RDF-based Knowledge Graphs (42). Typically, systems use a "question-answering" process wherein LLMs are used to generate SPARQL queries to fetch data that is to be included in a subsequent prompt (43). In answering this RQ, GraphRAG may be performed against remote data stores with guarantees that the data has certain provenance features that allow it to be "trusted" - such as being derived from facts issued by the UK Government and NHS Services.

There are also applications for near- and long-term semi-autonomous neurosymbolic AI agents (17). The near-term is to perform "trusted" GraphRAG over remote datasets within agentic AI architectures such as Charlie (44). Looking to the future, the authors hypothesise that epistemic memory modules (45) will emerge within LLM (46), LCM (47) or future Deep Learning architectures; and expect this work to play a role in supporting the accumulation of trusted knowledge for this memory.

3.2.4.2 Research Question 2

In answering this RQ, the authors expect to leverage Secure Multi-Party Computation (SMPC) in addition to ZKP to extend support for the SPARQL 1.2 interface developed in RQ1 to support querying over data distributed across two or more graph databases – whilst minimising the data that is disclosed between graph databases.

Extending the discussion of RQ1, the developments from this RQ can be applied to Digital Wallets to support asking, "Do Alice, Bob and Charlie jointly earn over £75k p.a. based on bank statements issued by trusted UK banks." Similarly, this work supports "trusted" GraphRAG over distributed datasets and in turn semi-autonomous neurosymbolic AI agents (17).

[5] https://github.com/jeswr/queryable-credentials

Further expect to apply this work to the following use cases:

- **Data Trusts** (48): Allowing the ergonomic collection of aggregate datasets from distributed data-stores whilst protecting user-privacy. Examples include the:
    - **Clinical Data Federation**: Our group is working on a project with the University of Oxfords Paediatrics Department; the goal is to enable the collection of aggregate data from physio studies – with the patient data stored in data-stores across a range of regional jurisdictions. In the current phase of the project, the developer is responsible for defining the Multi-Party Computation algorithm that is used on a per-query basis; which extends the existing work done by our group on Libertas (49). The work outlined in this research topic would enable the developer of the research platform to simply define a SPARQL query defining the aggregate result they want from data in the dataset.
- Derived credentials for groups, including:
    - Proof that a school is eligible for Pupil Premium Funding (50) based on aggregate properties of the student population.
    - Proving demographic discrimination with such as salary discrepancy within a given population.

3.2.4.3 Research Question 3

On the Web, there are a plethora of *identity* infrastructures, as well as *authentication* and *authorisation* flows. These are detailed in the literature review on Contemporary Identity Infrastructure. The author hypothesises that this is because:

- Identity solutions are being built with different use cases and requirements in mind,
- Those developing digital identity solutions have differing priorities over the *values* and in turn functional requirements for which the solution is built, and
- There are numerous ways of implementing protocol flows to satisfy a given set of functional requirements. Individuals and groups can come to rely on specific solutions for numerous reasons: including developers becoming attached to "their" solution, lack of knowledge of existing solutions, political challenges of standards harmonisation, and the technical burden of migrating to use new specifications.

This research question investigates the extent to which fixed *authentication* and *authorisation* flows such as OIDC can be eliminated and instead *negotiated* or *designed* during a query planning process (17) (18). This mitigates the need for humans to design *authorisation* and *authentication* flows – thus bypassing the third cause (numerous implementations) of flow proliferation. Moreover, these flows can be designed on-the-fly and tailored to specific configurations of *identity* infrastructures; thus, removing the need for all classes of use-cases to be anticipated by developers in advance of their occurrence.

To exemplify this, consider the use-case of transferring digital credentials; for which there are currently numerous *authorisation* and *transfer*. With the architectures developed for this

Research Question it would be possible for a resource endpoint to declaratively advertise that clients must include the following when attempting to fetch the resource:

- Delegated authority from a user that has been authenticated using:
    - an identity managed by the user using their own key infrastructure, or
    - an identity provider from the list of:
        - Google
        - GitHub
        - Facebook
- And have a signed confirmation from Alice, Bob or Charlie – that the user has authority to access the resource

Whilst User Managed Access (UMA) (51) and OIDC4VP (30) can support these use cases, they are achieved by having developer documentation on the resource server outline the above requirements in natural language – which developers of server clients then need to interpret and encode into the behaviour of the client application.

In answering the research question, architectures will be developed to enable software clients to determine which, and which order of actors (Alice, Bob, Charlie, Google, GitHub, Facebook, resource server) and what information needs to be exchanged within each interaction.

Support for the following use-cases will also be implemented whilst answering this Research Question:

- Dynamic Authentication Flows: Including the resource access use-case described above, and e-readers accessing authenticated documents with institutional authentication[6].
- Delegating Authorisation to AI Agents: As motivated by South et. al. (18).
- Agents performing scheduling (17).

[6] https://github.com/jeswr/queryable-credentials/blob/main/use-cases/e-reader-access.md

3.3 Research Plan

3.3.1 Ethics

This research does not involve studies with human participants and does not require any form of ethics approval.

3.3.2 Timeline

The researcher plans to sequentially address each research question, with outputs according to the below plan. This Gantt chart summarises the timeline; and includes only that work starting from the transfer assessment date. As discussed below, there is flexibility to present some topics listed separately below within the same paper.

		2025												2026												2027					
		J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J
RQ1	ISWC Zero Knowledge Proof SPARQL																														
	W3C Specification Documents																														
	UseNix submission(s) of SPARQL ZKP engine system descriptions																														
	SWJ submission(s) of SPARQL ZKP engine complexity																														
RQ2	UseNix Submission of MPC database (**optional**)																														
	ISWC Submission of MPC + ZKP SPARQL Engine (registry based)																														
	ISWC Submission of MPC + ZKP SPARQL Engine (query planning based)																														
RQ3	Registry for Authentication flows (potential UseNix submission)																														
	Query Planning for Authentication flows (potential UseNix submission)																														
NQ	Thesis write-up																														
	Buffer																														

3.3.2.1 Research Question 1

The researcher intends to publish this work as conference and journal papers to ISWC, Semantic Web Journal, UseNix, and ADBIS. The researcher is currently targeting the first deliverables of this piece of work for ISWC 2025 (Deadline May 13, 2025). The planned timeline and outputs are:

- **May 13, 2025** - A submission to ISWC answering the following Research Questions:
  - Which logical fragments of SPARQL 1.2 with built-ins for zero-knowledge verifiable provenance are implementable, regardless of computational cost?
  - Is it possible to write a zero-knowledge proof specification that is dependent on only one of the following:
    - The SPARQL Query Algebra
    - A SPARQL Proof Language (under development)
    - The Notation3 Query and Reasoning Semantics
    - The Notation3 Proof Language
  - What are the differing approaches that can be taken to implement proof of inference or query correctness?
- **July 2025** - A W3C CG Specification Documents
  - defining a SPARQL (1.2) Proof Language
  - formalising a the existing Notation3 Proof Language
  - describing a proposed set of new standards for the Verifiable Credentials API (52) to support derived disclosure.
- **Ongoing through to end of 2025** A system description paper to UseNix, similar to Tsarkov et. al. (53), for each *novel* ZKP proof engine which implemented. It is expected that a subset of the below matrix of engines will be implemented:
  - Implementing a
    - SPARQL 1.2 system
    - Notation3 system
  - which performs
    - query/inference
    - proof checking
  - within a
    - custom circuit (54)
    - Zero Knowledge Virtual Machine (37)
- **Ongoing through to end of 2025** A submission to the Semantic Web Journal (SWJ) or ADBIS performing a theoretical and experimental complexity analysis of each system implemented – these complexity analyses may also be part of the above UseNix submission(s).
  - The theoretical complexity analyses will be performed in the same manner as the SPARQL complexity analyses have been performed by Perez et. al. (55) for

SPARQL and Horrocks et. al. (56) for rule-based inference profiles such as SROIQ (56).
    - Empirical complexity analyses will extend traditional SPARQL benchmarks such as the Berlin SPARQL Benchmark (57) to add queries requesting verifiable provenance.

Existing work and notes and research work towards this answer are available on GitHub[7].

3.3.2.2 Research Question 2

The roadmap for RQ2 is as follows:

- Develop registry-based methods for determining the SMPC algorithms to use by taking libraries of MPC algorithms[8] and defining those algorithms in standardised declarative formats, including:
    - a clear prescription of the flow – i.e. *who* sends *what* when (including conditional logic), and
    - proofs of security assumptions and guarantees, such as whether the algorithm is susceptible to a 51% attack by malicious parties
- Investigate how query planning can be used to develop such SMPC algorithm descriptions on-the-fly

The following extensions are proposed for our work on research question 2:

- **Policy Aware**: Use ODRL (58) to describe the usage-controls on datasets coming from disparate sources and compute the usage controls that apply to the derived dataset. The researcher has confidence that they would be able to successfully and efficiently complete such work as they have experience using ODRL across industry and academia – as demonstrated in Wright et al. (16).
- **Croissant Dataset Generation**: As an extension to the data trust use-case, support the ability to generate datasets annotated with Croissant Metadata (59; 1) to support use in machine-learning applications.

The author plans the following output timeline for this research question:

- **January - March 2026**: Work on a UseNix submission which provides (*may be skipped if the prior work identified in the literature review effectively satisfies the requirements*):
    - A formal language for describing SMPC algorithms,
    - A formal language for describing the assumptions of SMPC algorithms – such as number of honest parties,
    - A database of formally described MPC algorithms – taken from existing papers and code.
- **March 2026 - May 2026**: Work on an ISWC submission which describes:

[7] https://github.com/jeswr/queryable-credentials
[8] https://github.com/rdragos/awesome-mpc

    - A functional federated query engine architecture that can generate plan SMPC approach to evaluate a range of SPARQL 1.2 queries with verifiable provenance built-ins, by referring to the SMPC algorithm database. This interface will include support for proof of query correctness, and properties of the provenance of the result using the interface developed as part of research question 1.
- **May 2026 - July 2026** - A submission to the Semantic Web Journal which:
    - Generates SMPC algorithms on the fly based on the SPARQL 1,2 query and security assumptions about other nodes in the network.

The contents of the last two papers are expected to:

- Describe the architecture,
- Perform a theoretical complexity analysis of the system when executing different query types – in the same manner as SPARQL complexity analyses have been performed by Perez et. al. (55) for SPARQL and Horrocks et. al. (56) for rule-based inference profiles such as SROIQ (56).
- Empirically evaluates the complexity of the system using traditional SPARQL benchmarks such as the Berlin SPARQL Benchmark (57).

These last two papers may be presented in a single journal submission

3.3.2.3 Research Question 3

The roadmap for RQ3 is as follows:

- **July 2026 to September, 2026** Develop registry-based method for defining authorisation protocol flows in standardised declarative formats, including:
    - a clear prescription of the flow – i.e. *who* sends *what* when (including any conditional logic), and
    - proofs of security assumptions / guarantees, answering the questions that would normally be asked in a W3C Security Review (60)
- **October, 2026 to December 2026** Investigate whether the query planning developed in Research Question 2 can be extended to develop such protocol flows on-the-fly

The researcher appreciates that these roadmap items lack many details – that is because the researcher has a clear view of the need, but not what the implementation details will look like for this research question. Consequently, there will be a development approach of starting with common flows of existing authentication protocols – primarily OIDC – and:

- For the registry method:
    - create a formal description of those flows,
    - create a formal description of the security assumptions and guarantees of those flows, and
    - generate a verifiable proof that the security assumptions and guarantees follow from those flows – ideally, reaching a point where the proof and

security assumptions can be automatically generated from the flow description
    - create a formal description of
        - what client(s) wants(s) to achieve
        - what a clients security requirements are for the flow
    - work to create a matching algorithm that can search the registry to identify flows that satisfy the client(s) requirements
- For the negotiation and flow-generation method:
    - Investigate if the formal descriptions of client requirements can be used to generate flows, rather than matching against existing flows using a matching algorithm,

It is expected that RQ3 is addressed by extending the algorithms and architectures developed in RQ1 and RQ2, in particular:

- generalising the MPC registry and query planning methods from RQ2 to implement the registry and protocol planning for this research question, and
- using the zero-knowledge proof SPARQL engine developed for RQ1 in many of the flows to support proving that agents satisfy properties such as "the user being over 21." Observe that Verifiable Credentials are being used in the current User Managed Access (UMA) (51) flows, in order to perform such attribute based authorisation in many contemporary authentication systems.

The roadmap for dissemination is as follows:

- **July 2026 - September 2026** - A submission to UseNix describing the registry-based method.
- **September 2026 - December 2026** - A journal submission describing the negotiation and query planning architecture for generating authorisation flows, including any novel algorithmic contributions.

These two papers may be folded into a single submission.

3.4 Collaborations

There are numerous collaborations that the researcher has in place to:

- Support the promotion and adoption of current work,
- Collaborate with on core research topics proposed, and
- To support on research topics outside of the authors core thesis

3.4.1 Verifiable Credentials Collaborations

The researcher is working heavily on numerous applications relating to digital credentials. Particularly, the researcher is leading work on the Solid Project (14) and Solid Pods are applicable as holder services for verifiable credentials. Through this work, and presentations

given[9] on current academic work on the topic the researcher is in contact with key players in the space – including the Linux Foundation Decentralised Trust[10], the Linux Open Wallet Foundation[11] and Mattr Global[12] who have led spec development work on a number of the Verifiable Credential standards. The author also has direct contact with; the core development team of Risc0 zero knowledge virtual machine with which some of this work has already been performed. The researcher is also in contact with the CTO of the Decentralised Identity Foundation who has a PhD from the Stanford Cryptography Group – supervised by Dan Boneh who co-developed the field of Zero Knowledge Cryptography.

3.4.2 LLM Communication

The researcher is collaborating on several topics around communication between LLM-based agents. In particular, the author has co-authored the paper A Scalable Communication Protocol for Networks of Large Language Models (18) and is also co-chairing an informal working group on lightweight standards for LLM-driven web agents[13].

3.4.3 Agentic AI operating over Personal Data Store

Off the back of a Dagsthul paper (61), the author is collaborating on a stream of work developing Computer Using AI agents that operate over Personal Knowledge Graphs.

3.5 Topics covered to date

The work on this DPhil began with a focus on *trustworthy* and *accountable* Personal neurosymbolic AI agents – which operate at Web scale and truly in the best interest of individuals. This can be seen in the following work that I have developed or contributed to:

1. AI-agents in Customer Experience for Vulnerable Consumers (Journal Paper – under review at Journal of Service Management)[14]
2. **Me want cookie**! Towards automated and transparent data governance on the Web (16) (NXDG Workshop Paper)
3. A scalable communication protocol for networks of large language models (18) (Full-length paper)
4. Towards Computer-Using Personal Agents (61) (Short Vision Paper)
5. **Here's Charlie**! Realising the Semantic Web vision of Agents in the age of LLMs (17) (ISWC Demo Paper)
6. EYE JS: A client-side reasoning engine supporting Notation3, RDF Surfaces and RDF Lingua (62) (ISWC Poster Paper)
7. Semi-Autonomous Agents at Web Scale (Doctoral Consortium Paper)[15]

[9] https://fosdem.org/2025/schedule/event/fosdem-2025-5970-are-current-standards-enough-towards-verifiable-credentials-with-expressive-zero-knowledge-query/
[10] https://www.lfdecentralizedtrust.org
[11] https://openwallet.foundation
[12] https://mattr.global
[13] https://las-wg.org
[14] https://drive.google.com/file/d/1mW-4Fw3wSHeCssTzpnv9XE6_DASsvOZL/view
[15] https://jeswr.solidcommunity.net/public/iswc_doctoral_consortium.pdf

The narrower focus of this transfer report remains in support of the vision for *trustworthy* and *accountable* Personal neurosymbolic AI agents; by providing an abstraction with which agents can fetch the "trusted" data they need.

3.6 Immediate next topics

Should the researcher have time remaining in their DPhil upon completing the above areas of research, the immediate topics the researcher would like to explore next are extracting ontological models from LLMs via mind-mapping[16], entity relationship and epistemic memory modules for LCMs[17], designing hybrid vector-graph databases[18], and suitable representations for agentic communication[19]. To further justify these "desirable" work items, observe that LLMs today do not have an explicit understanding or epistemology (63). This needs to change if they are to evolve from being effective "bullshit engines" (64) to become oracles providing epistemic insight as I envision in this article on trusted conversational interfaces[20] – as these topics are. These immediate next topics are towards a future with *epistemically accurate* and *transparent* Deep Learning Models – which may or may not be an evolution of existing Large Language Model (LLM) or Large Concept Model (LCM) architectures.

The work items that the researcher commits to in **RQ1, RQ2**, and **RQ3** are a necessary pre-requisite to support these architectures in effectively operating as part of an *epistemically accurate* and *transparent* real-time Global Information System.

[16] https://blog.jeswr.org/2025/02/17/research-topics#extractingontologicalmodelsfromllms
[17] https://blog.jeswr.org/2025/02/17/research-topics#symbolicconceptualmemorymodulesfordeeplearningmodels
[18] https://blog.jeswr.org/2025/02/17/research-topics#hybridkgvectordatabasearchitectures
[19] https://blog.jeswr.org/2025/02/17/research-topics#suitablerepresentations
[20] https://blog.jeswr.org/2025/03/22/conversational-interface-trusted-data

4 Chapter 2 – Literature Review

4.1 RDF, SPARQL and the Semantic Web

The Semantic Web (4) is a technology stack developed by the World Wide Web Consortium (W3C) to enable interoperability between systems. At its core is the concept of Linked Data, which is the "collection of interrelated [and machine readable] datasets on the Web." Linked Data is achieved by publishing and exchanging data in the standardised Resource Description Framework (RDF) (34).

The atom of a piece of data in RDF is called a triple. A triple relates a resource (a *subject*) to another resource or value (an *object*) through a predicate (a *verb*). A knowledge graph is a set of triples that link and describe such resources and entities, and the triple can be thought of as an edge in the graph. It is common to include multiple named graphs within a database (65), triples in named graphs are represented as quad - where the fourth value in the tuple is the identifier for the named graph. Vocabularies (or ontologies) are used to define concepts and relationships (predicates) in RDF datasets.

4.2 Transparency, Provenance and Privacy

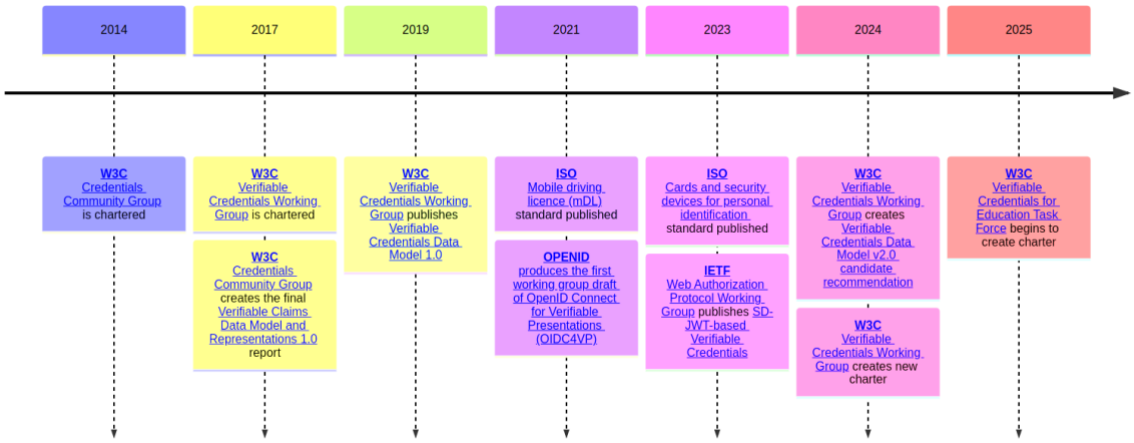
There is real-world evidence that these *values* are desired. In support of transparency and epistemology, C2PA (66) and Verifiable Credentials (31) are standards for maintaining provenance trails of media (e.g. photos, videos) and credentialed data (e.g. digital drivers' licence) respectively. C2PA is now widely adopted amongst most media organisations – as well as many social media platforms – as a means of identifying whether media content is real and whether it has been manipulated. As expounded in further sections – Verifiable Credentials are enjoying regulatory support from bodies including the European Union and UK Government as a means of effectively proving that authoritative statements have been issued.

Transparency and *epistemology* can often be at odds with *ownership* and *privacy*. The first two future research questions close this gap by producing mechanisms for data minimisation that maintain effective provenance trails. This enables sensitive information to be robustly minimised before being shared, whilst the recipient of the data can maintain confidence that the data is correct – or at least derived from claims made by trustworthy information sources.

4.3 Verifiable Credentials

In this section the author provides a condensed review of Wright's (27) work describing data wallets.

Verifiable Credentials refer to a suite of W3C, ISO and IETF standards for signing data. There are three entities typically defined within these standards: the **issuer** – responsible for creating and signing the data, the **holder** – who collects signed data from the issuer and is usually the data subject, and the **verifier** – who the holder forwards the signed data to as needed.



4.3.1 W3C Standards

The W3C were first to work on many standards around Digital Credentials, after the formation of a Credentials Community Group in 2014. By 2017, this group had published their Verifiable Claims Data Model and Representations 1.0 (31) which defined how to express signed credentials. This specification was prescriptive of core functionality such as *how to sign* credentials, describe core "metadata" such as *who issued the credential*, *when the credential was issued* and *who the credential is about*. The specification intentionally left the task of defining the data structures of domain specific credentials - such as a *diploma credential* or *digital driver's license* out of scope. Instead, allowing arbitrary credential types to be listed.

Even *within* this W3C specification there is a tension in the *format* that should be used to describe the content of credentials. The specification provided a description of how to describe credentials using both JSON and JSON-LD (35). The Linked Data (67) community advocated for the use of an RDF (34) data model for its semantic richness, extensibility, and interoperability, aligning credentials with the broader Semantic Web vision (68) - and compromised to use JSON-LD as the encoding for this data model.

4.3.2 ISO Standards

The ISO Verifiable Credential Standards such as ISO/IEC 18013-5:2021 Mobile driving licence (mDL) application (32) describe a fixed schema for describing approximately 30 attributes in digital driver's license's - such as the drivers name, address, date of birth and the expiry date of the license. The specification expects attributes to be serialized using JSON or CBOR (69) and thus lacks the out-of-the-box interoperability that comes with linked-data formats.

This means that it is very well-defined how to build an architecture specifically for mDL licenses. The trade-off is that implementors need to build custom transmission flows, and query engines to support the specification. This both increases implementation burden and hinders interoperability with non-mDL credentials.

ISO is also working on several other Verifiable Credential standards - including Cards and security devices for personal identification (70) designed to standardize core features for electronic identity document including drivers licenses, passports, residency permits, and

building passes. The underlying goal of the standard is to support interoperability between electronic identity (eID) systems. This standard also defines a range of attributes that may be required in different eID systems - extending those attributes found in the mDL license with attributes such as Business Name, Profession, and Academic Title to support workplace passes, as well as other attributes such as telephone number and email address. This specification also targets JSON and CBOR formats for encoding data in credentials - meaning that there are still interoperability challenges with systems that need to define attributes that are not defined within this document.

4.3.3 IETF Specifications

The IETF is also producing a set of JSON based verifiable credentials called SD-JWT-based Verifiable Credentials (71).

JSON Web Tokens (JWT's) are commonly used on the Web today for a range of tasks requiring signed data - for instance they are often used to prove to a website that you are logged in and allowed to access private information on a website.

Selective Disclosure (SD), as discussed above, is a mechanism for proving that a subset of information within a credential is true - without revealing the whole credential to a verifier.

As the Internet Engineering Task Force (IETF) is responsible for producing a number of Internet Standards - the SD-JWT-based Verifiable Credentials has been produced with the goal of allowing digital credentials to be easily integrated into existing internet systems - such as OAuth authentication flows - which are commonly used for single sign on.

4.3.4 European Digital Identity (EUDI) Regulation

After three years in the making, the EUDI (European Digital Identity) wallet officially came into force on May 20, 2024 - through the eIDAS (Electronic Identification, Authentication, and Trust Services) 2 regulation (72). EUDI promises to make "EU Digital Identity […] available to EU citizens, residents, and businesses who want to identify themselves or provide confirmation of certain personal information." By 2026, every EU Member State will be required to make at least one Digital Identity Wallet available to all citizens and residents. There are three core types of credentials that are to be made available under eIDAS 2 regulation:

- Electronic Attestation of Attributes (EAA) - which can be issued by *any* organisation that wants to make statements about a particular entity (e.g., they have a concert ticket, gym membership, or student card)
- Qualified Electronic Attestation of Attributes (QEAA) - which can be issued *only* by Qualified Trust Service Providers to create legally binding credentials such as professional qualifications, birth certificates, marriage licenses, property deeds and business operating licenses.
- Personal Identification Data (PID) - which can be issued *only* by government authorities and serve as a proof of identity.

The European Union has produced an Architecture and Reference Framework (73) which specifies:

- How to issue PID data using both the ISO mDL specification and the IEEE SD-JWT specification can be used to format the data, and how verifiers can request data using OID4VP.
- That (Q)EAA's MUST be issued in accordance with either the ISO mDL data model or the W3C Verifiable Credential Data Model.

4.3.5 UK Digital Verification Scheme – and the Digital Identity and Attributes Framework

The Data (Use and Access) Bill is proposed legislation currently at committee stage in the House of Commons. One mandate of the bill is to create a Digital Verification Services (DVS) Trust Framework - driven by the Secretary of State maintaining a register of *service providers* accredited to provide some "digital verification services" in the UK.

The Digital Identity and Attributes Framework (DIATF) has been created by the Department of Science and Technology (DSIT) in the UK, as a framework defining the *services* that different service providers in the UK can implement and become registered as a DVS service. The DVS may be seen as the UK's equivalent to eIDAS regulation, whilst the DIATF may be seen as equivalent to the EU's Architecture and Reference Framework (73).

Notably, the DIATF is less prescriptive of which standards must be used - and places more of a focus on the roles of different service providers. In the latest iteration of this framework, 5 service providers were defined – Identity, Attribute, Holder Orchestration and Component.

4.3.6 The goal of regulation

Fundamentally, the goal of these technologies is to *build trust*. By empowering:

- Organisations as *authoritative sources* to assert information – such as marriage licenses, and
- Providing technical architectures to *prove* that individuals or organisations made assertions

4.3.7 Selective Disclosure

Many headlines surrounding digital credentials - promise the ability to "prove your age without revealing any other information."

To enable this, some Verifiable Credentials are built with the capacity to perform Selective Disclosure (SD). In short, this allows one to take a Verifiable Credential containing lots of information, such as a Resident Card credential - and forward only part of the information, such as the birthDate to the *verifier*, whilst enabling the *verifier* to confirm that the date of birth was contained in a validly signed Verifiable Credential.

Selective disclosure is typically supported using Zero Knowledge Cryptography. In particular, the W3C Verifiable Credential Standards perform selective disclosure in the Data Integrity BBS Cryptosuites v1.0 (74) specification by: parsing the credential as set of facts (specifically

RDF triples or quads), individually hashing each fact and then signing the set of hashes using the BBS Signature Scheme (75). This signature scheme enables proof that a subset of messages signed using the scheme are true, without needing to reveal all the signed messages. In turn, this enables proof that a subset of the facts that are in a digital credential are true. This signature scheme is derived from the 2004 work of Boneh et. al (36).

Despite the further advances in Zero Knowledge Cryptography which will be expounded in later sections – selective disclosure is the only form of zero knowledge proof currently supported within digital credential standards. This means that it is not possible to prove that computed or inferred results are true – for instance, it is not possible to generate a proof that one is over 18 from a credential containing a signed date of birth.

Digital Driver's license do promise the ability to do proof of age verifications (76). Taking a deeper look at the ISO Mobile driving license (mDL) (32) reveals that the **issuer** (e.g. the DVLA) has to *explicitly sign statements* about your age. Practically, this means that:

- One must tell the issuer (DVLA) that I want to prove I'm over 18 - when this isn't something they need to know.
- One is *reliant* on the *issuer* (DVLA) to issue these statements - so if a driving authority doesn't want to issue is_over_21 statements; one may be forced to reveal their age. Whilst this is less problematic - and less likely - in the case of age; it is an issue when trying to any *non-standard* derivation. For example, proving non-caucasian ethnicity, without revealing the minority population that one belongs to.
- One cannot tell the verifier about information that can be derived from multiple credentials. For instance, it is not possible to prove to a car hire agency that one can drive in the UK without giving them details from ones license, visa and passport.

4.4 Self-Sovereign Data and Personal AI Infrastructure

The drivers for developing this work which support sovereign data and AI architectures are largely to address social, rather than technical challenges. The architectures that different individuals and groups strive for are often informed by their *values* and *philosophies*. Moreover, Verhulst (77) highlights that the operational deployment of these systems often reflects and influences the hierarchies and relationships of power in society.

We are not necessarily discussing *individual* sovereignty. Other groups or entities that may be characterised as desiring sovereign architectures include *countries*; *indigenous populations*; *consumers*; *private-sector, governmental and non-governmental organisations; experts/professionals*; *societies*; *intergovernmental organisations* and *patients* (Hummel et. al (78)).

Far from being hypothetical, we can observe these power-play phenomena within modern digital infrastructure. Floridi (12) points to examples including centralised social media becoming a vector for mass social and political messaging – additional to being a valuable source of revenue and facing numerous threats of being shut-down by governments. More recently, we have seen the UK Government exercise regulatory powers that saw Apple

remove end-to-end encryption from its cloud services; in this instance eroding privacy in a manner uncontrollable by most of the above sovereign groups.

Verhulst (77) abstracts this and observes three key asymmetries "especially for already vulnerable and marginalized groups: data asymmetries, information asymmetries, and agency asymmetries. These asymmetries limit human potential, both in a practical and psychological sense, leading to feelings of disempowerment and eroding public trust in technology." whilst Hummel et. al (78) observe that most often "the following kinds of mutually connected issues are salient: constitutive, technical, epistemic, and legal challenges."

In their survey paper, Hummel et. al (78) identify 15 values which drive and inform the design of sovereign infrastructure. These are: *control and power*; *security and non-maleficence*; *deliberation, representation and inclusion*; *privacy*; *ownership*; *transparency, epistemology*; *effectiveness, complexity*; *autonomy*; *autarchy*; *beneficence*; *dignity, fundamental rights, identity*; *emancipation, empowerment*; *trust, reliability*; *justice*; *responsibility* and *recognition, respect*. Of these values, the architectures we now drive in this thesis most directly support **transparency, epistemology, ownership** and **privacy** of data – whilst supporting user **autonomy** and **empowerment** in their use of AI systems.

4.5 Value Centred (Sensitive) Design

Value Centred Design (VCD) – also often referred to as Value Sensitive Design (VSD) – emerged as a design philosophy within the field of HCI which prioritises user, enterprise and societal (often termed stakeholders) values in the design process (9). Friedman et al. (79) defined VSD as "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process."

Values reflected through VCD may include *privacy*, *autonomy*, *fairness*, and *transparency*. By accounting for human values in a principled manner, VCD intends to ensure that technologies not only meet practical user-needs – as is the focus of user-centred design – but also to ensure that systems are socially sustainable and beneficial (80).

Sovereign Digital Architectures (SDA) are built with primarily value-based goals in mind, rather than to solve physical or technical challenges (12). Often these technologies are developed with the designer(s) considering themselves as the primary user; and thus, the values these systems reflect are often those of the developer as the sole-stakeholder (81).

4.6 Value Centred (Sensitive) Design of Sovereign Digital infrastructures

There appears to be little work at the intersection of VCD and SDA despite both having values-based approaches to technology design, and VCD having emerged from the study of information systems. The researcher hypothesises that this is because the developer community developing SDAs often build for *their* values – whilst VCD is generally a more principled approach and with a focus on the stakeholder values. That said, many values including *privacy* and *freedom* are commonly emphasised across the two domains. Ishmaev et. al. (81) started to explore this idea of applying VCD to SDA specifically in the context of

Self Sovereign Identity (SSI) on the basis that "The problem of ethical issues in identity management solutions is an underdeveloped topic, and yet, one of the most critical concerns in our increasingly digitalised society." Ishmaev et. al. (81) explicitly identified the fact that there was no prior work on the application of VSD methods to SSI solutions. Further these authors identified the following set of stakeholder values that may be met with SSI infrastructures: *autonomy, control, agency, transparency, trust, privacy,* and *security*.

4.7 State of the Art in Zero Knowledge Proof

In this section we discuss the three main paradigms (abstractions) for creating Zero-Knowledge Succinct Non-Interactive Argument of Knowledge (zk-SNARK) (82). zk-SNARKs are a specific form of Zero Knowledge Proof which allow the proof issuer to generate an entire proof without interacting with the proof verifier in any way. The three core paradigms (abstractions) we discuss are **Selective disclosure of messages with BBS style signatures** (36), **Zero Knowledge Circuit Builders** (83) and **Zero Knowledge Virtual Machines (ZKVMs)** (37).

We have already discussed selective disclosure signature schemes such as BBS signatures in the above section on Verifiable Credentials, so we omit further discussion here.

Zero Knowledge Arithmetic Circuits builders (ZK Circuits) - such as circom (83), plonky3 and halo2 (54) – provide a layer of abstraction for building Zero Knowledge Proofs. Specifically, circuit builders allow programmers to describe a "circuit" which is a set of constraints. The expressivity of the constraints is dependent on the circuit builder – circom for instance has support for quadratic constraints. The circuit builder can generate a zkSnark indicating whether a particular set of constraints is satisfied with a given parameterisation – without revealing any further information.

There is existing work demonstrating that ZK Circuits can be used to prove correct execution of query engines, particularly Gu et. al (84), proved that it was possible to generate Non-interactive Zero-Knowledge Proofs for Arbitrary SQL-Query Verification in developing PoneglyphDB (84) with the Halo 2 Circuit Compiler (54) .

Zero Knowledge Virtual Machines (ZKVMs) enable the proof that arbitrary application code has been correctly executed and produced a given result. Risc0 (37) is one such Zero Knowledge Virtual Machine and is used to prove correct execution of RISC-V instruction sets (38). Since higher level languages including Rust can be compiled into RISC-V instruction sets, it is possible to use the Risc0 ZKVM to prove correct execution of Rust code. There numerous other ZKVMs including Ceno (85), SP1 (86), Nexus[21], Powdr (87) and ZkMIPS (88).

4.8 Multi-party computation over query languages

There is substantial prior work on standard languages for representing MPC, more than a dozen Domain Specific Languages (DSLs) have been developed to describe MPC. Notably, the *Secure Multiparty Computation Language (SMCL)* (89) has been developed as *declarative programming language for Secure Multiparty Computations*, but does not contain descriptions for the security assumptions required for MPC calculations. Wys* (90) - co-

[21] https://docs.nexus.xyz/home

developed by Microsoft Research and the University of Maryland presents a DSL for Multi-Party Computation which provides program logic to reason about the correctness and security of MPC programs. Consequently, we will first do a systematic review of these systems to establish if any meet our needs.

When it comes to SMPC for databases and query languages – most related literature is around Secure Multiparty Computation (SMPC) (91) over relational databases – such as those query able via SQL query interfaces. Such works include SMCQL (92), Conclave (93) and Senate (94).

SMCQL is a framework for executing SQL series over a Private Data Network (PDN), where a user submits a query to an honest broker which the orchestrates the Secure Multi-Party Computation over the Private Data Network with an honest-but-curious threat model. SMCQL supports joins, aggregations and group-by queries. Conclave supports a similar set of operations to SMCQL but allows weakening of its security model to achieve improved performance. Senate was developed after SMCQL and Conclave, and through the planning protocol developed – which enables more parallelisation of computation, and compartmentalisation to identify when subsets of nodes work towards a particular result – has a performance that is orders of magnitude better than SMCQL and Conclave. Senate additionally supports a stronger *malicious security* guarantee.

There does also exist some work towards supporting SMPC over fragments of SPARQL. Goose: A Secure Framework for Graph Outsourcing and SPARQL Evaluation (95) uses secure multi-party computation to achieve the following features:

(i) *no cloud node can learn the graph,*
(ii) *no cloud node can learn at the same time the query and the query answers, and,*
(iii) *an external network observer cannot learn the graph, the query, or the query answers*

However, GOOSE is limited to support Unions of Conjunctions of Regular Path Queries (UCRPQ) and does not support common numeric or build-in operations such as COUNT, SUM and AVG. Further, GOOSE requires an honest broker to design the query plan and communicate it to the compute cluster of graph databases and has a fixed assumption that the graph databases executing the query are honest-but-curious. That is, the databases can be trusted to execute the plan given to them by the broker.

SMPG: Secure Multi Party Computation on Graph Databases (96) has been produced as a position paper and prototype for automatically executing MPC evaluation of Cypher queries (22) over Neo4J (97) databases. SMPG is built using Conclave and so has matching weak security assumptions, performance and expressivity challenges.

Cypher is problematic for distributed queries as identifiers are local to the database. Consequently, queries often must explicitly disambiguate entities by identifying the which node it occurs in within queries MATCH(node1:label1). This is not amenable to one of our driving goals which is to abstract away all underlying architectures to the greatest extend possible and have a pure data-layer for systems to work with.

4.9 Contemporary Identity Infrastructure

Generally, the goal of identity architectures is to support online services in establishing confidence that an agent – which could be a person, persona, or any piece of digital architecture from an AI agent to small script – is the agent they claim to be. This can be further generalised to establishing confidence that the agent you are interreacting with satisfies a set of properties – such as being over the age of 18.

Identity, and digital identity, are complex topics - fraught with extensive debate. The matrix presented by Wright (98) provides an overview of the values most commonly desired from self-sovereign identity (SSI) infrastructure, and how a range of existing identity solutions satisfy those values. The three most common identity models are: **centralised identity** – i.e. traditional log-in where you must directly authenticate with the service you are using, by providing proof of identity such as email and password; **federated identity** – i.e. Single Sign on Flows which are commonly seen today, where service providers *trust* platforms like Google to authenticate you and attest to your identity; and **self-sovereign identity** - where users prove their identity directly, e.g., through the use of self-managed public-private key pairs, reducing reliance on trusted intermediaries – and improving privacy as these intermediaries no longer need to be privy to online transactions.

To further this complexity, there are a diverse range of protocols and associated implementations for these, and other, identity models. The Decentralised Identity Foundation (DIF) alone supports dozens of identity specifications and has hundreds of member organisations.

4.10Formalising and Generalising Authentication protocols

There are two primary streams of work we see as related to research question 3. The first is formal representations of the Web architecture, and specific authentication protocols for the purpose of performing security research. The second is work on support for dynamic authorisation mechanisms which allow clients to negotiate with servers to determine the attributes they must present to a server to be authorised to access a resource.

Whilst these two domains can be drawn upon in working towards the completion of this research topic, the formal verification work lacks complete mechanisms to fully describe flows such as authentication and authorisation flows in a manner that is fully machine interpretable. The dynamic authorisation protocols that we shall discuss fall short as they are built upon existing opinionated authorisation protocols such as OAuth and User-Managed Access (UMA). Consequently, they are still tied to very specific architectural choices around identity architectures – such as being reliant on centralised IDPs for identity provision by design in many cases. Most dynamic authorisation protocols we reference are also built with the specific use-case of a client, such as a website, obtaining delegated authorisation (i.e. from a User) to access an authorised resource on the Web; meaning that the approach still requires fixed flows for many parts of the authentication process – such as establishing how clients obtain proof that they can obtain access to a resource from an authority (e.g. the user) – let alone allowing for the adaptable creation of flows that are not strictly related to authentication.

Turning to the formal verification literature, we have found OAuth2.0, OpenID Connect (OIDC), FIDO UAF and SAML to be the authentication and authorisation protocols that have most relevant work on formal verification. The most useful pieces of work have (some) of the protocol described using ProVerif. ProVerif (99) is a cryptographic protocol verifier that can prove the *secrecy*, *authentication*, *strong secrecy* and *equivalences* of authentication protocols that have been described in pi-calculus. ProVerif is based on the Dolev-Yao (100) formal model for interactive cryptographic protocols. Bansal et. al. (101) use ProVerif in their analysis of OAuth2.0, as do Feng et. al. in their analysis of FIDO (102).

The second most popular approach, though of less utility in our work, is works use Fett et. al's (103) formalisms for Web architectures to perform analysis of potential attack vectors in authentication mechanisms on the Web. This formalism is purely mathematical and does not immediately lend itself as a descriptive language for describing the above authentication procedures. Whilst it may be possible to develop a formal verification system on top of Fett et. al's (103) work, none have been encountered in our research. In particular, Fett has a applied this formalism to analyse OAuth 2.0 (104) and OIDC (105).

It is also prudent to observe that increasingly authentication and authorisation standards have overlapping requirements, and specs used to verifiable credential standards. After all, credentials are often used to prove that an entity has attributes, which includes attributes such as "is over 18" or "is a UK Doctor" which allows identity and role-based access to authorised resources on the Web. Specifically, this is often used within the OAuth 2.0 extension – User Managed Access (UMA) 2.0 (51), where W3C Verifiable Credentials can be included within auth scopes. UMA is now used across a range of projects including – and is deferred to by several W3C standards including Solid-OIDC (106). Credentials are also being used as the mechanism to provide AI Agents with delegated access (107) to perform operations on behalf of a human-entity.

IETF Standard Grant Negotiation and Authorization Protocol (GNAP) (108) has overlapping goals with UMA. Both in terms of having a constrained scope to allow permissioned access to resources and allowing clients to negotiate with resource servers to prove that they have appropriate levels of delegated authorisation to be able to access said resource. Whilst this ability to negotiate the information that is revealed provides a small reduction in how static the authorisation process is, most of the flow is still pre-defined and hence is not a sufficient starting point for this work.

There is also a range of work on *protocol languages.* These include the Blindingly Simple Protocol Language (BSPL) (109) which follows an Interaction-Oriented Programming (110) approach. Whilst such protocol languages may be useful within our work on research topic 3, they are limited to the expression of potential message ordering – and thus do not bridge the gap of automating the generation of an authentication flow from the description of a *goal*.

4.11 Personal AI Agents

The discussion for this section is applicable to the researcher's early work throughout their thesis, and is thus included for completeness – however, it is not directly relevant to items discussed in this transfer report.

4.11.1 Overview

The notion of a Personal AI Assistant is not new. Wooldridge (111) gives an example of a Personal Digital Assistant (PDA) which "converses with several different Web sites, which sell services such as flights, hotel rooms, and hire cars. After hard negotiation on your behalf with a range of sites, your PDA presents you with a package holiday" (111 p. 6) use the term Virtual Personal Assistant to describe "any **device** […] that provides professional, technical, or social assistance to automate or simplify daily tasks" (112 p. 1), and Searls (113) use the term Vendor Relationship Management to describe the "customer-side counterpart of CRM, or customer relationship management […] that would make individuals both independent of vendors, and better able to engage with them." (113 p. xii).

The concept of AI existed as early as 1955, first coined by American Computer Scientist John McCarthy the focus was to "find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (114). Today the scope and definition of AI is largely undefined. We view an 'AI system' to be a "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." This means that a wide range of systems are in scope for discussion, from non-interpretable 'black-box' systems such as generative Large Language Models (LLMs) (40), geometric deep learning networks, and logistic regression classifiers which 'learn' to generate or predict outputs based on masses of sample training data, through to interpretable and predictable rules-based systems which execute a fixed set of instructions explicitly set by humans.

4.11.2 AI agents

So, what distinguishes AI and an AI agent? An agent is "a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives" (111 p. 16) where autonomous action is the capability of agents "deciding for themselves what they need to do in order to satisfy their design objectives" (111 p. xi). Further, we expect the agents to be intelligent agents, characterized with **reactivity** – being able to understand, and effectively respond to their environment, **proactivity** – taking the initiative to service users, and **social ability** – being able to interact with the human they represent, as well as other agents.

What can constitute the environment for an agent is quite broad. For the purposes of this paper, there are two important environments in question – first, is the set of other agents that an agent will interact with, by messaging them on the Web. As we shall detail later in this section these agents will typically be either "Service Provider AI Agents" or other "Personal AI Agents" that represent different people. This is a typical multi-agent system (MAS) construction where agents communicate "not simply by exchanging data, but by engaging in

analogues of the kind of social activity that we all engage in every day of our lives: *cooperation*, *coordination*, *negotiation*, and the like" (111 p. xi). By Wooldridge's characterisation of agentic environments, this environment is *inaccessible* – as the agent does not have access to complete information about the action space or knowledge of other agents; *non-deterministic* as there are no guarantees as to how other agents are built – and hence respond, *dynamic* as the members of the agent system can change over time, and *continuous* as this multi-agent system is responsible for handling ongoing service interactions. Searls suggests that Personal AI Agents should primarily perform *intent casting* in this environment – for example, broadcasting the message "I want to buy 2 plane tickets from London to Berlin on Sunday Feb 9, 2025, departing between 6 and 9pm," which Service Provider Agents representing airlines would bid serve. Intent can also be broadcast to other personal agents – for instance "I would like to meet with Janet on w/c Feb 12, please suggest times that would suit."

The second environment we consider is the "real-world" environment in which the agent interacts with the user. This environment is the set of inputs provided by the user and their auxiliary devices, and the means by which the agent can respond or prompt. These auxiliary devices can range from an air-quality sensor providing data to the agent at fixed intervals, through to a voice assistant the agent can interact with, or a humanoid robot controlled by the agent. Additional user and auxiliary data may be made available to agents with access to Personal Data Stores such as a Solid Pod (14) – enabling agents to access any digital information collected about users, within the bounds of what users consent for the agent to access. By Wooldridge's characterization of agentic environments, this environment is *inaccessible* – as the agent does not have access to complete information of the users' world; *non-deterministic* due to the unpredictability of users and their environments, *dynamic* as users and their environment change over time, and *continuous* as the action space of the agent is not fixed nor is it finite.

More recently, Gartner defined Agentic AI as "Autonomous AI can plan and take action to achieve goals set by the user" and have identified this as the top strategic trend for 2025 (Alvarez, 2024). However, Gartner envision these agents taking on normative roles (Zhi-Xuan *et al.*, 2024) within organizations – such as being integrated into a SaaS platform to replace some of the functions of a customer service representative. This is *not* in alignment with our vision of Personal AI Agents are *decoupled* from service providers and strictly represent the "best interests" of the consumer.

4.11.3 How consumers interact with agents

There are various modalities by which these AI systems may receive input and produce output. At first, there were algorithms called on demand by programmers running commands on their machine. This has significantly evolved over the last four to five decades, with the emergence of chat-like interfaces to interact with LLMs such as ChatGPT in 2021/2022, and now a rise in popularity of Embodied AI (EAI) (115). EAIs are AI systems with some form of physical embodiment – be it a webcam and screen providing a visual interface, or a full humanoid system with sensors that can capture all five human senses of sight (vision), sound (hearing), smell (olfaction), taste (gustation), and touch (tactile perception).

We consider both Embodied and non-Embodied Personal AI Agents to be within scope. We expect that just as with the AI services we interact with today, the modality with which it is

appropriate to interact with an agent will be highly circumstantial. For instance, when consenting to having an agent purchase one's weekly shopping may take the form of hitting accept on a mobile notification; while planning a trip may involve a verbal discussion with a voice assistant to illicit preferences and allow a range of decisions to be made in a short period of time – much like working with a human travel agent. This multi-modality is a crucial feature when building personal AI Agents for vulnerable individuals.

4.11.4 Delegated control and decision making

When it comes to non-interpretable AI systems such as LLMs, there is increasing discussion around the topic of alignment. The traditional *preferentialst* approach to alignment seeks to have AI systems understand the preferences of one, or more, users of the system and act in line with these preferences (116). In cases where personal AI Agents have delegated authority and decision-making power (107) this means that alignment results in a best-effort approach to emulate the decisions that the user would have made. More deterministic and rules-based systems is implemented by having users explicitly define what tasks an agent can *autonomously* perform; and the decision criteria that should be used when performing the task. A naïve instance of such an agent would be an email filter, which has a fixed set of rules to determine in which folder an email should be placed based on the sender and content of the subject. The kind of personal AI agents that are the focus of this paper, the constraints of what an agent is authorized to perform may be rules such as "do not spend more than $100 over the course of a week without my [the user's] authorisation," and the decision making criteria would be largely outline fixed preferences within particular task-scopes "when booking travel pick the cheapest hotel listed on my approved companies travel list, within a 500m radius of the conference."

For more contemporary machine-learning systems, a range of approaches are applied to align decision making preferences. One such emulative approach includes task-specific predictive systems – for instance, a machine learning system that identifies the products a user would buy by collecting a dataset describing the browsing history of a range of users, and the purchases they made – and then training a machine learning model to predict purchases based on user interaction with the browser over time. Note that this is the kind of predictive machine learning that powers targeted advertising in online platforms.

Similarly, the more generalist ChatGPT has been trained by "predicting" the sample output of a set of input text; and then having the response refined using Reinforcement Learning from Human Feedback (RLHF) such that the output is "defined by human judgment, building a model of reward by asking humans questions" (117) (118). In cases such as that of ChatGPT, this process of RLHF is *not* done to align the system to a set of individual user preferences; instead, the system is being trained to comply with specific normative criteria (116) including helpfulness, harmlessness, and truthfulness (119) (120). These normative roles are communicated to the human employees and contractors of OpenAI tasked with providing the system feedback for RLHF.

In both the predictive-purchasing, and ChatGPT example; these systems are *not* being designed to emulate the preferences of an *individual user* but rather be predictive of the behaviors of a population at-large. In contrast, we expect that if a personal AI Agent uses

machine learning and is *preferentialist* then the system would specifically try to emulate the *user intent* when decisions are delegated to the agent.

This both calls into question how we align with user intent and whether we should be aligning with user intent at all. As to whether it is possible to *align with user intent*, Zhi-Xuan *et al.* (116) observe that the traditional *preferentialist alignment* (121) approach for machine-learning AI systems makes the false assumption that *humans are themselves rational decision makers*, that can *capture their values as a set of preferences* and *always act to maximize those preferences*. When this assumption breaks; it becomes very difficult for a system to discern a clear set of criterion upon which to establish if it is following user intent – much as a human personal assistant can only roughly guess the decision-making procedures of their superior, and never fully emulate them.

There is also a further discussion of whether we should be instead building systems that are not *value* or *preference* aligned, but instead "optimised" in other ways – such as making decisions that are in the interest of the users' long-term wellbeing. Zhi-Xuan *et al.* (116) suggest that systems should always be designed to fulfil normative societal roles – such as a travel planner, psychologist or manager. Some argue that we should perform **thick value alignment** to ensure AI is aligned with human values at large (122 p. 137). *Ji et al.* (123) suggest that when doing such **thick alignment** there are four guiding design principles to be accounted for Robustness, Interpretability, Controllability, and Ethicality (RICE). We highlight this as a critical open ethical question in the design of personal AI Agents.

Noting all the above alignment challenges, we expect that in the near term, most Personal AI Agents will be a hybrid of deterministic rules-based systems and black-box symbolic systems – a simplistic example of this is presented in (17). For the most part, we expect that the user delegates control to the agent using rules-based "authorisation controls" (107) and within these bounds a neurosymbolic system performs decision making according to some form of alignment.

4.11.5 Distinguishing personal and personalized AI

Personalized AI is characterized by being in some way tailored, or in some way self-tailoring for a particular user. Examples of Personalized AI Agents include *recommender systems* (Ko *et al.*, 2022), *smart home assistants* (112) (124)*, conversational LLM's such as ChatGPT* and *Computer Using Agents (CUAs)* such as OpenAIs operator agent[22]. What most of these personalized agents have in common, is access to some degree of personal data with which to inform their interactions with users. For *recommender systems* it is previous watch history to prescribe suggested shows, *home assistants* have access to calendar data to alert you of upcoming events *Amazon Alexa* further supports contextualized discussions – such as about one's interests, and learns repeated user behaviors and notifying them with a "hunch" that they may have forgotten something. Another common feature is tailored mannerisms. Voice assistants such as *Amazon Alexa* which have customized voice profiles, *ChatGPT* - which uses past conversations with a user to provide context to the current conversation, thus making the result more *relevant* to users; and also makes the conversational LLM begin to act *more*

[22] https://openai.com/index/introducing-operator/

like the user[23].

While we expect all *Personal AI Agents* to be *Personalized AI* – the converse is rarely true. The earlier discussion around alignment is what fundamentally distinguishes the personal AI Agents we discuss in this paper from personalized AI, which is more prevalent in the existing service literature. Alexa is a good example where the system is not aligned with user *intent* or interests – as users are often recommended to buy products by the device; not because they are what the user would normally choose to buy, or are necessarily in their best interest to buy, but instead because the system is marketing a product to them. This is exactly why there is a need for *personal* agents which advocate for users.

Modern Personal AI Agents are beginning to emerge. Kwaai.ai[24] for instance is a non-profit lab building "a [self-sovereign] Personal AI Operating System to allow you to train your own personal assistant, privately [and] securely."[25] It is led by Doc Searls who invented the concept of the *intention economy* and *vendor relationship management.* To some extent open source frameworks such as BabyAgi[26] may also be considered to be working towards Personal AI Agents, by laying the groundwork for end-users to design custom AI agents for themselves.
We need to consider how we can ensure that Personal AI Agents are not operating with ulterior motives when deployed in practise – for instance, how can we know that our personal AI Agents are not just *Personalized AI Agents* in disguise and ultimately working in the interest of a particular organization by "manipulating" us to buy specific products or services, much as current "attention economy" services do today (113). One approach is to encourage the development of open-source implementations of Personal AI Agents such as kwaai.ai, which can then be deployed locally on individuals' devices – while a nice ideal, this still requires most end-users to trust opensource developers in their design and implementation of such agents, with very little means to understand what has been done. A more compelling answer may lie in making companies that implement services for Personal AI Agents legally accountable – and subject to fines if the agents they implement are found to be in anyway make decisions to better the commercial interests of the company rather than the user once deployed. If too heavily regulated, however, this risks stifling the development of such agents.

4.11.6 What is meant by Personal AI Agents

To summarize a personal AI agent is an agent operating within a multi-agent system of personal AI agents, and service provider agents. The ideal Personal AI agent *must* represent user interests, that is, have **alignment** with the values and intentions of the individual user when given authority to act autonomously, and support their self-determination (125). The scope, or granularity, at which agents may which act autonomously is to be user defined – if deployed at scale, we anticipate that there will be a range of preferences that users have for the degree of autonomy they wish to delegate; some customers we expect to place expectations such as "notify me before making any *social, legal* or *contractual agreement*," while others may set looser bounds for autonomy "bring me into the loop if you plan to spend more than £200 in the course of a week."

[23] https://help.openai.com/en/articles/8590148-memory-faq
[24] https://www.kwaai.ai
[25] https://drive.google.com/file/d/1lHxy0q3z5krG8hBwYIG6bDloIuzagSce/view
[26] https://babyagi.org

In contrast to most AI Agents, and Personalized AI the service literature, these the ideal versions of these agents *are not* to be implemented by providers of a particular service – but instead interact with the service provider while representing user interests.

4.12 Further Reading

Please see Wright (27) for a literature review of Verifiable Credential standards and Data Wallet Regulation (126), Wright et al. (16) for a literature review on automated data governance, and Wright (127) for a position piece on communication in multi-agent systems.

5 Appendix

5.1 Figures

5.1.1 Figure 1

For visibility – I give a brief overview of other research which I have **partially completed** in the first year of my DPhil, **ongoing collaborations** and **adjacent work** I am involved with and **long-shot work items** which are pieces of research with a low likelihood of success that I would like to continue pursue once I have satisfied the minimal requirements for my DPhil – which are the focus of this Transfer of Status report.

5.2 Partially Completed Research

There are several research directions that we have *partially completed* and may re-visit once the next three planned papers are complete:

1. Generating formal **ODRL Policy** (58) descriptions for cookies: There is extensive work on making the Web aware of policies, including Oshani et al. (128), and privacy preferences (129) such that policy engines (130) may be used to enforce usage controls upon data sent over the Web. As Wright (16) discusses, one such context

where we wish to apply usage controls is upon cookies sent from browsers to websites. This requires cookies to have a formal description of their usage controls, these codebases have been developed towards that goal:

   a. https://github.com/jeswr/cookie-purpose
   b. https://github.com/jeswr/purpose_classification
   c. https://github.com/jeswr/cookie-analysis
   d. https://github.com/jeswr/cookie-terms-of-use-extension
   e. https://github.com/jeswr/who-do-i-trust-extension

   I would like to complete this work, and submit with a title along the lines of **Extending DPV to describe Cookie Purposes** as a resource paper to ISWC this year.
2. Interactions between Trusted Web Agents towards Semi-Autonomous agents at Web Scale as outlined in this Doctoral Consortium vision paper and Graduate Scholarship Application.
   a. https://github.com/jeswr/nhs-trusted-hospitals
   b. https://github.com/jeswr/who-do-i-trust-extension
3. Logical Dialogues between systems:
   a. EYE QA
   b. Dialogical Investigation
4. RDF Representation of mathematics
   a. https://github.com/jeswr/lean2rdf
   b. https://github.com/jeswr/RDF.lean

5.3 Ongoing Collaborations and Adjacent Work

1. Computer Using Personal Agents over Personal Knowledge Graphs: Off the back of this Dagsthul Vision Paper (61) we are working to develop this vision.
2. Communication between LLM-based agents and LLM agent safety.
   a. https://arxiv.org/abs/2410.11905 (LLM Communication)
   b. https://arxiv.org/pdf/2409.04465 (LLM + Semantic Web Agent Communication)

5.4 Long-shot work items

I have a blog post here describing a series of other topics that I am interested in. **I would still like to collaborate on these and other topics and produce first-author publications on these topics once I have published work on the topics discussed below**.

5.5 Publications produced since starting the DPhil:

8. AI-agents in Customer Experience for Vulnerable Consumers (Journal Paper – under review at Journal of Service Management)
9. **Me want cookie**! Towards automated and transparent data governance on the Web (16) (NXDG Workshop Paper)
10. A scalable communication protocol for networks of large language models (18) (Full-length paper)

11. Towards Computer-Using Personal Agents (61) (Short Vision Paper)
12. **Here's Charlie**! Realising the Semantic Web vision of Agents in the age of LLMs (17) (ISWC Demo Paper)
13. **N3. js** Reasoner: Implementing reasoning in **N3. js** (131) (ISWC Poster Paper)
14. EYE JS: A client-side reasoning engine supporting Notation3, RDF Surfaces and RDF Lingua (62) (ISWC Poster Paper)
15. Semi-Autonomous Agents at Web Scale (Doctoral Consortium Paper)
16. Transforming Service Contract Management through an Intention Economy Model: The Case for Semi-Autonomous Web Agents (Service Management Forum)

5.6 All venues where I have submitted to or am considering submission to

- NXDG – Next Generation of Data Governance Conference, co-located with SEMANTICS24 in 2024, Amsterdam, Netherlands; 17 – 19 September 2024. Future years anticipated.
    - Link: https://nxdg-workshop.github.io/2024/
- SMF – Service Management Forum: "Shaping the Future of Service Management", Cambridge Service Alliance, Institute for Manufacturing, University of Cambridge, UK *23-24 September 2024*
    - Link: https://www.servsig.org/wordpress/2024/03/phd-students-forum-in-cambridge-shaping-the-future-of-service-management/
- ISWC – The 23rd International Semantic Web Conference, November 11–15, 2024, Hanover, MD.
    - Link: https://iswc2024.semanticweb.org/
- ESWC – 22nd Extended Semantic Web Conference, June 1–5, 2024, Portoroz, Slovenia.
    - Link: https://2025.eswc-conferences.org/history/
- TheWebConf – The Web Conference 2025, Sydney, Australia 28 April - 2 May 2025
    - https://www2025.thewebconf.org
- CHI – Conference on Human Factors in Computing Systems
    - Link: https://chi2024.acm.org/
- JOSM – Journal of Service Management
    - Impact Factor: (2023): 7.8; 5-Year (2023): 10.1
    - Link: https://www.emeraldgrouppublishing.com/journal/josm
- ICLR – International Conference on Learning Representations
    - Link: https://iclr.cc/
- UseNix
    - https://www.usenix.org
- Semantic Web Journal (SWJ)
    - https://www.semantic-web-journal.net
- European Conference on Advances in Databases and Information Systems

- https://adbis2025.github.io

5.7 Blog posts elaborating on some of those publications:

1. https://blog.jeswr.org/personal-ai.pdf extends the definition of Personal AI for AI-agents in Customer Experience for Vulnerable Consumers
2. https://blog.jeswr.org/2025/05/17/mas-communication provides a more thorough discussion of multi-agent communication for A scalable communication protocol for networks of large language models

5.8 Other artefacts produced so far within the scope of the DPhil:

1. Original DPhil Application (for reference).
2. https://github.com/jeswr/RDF.lean an RDF Library for the Lean4 theorem prover.
3. https://solid-catalog.jeswr.org a catalogue of all applications used in the Solid Project.
4. Collaborated on the development of Doctelligence – an opensource decentralized architectures for AI and data marketplaces, enabling secure, peer-to-peer collaboration without data movement.
5. All the opensource work which can be seen here.
6. Standards Work – I am extensively involved in standards development, particularly W3C standards development, including:
    a. W3C Engagements (Community Groups)
        i. Autonomous Agents on the Web Community Group
        ii. Data Privacy Vocabularies and Controls Community Group
        iii. Notation 3 (N3) Community Group
        iv. RDF JavaScript Libraries Community Group
        v. RDF Surfaces Community Group
        vi. RDF Test Suite Curation Community Group
        vii. SHACL Community Group
        viii. Solid Community Group
    b. W3C Engagements (Working Groups):
        i. RDF-Star
        ii. RDF Dataset Canonicalization and Hash Working Group
        iii. Verifiable Credentials Working Group
        iv. Linked Web Storage Working Group (intend to become specification editor)
        v. Data Shapes Working Group (editor of SHACL-C specification)
    c. Industry Consortia
        i. Lightweight Agent Standard Working Group (co-chair)
        ii. Data Products Working Group
7. Reviewing:
    a. ICLR
    b. Solid Symposium Privacy Session
8. White papers

a. Contributed to Linux Foundation Paper on Linux Foundation paper on Decentralized Platforms and AI

6 References

1. *Man-Computer Symbiosis.* **Licklider, Joseph Carl Robnett.** IRE Transactions on Human Factors in Electronics, s.l. : IEEE, 1960, Vols. HFE-1. 10.1109/THFE2.1960.4503271.
2. *A research center for augmenting human intellect.* **Englebart, Douglas C. and English, William K.** San Francisco California : Association for Computing Machinery, 1968.
3. **International Telecommunication Union (ITU).** Individuals using the Internet . *International Telecommunication Union (ITU).* [Online] International Telecommunication Union (ITU). [Cited: 9 April 2025.] https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.
4. **Berners-Lee, Tim.** *WWW: Past, Present and Future.* s.l. : IEEE, 1996. 10.1109/2.539724.
5. **Dizikes, Peter.** Study: On Twitter, false news travels faster than true stories. *MIT News.* [Online] 8 March 2018. [Cited: 10 April 2025.] https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308.
6. *Social media addiction: Its impact, mediation, and intervention.* **Hou, Yubo, et al.** s.l. : Cyberpsychology: Journal of psychosocial research on cyberspace, 2019, Vol. 13. 10.5817/CP2019-1-4.
7. **Kleinman, Zoe.** Apple pulls data protection tool after UK government security row. *BBC News.* [Online] 21 February 2025. [Cited: 10 April 2025.] https://www.bbc.co.uk/news/articles/cgj54eq4vejo.
8. **Milmo, Dan.** Why are creatives fighting UK government AI proposals on copyright? *The Guardian.* [Online] 25 February 2025. [Cited: 10 April 2025.] https://www.theguardian.com/technology/2025/feb/25/why-are-creatives-fighting-uk-government-ai-proposals-on-copyright.
9. **Friedman, Batya, Kahn, Peter and Borning, Alan.** *Value sensitive design: Theory and methods.* s.l. : University of Washington technical report, 2002.
10. *Value Sensitive Design and Information Systems.* **Friedman, Batya, et al.** Early engagement and new technologies: Opening up the laboratory, Dordrecht : Springer, Dordrecht, 2013. 10.1007/978-94-007-7844-3_4.
11. **Lessig, Lawrence.** *Code is law.* s.l. : Harvard magazine, 2000.
12. **Floridi, Luciano.** The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU. s.l. : Springer Nature, 2020. 10.1007/s13347-020-00423-6.
13. **Preukschat, Alex and Reed, Drummond.** *Self-sovereign identity.* s.l. : Manning Publications, 2021.
14. **Sambra, Andrei Vlad, et al.** *Solid: a platform for decentralized social applications based on linked data.*
15. *Privacy enhancing technologies: A review.* **Shen, Yun and Pearson, Siani.** s.l. : Citeseer, 2011, Vol. 2739.
16. **Wright, Jesse, Esteves, Beatriz and Zhao, Rui.** *Me want cookie! Towards automated and transparent data governance on the Web.* s.l. : CEUR, 2024.
17. **Wright, Jesse.** *Here's Charlie! Realising the Semantic Web vision of Agents in the age of LLMs.* s.l. : CEUR WS, 2024.
18. **Marro, Samuele, et al.** *A Scalable Communication Protocol for Networks of Large Language Models .* s.l. : arxiv, 2024. 10.48550/arXiv.2410.11905.
19. *Semantics and Complexity of SPARQL.* **Pérez, Jorge, Arenas, Marcelo and Gutierrez, Claudio.** Athens, Greece : Springer, Berlin, Heidelberg, 2006. 10.1007/11926078_3.

20. **Date, Chris J.** *A Guide to the SQL Standard.* s.l. : Addison-Wesley Longman Publishing Co., Inc., 1989.
21. **Seaborne, Andy and Harris, Steve.** *SPARQL 1.1 Query Language.* 2013.
22. *Cypher: An Evolving Query Language for Property Graphs.* **Francis, Nadime, et al.** s.l. : Association for Computing Machinery, 2018. 10.1145/3183713.3190657.
23. *GraphQL.* **GraphQL contributors.** 2025.
24. **Drummond, Nick and Shearer, Rob.** *The open world assumption.* 2006.
25. **DuCharne, Bob.** *Learning SPARQL: Querying and Updating with SPARQL 1.1. (p. 23).* s.l. : O'Reilly Media, Inc., 2013.
26. **Hartig, Olaf, et al.** *SPARQL 1.2 Query Language.* s.l. : W3C, 2025.
27. **Wright, Jesse.** Disambiguating Data Wallets. *Jesse Wright.* [Online] 13 March 2025. https://blog.jeswr.org/2025/02/14/data-wallets.
28. **Herman, Ivan, et al.** *Verifiable Credentials Data Model v2.0.* 2025.
29. **Monero, Nicolás Peña.** *Federated Credential Management API.*
30. **Terbu, O., et al.** *OpenID for Verifiable Presentations.*
31. **Sporny, Manu and Longley, Dave.** *Verifiable Claims Data Model and Representations 1.0.* s.l. : W3C, 2017.
32. **ISO: the International Organization for Standardization .** *Personal identification — ISO-compliant driving licence - Part 5: Mobile driving licence (mDL) application.* s.l. : ISO: the International Organization for Standardization , 2021. ISO/IEC 18013-5:2021.
33. **United Nations Economic Commission for Europe.** *UN Transparency Protocol.* s.l. : United Nations Economic Commission for Europe , 2025.
34. **Cyganiak, Richard, Wood, David and Lanthaler, Markus.** *RDF 1.1 Concepts and Abstract Syntax.* s.l. : W3C, 2014.
35. **Sporny, Manu, et al.** *JSON-LD 1.1: A JSON-based Serialization for Linked Data.* s.l. : W3C, 2020.
36. *Short Group Signatures.* **Boneh, Dan, Boyen, Xavier and Shacham, Hovav.** Santa Barbara, California, USA : Springer Berlin Heidelberg, 2004.
37. **RISC Zero.** RISC Zero. *RISC Zero.* [Online] https://risczero.com. [Cited: 7 April 2025.] https://risczero.com.
38. **Kanter, David.** *RISC-V offers simple, modular ISA.* s.l. : Microprocessor Report, 2016.
39. **Wright, Jesse.** Are current standards enough? Towards Verifiable Credentials with expressive zero knowledge query. *FOSDEM 2025.* [Online] 2 February 2025. [Cited: 9 April 2025.] https://fosdem.org/2025/schedule/event/fosdem-2025-5970-are-current-standards-enough-towards-verifiable-credentials-with-expressive-zero-knowledge-query/.
40. **Thirunavukarasu, Arun James, et al.** *Large language models in medicine.* s.l. : Nature Publishing Group US New York, 2023. 10.1145/3641289.
41. *Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering.* **Zhentao, Xu, et al.** s.l. : Association for Computing Machinery, 2024.
42. *SGPT: A Generative Approach for SPARQL Query Generation From Natural Language Questions.* **Rony, Md Rashad Al Hasan, et al.** 10, s.l. : IEEE Access, 2022. 10.1109/ACCESS.2022.3188714.
43. **Allemang, Dean and Sequeda, Jaun F.** *INCREASING THE LLM ACCURACY FOR QUESTION ANSWERING: ONTOLOGIES TO THE RESCUE!* s.l. : arxiv.
44. **Berners-Lee, Tim.** *Let's knock down social media's walled gardens.* s.l. : Financial Times, 2025.

45. **Yang, Hongkang, et al.** *Memory3: Language Modeling with Explicit Memory.* s.l. : arXiv, 2024. 10.4208/jml.240708.
46. **Zhao, Wayne Xin, et al.** *A survey of large language models.* s.l. : arXiv, 2023.
47. **Barrault, Loïc, et al.** *Large Concept Models: Language Modeling in a Sentence Representation Space.* s.l. : arXiv, 2024. 10.48550/arXiv.2412.08821.
48. **Hardinges, Jack.** What is a data trust? *The Open Data Institute.* [Online] The Open Data Institute, 10 July 2018. [Cited: 13 April 2025.] https://theodi.org.
49. **Zhao, Rui, et al.** *Libertas: privacy-preserving computation for decentralised personal data store.* s.l. : arXiv, 2023. 10.48550/arXiv.2309.16365.
50. **Department of Education, United Kingdom.** Pupil premium: overview . *GOV.UK.* [Online] 25 March 2025. [Cited: 13 April 2025.] https://www.gov.uk/government/publications/pupil-premium/pupil-premium.
51. *User-Managed Access to Web Resources.* **Machulak, Maciej P., et al.** Chicago, Illinois, USA : Association for Computing Machinery, 2010. 10.1145/1866855.1866865.
52. **W3C.** *Verifiable Credentials API.* s.l. : W3C, 2025.
53. *FaCT++ Description Logic Reasoner: System Description.* **Tsarkov, Dmitry and Horrocks, Ian.** Lecture Notes in Computer Science, Berlin, Heidelberg : Springer, Berlin, Heidelberg, 2006, Vol. 4130. 10.1007/11814771_26.
54. **halo2 contributors.** halo2. *The halo2 Book.* [Online] [Cited: 14 April 2025.] https://zcash.github.io/halo2/.
55. *Semantics and Complexity of SPARQL.* **Pérez, Jorge, Arenas, Marcelo and Gutierrez , Claudio.** Springer, Berlin, Heidelberg : Athens, GA, USA, 2006.
56. *The Even More Irresistible SROIQ.* **Horrocks, Ian, Kutz, Oliver and Sattler, Ulrike.** Lake District, UK : Proceedings of the KR, 2006.
57. *The berlin sparql benchmark.* **Bizer, Christian and Schultz, Andreas.** s.l. : IGI Global, 2009.
58. **Iannella, Renato, et al.** *ODRL Version 2.2 Ontology.* s.l. : W3C, 2017.
59. *Croissant: A metadata format for ml-ready datasets.* **Akhtar, Mubashara, et al.** Advances in Neural Information Processing Systems, Vancouver, Canada : Curran Associates, Inc., 2024, Vol. 37.
60. **W3C.** Web Security. *W3C.* [Online] W3C. [Cited: 14 April 2025.] https://www.w3.org/mission/security/#ping.
61. *Towards Computer-Using Personal Agents.* **Bonatti, Piero, et al.** s.l. : City University of London, 2025.
62. *EYE JS: A client-side reasoning engine supporting Notation3, RDF Surfaces and RDF Lingua.* **Wright, Jesse.** s.l. : CEUR WS, 2024.
63. **Heersmink, Richard, et al.** *A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness.* s.l. : Ethics and Information Technology, 2024. 10.1007/s10676-024-09777-3.
64. **Johnson, Sandra and Hyland-Wood, David.** *A Primer on Large Language Models and their Limitations.* s.l. : arXiv, 2024. 10.48550/arXiv.2412.04503.
65. **Carothers, Gavin.** *RDF 1.1 N-Quads: A line-based syntax for RDF datasets .* s.l. : W3C, 2014.
66. **Rosenthol, Leonard.** C2PA: the world's first industry standard for content provenance (Conference Presentation). [book auth.] SPIE. *Applications of Digital Image Processing XLV.* 2022.

67. **Bizer, Christian, Heath, Tom and Berners-Lee, Tim.** Linked data-the story so far. *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web.* s.l. : Manning publications co., 2023.
68. **Lasilla, Ora, Hendler, James and Berners-Lee, Tim.** *The Semantic Web.* s.l. : Scientific American, 2001.
69. **Bormann, Carsten and Hoffman, P.** *Concise Binary Object Representation (CBOR).* s.l. : IETF, 2020. RFC 8949.
70. **ISO: the International Organization for Standardization.** *Cards and security devices for personal identification — Building blocks for identity management via mobile devices - Part 1: Generic system architectures of mobile eID systems.* s.l. : ISO: the International Organization for Standardization, 2023. ISO/IEC 23220-1:2023.
71. **Terbu, O., Fett, D. and Campbell, B.** *SD-JWT-based Verifiable Credentials (SD-JWT VC).* s.l. : IETF, 2024.
72. **European Union.** *Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC.* 2014.
73. **European Commission.** *The European Digital Identity Wallet Architecture and Reference Framework.* s.l. : European Commission, 2023.
74. **Bernstein, Greg and Sporny, Manu.** *Data Integrity BBS Cryptosuites v1.0.* 2025.
75. **Looker, T., et al.** *The BBS Signature Scheme.* s.l. : Decentralised Identity Foundation, 2025.
76. **Department for Science, Innovation and Technolog.** Digital driving licence coming this year . *GOV.UK.* [Online] GOV.UK, 21 January 2025. [Cited: 13 April 2025.] https://www.gov.uk/government/news/digital-driving-licence-coming-this-year.
77. **Verhulst, Stefaan G.** *Operationalizing digital self-determination.* Cambridge : Cambridge University Press, 2023. 10.1017/dap.2023.11.
78. **Hummel, Patrik, et al.** *Data sovereignty: A review.* s.l. : SAGE Publications, 2021. 10.1177/2053951720982012.
79. **Friedman, Batya and Hendry, David G.** *Value Sensitive Design.* s.l. : The MIT Press, 2019. 9780262039536.
80. **Norman, Donald A.** *Design for a Better World: Meaningful, Sustainable, Humanity Centered.* s.l. : MIT Press, 2023. 978-0262047951.
81. **Ishmaev, Georgy, et al.** *Value Sensitive Design for Self-Sovereign Identity Solutions: Conceptual Investigation of uNLock Use Case.* s.l. : Springer Nature Link, 2023. 10.1007/s44206-023-00046-2.
82. *Succinct Non-Interactive Zero Knowledge for a von Neumann Architecture.* **Ben-Sasson, Eli, et al.** s.l. : USENIX Association, 2013.
83. *Circom: A circuit description language for building zero-knowledge applications.* **Bellés-Muñoz, Marta, et al.** s.l. : IEEE, 2022.
84. *PoneglyphDB: Efficient Non-interactive Zero-Knowledge Proofs for Arbitrary SQL-Query Verification.* **Gu, Binbin, Fang, Juncheng and Nawab, Faisal.** Berlin, Germany : Association for Computing Machinery, 2025. 10.1145/3709713.
85. *Ceno: Non-uniform, segment and parallel zero-knowledge virtual machine.* **Liu, Tianyi, et al.** s.l. : Springer, 2025.
86. **Roy, Uma.** Introducing SP1: A performant, 100% open-source, contributor-friendly zkVM. *Succinct: Research and Announcements.* [Online] 14 February 2024. [Cited: 14 April 2025.] https://blog.succinct.xyz/introducing-sp1/.

87. **powdr contributors.** powdr*. powdr.* [Online] [Cited: 14 April 2025.] https://docs.powdr.org.
88. **ZKM.** ZKM Architecture. *ZKM.* [Online] 21 February 2025. [Cited: 14 April 2025.] https://docs.zkm.io/zkm-architecture.
89. **Marius-Călin, Silaghi.** SMC: Secure Multiparty Computation Language. [Online] 25 November 2004. [Cited: 7 April 2025.] https://cs.fit.edu/~msilaghi/SMC/tutorial.html.
90. **Rastogi, Aseem, Swamy, Nikhil and Hicks, Michael.** *Wys⋆: A DSL for Verified Secure Multi-party Computations.*
91. *How to generate and exchange secrets.* **Yao, Andrew Chi-Chih.** Toronto, ON, Canada : IEEE, 1987. 10.1109/SFCS.1986.25.
92. *SMCQL: Secure Querying for Federated Databases.* **Bater, Johes, et al.** s.l. : VLDB Endowment, 2017. 10.14778/3055330.3055334.
93. *Conclave: secure multi-party computation on big data.* **Volgushev, Nikolaj, et al.** Dresden, Germany : Association for Computing Machinery, 2019. 10.1145/3302424.3303982.
94. *Senate: A Maliciously-Secure MPC Platform for Collaborative Analytics.* **Poddar, Rishabh, et al.** s.l. : USENIX Association, 2021. 978-1-939133-24-3.
95. *GOOSE: A Secure Framework for Graph Outsourcing and SPARQL Evaluation.* **Ciucanu, Radu and Lafourcade, Pascal.** s.l. : Springer, Cham, 2020. 10.1007/978-3-030-49669-2_20.
96. *SMPG: Secure Multi Party Computation on Graph Databases.* **Aljuaid, Nouf, Lisitsa, Alexei and Schewe, Sven.** s.l. : SCITEPRESS – Science and Technology Publications, 2022. 10.5220/0010876200003120.
97. *Graph database applications and concepts with Neo4j.* **Miller, Justin J.** s.l. : Association for Information Systems, 2013.
98. **Wright, Jesse.** Comparative Analysis of Digital Identity Architectures. *Jesse Wright.* [Online] 27 March 2025. [Cited: 10 April 2025.] https://blog.jeswr.org/genai-identity.html.
99. *Modeling and Verifying Security Protocols with the Applied Pi Calculus and ProVerif.* **Blanchet, Bruno.** s.l. : Foundations and Trends in Privacy and Security, 2016. 10.1561/3300000004.
100. *On the security of public key protocols.* **Dolev, Shlomi and Yao, Andrew Chi-Chih.** s.l. : IEEE, 1983. 10.1109/TIT.1983.1056650.
101. *Discovering concrete attacks on website authorization by formal analysis.* **Bansal, Chetan, et al.** s.l. : Journal of Computer Security, 2014. 10.3233/JCS-140503.
102. *A Formal Analysis of the FIDO UAF Protocol.* **Feng, Haonan, et al.** 2021. 10.14722/ndss.2021.24363.
103. *An Expressive Model for the Web Infrastructure: Definition and Application to the BrowserID SSO System.* **Fett, Daniel, Küsters, Ralf and Schmitz, Guido.** Berkeley, CA, USA : IEEE, 2014. 10.1109/SP.2014.49.
104. *A Comprehensive Formal Security Analysis of OAuth 2.0.* **Fett, Daniel, Küsters, Ralf and Schmitz, Guido.** Vienna, Austria : Association for Computing Machinery, 2016. 10.1145/2976749.2978385.
105. *The Web SSO Standard OpenID Connect: In-Depth Formal Security Analysis and Security Guidelines.* **Fett, Daniel, Küsters, Ralf and Schmitz, Guido.** Santa Barbara, CA, USA : IEEE, 2017. 10.1109/CSF.2017.20.
106. **Coburn, Aaron, Pavlik, elf and Zagidulin, Dmitri.** *Solid-OIDC.* s.l. : W3C, 2022.
107. **South, Tobin, et al.** *Authenticated Delegation and Authorized AI Agents.* s.l. : arxiv, 2025. 10.48550/arXiv.2501.09674.

108. **Richer, Justin and Imbault, Fabien.** *Grant Negotiation and Authorization Protocol (GNAP).* s.l. : IETF, 2024. RFC 9635.
109. **Singh, Munindar P.** *Information-Driven Interaction-Oriented Programming: BSPL, the Blindingly Simple Protocol Language.* Richland, SC : International Foundation for Autonomous Agents and Multiagent Systems, 2011. 10.5555/2031678.2031687.
110. **Chopra, Amit K., Christie V, Samuel H. and Singh, Munindar P.** *Interaction-Oriented Programming: An Application Semantics Approach for Engineering Decentralized Applications.* Virtual Event, Italy : Association for Computing Machinery, 2021. 10.1145/3465084.3467486.
111. **Wooldridge, Michael.** *An Introduction to Multiagent Systems.* s.l. : Wiley, 2009. 978-0470519462.
112. **Saad, Umair, et al.** *A model to measure QoE for virtual personal assistant.* s.l. : Springer Nature Link, 2016. 10.1007/s11042-016-3650-5.
113. **Searls, Doc.** *The Intention Economy: When Customers Take Charge.* USA : Harvard Business Press, 2012. 978-1422158524.
114. **McCarthy, J., et al.** *A proposal for the Dartmouth summer research project on artificial intelligence.* s.l. : AI Magazine, 1955.
115. **Franklin, Stan.** *Autonomous Agents as Embodied AI.* s.l. : Taylor & Francis, 2010. 10.1080/019697297126029.
116. **Zhi-Xuan, Tan, et al.** *Beyond Preferences in AI Alignment.* s.l. : arxiv, 2024. arXiv.2408.16984.
117. *Deep reinforcement learning from human preferences.* **Christiano, Paul, et al.** Long Beach, CA, USA : Curran Associates Inc., 2017. 10.5555/3294996.3295184.
118. *Fine-Tuning Language Models from Human Preferences.* **Zeigler, Daniel M., et al.** s.l. : arxiv, 2020. 10.48550/arXiv.1909.08593.
119. *Training language models to follow instructions with human feedback .* **Ouyang, Long, et al.** Red Hook, NY, USA : Curran Associates Inc., 2022. 10.5555/3600270.3602281.
120. *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.* **Bai, Yuntao, et al.** s.l. : arXiv, 2022. 10.48550/arXiv.2204.05862.
121. *Preference-satisfaction.* **Barber, Harriet E.** s.l. : Encyclopedia of Global Justice, 2011.
122. *Artificial intelligence: a modern approach.* **Russell, Stuart Jonathan and Norvig, Peter.**
123. *AI Alignment: A Comprehensive Survey.* **Ji, Jiaming, et al.** s.l. : arxiv, 2025. 10.48550/arXiv.2310.19852.
124. *Analysis of Smart Home Systems in the Context of the Internet of Things in Terms of Consumer Experience.* **Turkyilmaz, Serap and Altındağ, Erkut.** s.l. : International Review of Management and Marketing, 2022. 10.32479/irmm.12709.
125. *Trust, accountability, and autonomy in knowledge graph-based AI for self-determination.* **Ibáñez, Luis-Daniel, et al.** Dagstuhl, Germany : Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, 2023. 10.4230/TGDK.1.1.9.
126. **Wright, Jesse.** Data Governance in Data Wallets. *Jesse Wright.* [Online] 14 March 2025. [Cited: 9 April 2025.] https://blog.jeswr.org/2025/02/14/wallet-governance.
127. —. Communication in Multi Agent Systems. *Jesse Wright.* [Online] 17 March 2025. [Cited: 9 April 2025.] https://blog.jeswr.org/2025/05/17/mas-communication.
128. *Policy-Aware Content Reuse on the Web.* **Seneviratne, Oshani, Kagal, Lalana and Berners-Lee, Tim.** Berlin, Heidelberg : Springer Berlin Heidelberg, 2009. 978-3-642-04930-9.
129. **Cranor, Lorrie.** *Web privacy with P3P.* Sebastopol, California, USA : O'Reilly Media, Inc., 2002. 0-596-00371-4.

130. **Slabbinck, Wout.** *ODRL Evaluator.* [Software] 2025. https://doi.org/10.5281/zenodo.14265266.
131. *N3.js Reasoner: Implementing reasoning in N3.js.* **Wright, Jesse.** s.l. : CEUR WS, 2024.

Jesse Macleay Wright – DPhil Research Transfer of Status Report – Hilary 2025
22 April 2025