

# Semi-Autonomous Agents at Web Scale

Wright, Jesse<sup>1</sup>

<sup>1</sup>Computer Science Department, University of Oxford, UK

## Abstract

Large Language Models (LLMs) show potentials for a vast range of applications such as personal assistants, contract negotiators, and transport network coordinators. A key challenge is how we can establish trust in these applications.

Our work considers a broader class of software agents, which we term *semi-autonomous web agents*. These are software agents which individual or organisational entities entrust to autonomously carry out tasks on their behalf. However, when the agent does not have sufficient context or confidence to proceed working autonomously, the agent consults the user. This creates a user-agent dialogue that allows the user to teach the agent about the information sources they believe, their data sharing preferences, and their decision-making preferences. Ultimately, this enables the user to maximise control over their data and decisions whilst retaining the convenience of using agents.

Our research aims to develop near-term solutions for developing trustworthy and reliable semi-autonomous web agents. In particular, we aim to contribute to answering the question: “How do we build a network of semi-autonomous agents which represent individuals and organisations on the Web?”. To this end, we take a software architecture approach to define the functional, and non-functional requirements of a range of semi-autonomous web agents, specify architectures for these agents, and then define the requirements of a web protocol by which these agents would communicate.

This research so far has identified several key requirements. The communication protocol should support logically sound descriptions of (1) usage restrictions on exchanged data (2) data origins and provenance, and (3) transactional outcomes of dialogues. Using the architectural designs for agents, which conform to a protocol satisfying the above requirements - and currently make extensive use of the Semantic Web stack - we have methodically identified a range of research challenges which must be solved in order to implement components of the software architectures.

The first core project of our thesis aims to address the primary challenge identified: “How do we build a conceptual model of a users trust perception(s) for use by semi-autonomous agents?”.

## Keywords

Agent, Dialogue, LLM, Data Privacy, Trust, Semantic Web, Solid

## 1. Related Work

We begin with a discussion of agentic systems and then turn to our research into trust in such systems in Section 3. Communication protocols for multi-agent systems have decades of history in academia. Research into such systems at web scale is as old as the Semantic Web itself [1, 2, 3] which came with a vision of Charlie an “AI that works for you”. Yet, the 2006 lamentation that “[b]ecause we haven’t yet delivered large-scale, agent-based mediation, some commentators argue that the Semantic Web has failed to deliver” [4] still rings true today. More recently, the


---

*The 23rd International Semantic Web Conference, November 11–15, 2024, Hanover, MD*

✉ [jesse.wright@cs.ox.ac.uk](mailto:jesse.wright@cs.ox.ac.uk) (W. Jesse)

🌐 <https://www.cs.ox.ac.uk/people/jesse.wright/> (W. Jesse)

🆔 0000-0002-5771-988X (W. Jesse)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

LLM community has taken an interest in building multi-agent dialogues between LLMs [5, 6]. The LLM community identified that many of their open research challenges lie in building Trustworthy and Reliable Web Agents [7, 8]. Semantic Web research is already showing its strengths in complementing emergent LLM technologies. For instance, the use of Retrieval Augmented Generation (RAG) with Knowledge Graphs has become an effective and popular technique for grounding the responses of semi-autonomous agents [9]. We expect that a similar phenomenon will occur with LLM agents on the Web, provided our community can solve the challenges LLM researchers face today.

Thus, the context in which we develop multi-agent protocols is rapidly transforming. Advancements in LLM capabilities are compelling multi-agent communication protocols to accommodate more unstructured content than they have in the past. We also need to account for regulatory changes, as systems must ensure that the flow and analysis of data is compliant with the likes of the European Union General Data Protection Regulation (GDPR) and AI Act [10].

## 2. Use-Cases: Instances of Semi-Autonomous Agents

We first contextualise our specific research challenges by outlining the sample flow and functional requirements of a semi-autonomous agent we want to support. Running demos and flow-diagrams for these flows can be found at <https://linktr.ee/semiautoweb>. The agent is a generic personal assistant and the sample flow is scheduling a meeting based on a user prompt:

1. The user (Nigel) types into a chat “Please schedule a dinner with Jun during ISWC”.
2. If any of Nigel’s personal data needs to be used in a way that he has not already permitted, the agent requests the relevant permissions which Nigel can approve, deny or modify. Note that the agent should time any follow up questions at a time convenient for the user e.g. asking them a batch of questions whilst they are at the gym.
3. If Nigel’s agent cannot automatically determine whether data coming from Jun’s agent can be trusted, Nigel is prompted to answer questions to decide whether to trust Jun’s agent.
4. Depending on user preference Nigel is prompted to confirm a proposed meeting time before it is added to his calendar.
5. Once we also have agents representing organisations, the restaurant and transport can also be reserved as part of the flow, ensuring that (1) it is at a restaurant which has coeliac friendly and lactose free menu items, because Nigel has coeliac disease and Jun is lactose intolerant and (2) if the booking requires payment, users may be prompted before booking is confirmed, depending on their preferences.

The above use-cases aid in identifying the following non-functional requirements for a communication protocol between semi-autonomous web-agents:

1. Entities (individuals or organisations) must be *identifiable* on the Web.
2. It must be possible to deterministically discover the *agents* representing an *entity* from the *Web identity* of the *entity*.
3. It must be possible for agents to describe, and agree to, any usage controls associated with the data they are sending between one another.

4. It must be possible for agents to describe the origin and provenance of data they exchange.
5. It must be possible to *unambiguously* describe the outcomes of a communication that requires an agreement or transaction.
6. Serendipity: it must be possible for semi-autonomous agents to contextualise the task they are working on, such that the agents they are interacting with can introduce new solution spaces or actors to negotiate with.

As part of the thesis we plan to make these requirements more rigorous by (1) performing a requirements gathering engineering approach to gather an extensive set of non-functional requirements for personal semi-autonomous agents from users and (2) engaging with industry to gather the functional requirements for a range of specialised semi-autonomous agents for industry. This will enable us to refine the requirements for the communication protocol and the research challenges that we are addressing.

## 2.1. Sample Protocol Flow

1. A semi-autonomous agent is tasked to perform an action by the entity (person or organisation) for whom it acts.
2. That agent identifies:
  - a) the external entities with which it needs to converse (identified using WebIDs); and
  - b) the necessary information it needs to disclose to those entities.
3. The agent discovers the external agents with which it needs to converse using the WebID-Profiles of the entities established in step 2.
4. The agent negotiates with the other agents to establish terms of use for the data in anticipation of it being shared between them.
5. The agents then negotiate using RDF and unstructured data packaged with provenance and agreed-upon usage controls.
6. The dialogue completes with an agreed upon structured result.

## 3. Research Challenges

At the current stage in our research we have built a reference architecture and implementation of the generic personal assistant using a protocol that conforms to the functional requirements and allows for the sample flow outlined in Section 2. From this, we have identified the following research challenges associated with implementing the *generic personal assistant* architecture and the protocol with which it conforms.

### 3.1. Conceptual Models of Trust

#### 3.1.1. Challenge

Design conceptual (ontological) models of trust specialised to be used internally within semi-autonomous agents to enable:

|                                      | Data Integrity | Data Usage |
|--------------------------------------|----------------|------------|
| Risk (Internal Assumption Modelling) | <b>IDI</b>     | <b>IDU</b> |
| Reality (Exchanged Between Systems)  | <b>EDI</b>     | <b>EDU</b> |

**Figure 1:** The four categories of conceptual models that we expect a semi-autonomous agent to use in order to *believe* and *exchange* data.

1. A semi-autonomous agent to decide whether to *believe* (**IDI**, c.f. Figure 1) information in the context of a task it is trying to perform based on the provenance that is *available* (**EDI**) and the provenance it can obtain. This context should encompass the *type* of data that is being sent to the agent and the level of *tolerance* the agent has for the information being incorrect given the task being performed.
2. A semi-autonomous agent to decide whether to *disclose* (**IDU**) information in the context of a task it is trying to perform, based on applicable usage policies (**EDU**) and the probability of external parties complying with these policies.

### 3.1.2. Context

The challenge of trust modelling similarly arises from our proposed agent architectures. In particular, recall in step 2a of section 2.1 that we cannot have agents arbitrarily conversing with any service it finds on the Web, as there are malicious or unreliable services and entities with which we may disagree. For this reason, the agent has an internal model of (1) the trustworthiness of the claims made by other entities and (2) the probability of an agent's compliance with applicable usage policies.

### 3.1.3. Solution Space

We have reviewed the extensive research on theories of trust and trust modelling, including Trust Management for the Semantic Web [11], the Web Services Trust Ontology (WSTO) [12] and A Trust Ontology for Semantic Services [13]. The work most aligned to this use case is the Reference Ontology of Trust (ROT) [14]. However, the ROT, and all other conceptual models that we have reviewed, fail to support non-functional requirements for **IDI** conceptual modelling such as:

1. Qualifying how trust in claims from a given source is influenced by the content of the claims (e.g. I trust that airlines can make claims about flight times but not medical data).
2. Qualifying the strength of a proof that a given source made a claim, or that a given claim can be derived from a set of sources.

For **IDU** conceptual modelling the following non-functional requirements are unsupported:

1. Modelling the repercussions of broken trust on each party. For example, there may be no need to trust an entity if we can assume that their breach will be readily identified and financially compensated for.
2. Modelling external parties making poor judgements in risk assessments when forwarding to other parties.

Consequently, we are applying the NeOn [15] methodology to develop and evaluate ontologies for **IDI** and **IDU**; where the functional and non-functional requirements are derived from our agent architectures. Existing work into **EDU** conceptual modelling is substantive with a range of work [16, 10] including ODRL [17] garnering interest in decentralised data storage efforts such as Data Spaces [18] and Solid [19]. Thus we expect to be able to make use of these existing efforts. Work on **EDI** conceptual modelling is largely done at the lower level, such as ontologies for describing public/private key schemes [20, 21, 22, 23, 24], and the Verifiable Credential data model [25]. However, it lacks an upper level ontology to unify common concepts and allow for reasoning with **IDI** concepts. Thus, we are also applying the NeOn [15] methodology to develop and evaluate an upper level **EDI** ontology.

## 3.2. Instantiating Models of Trust

### 3.2.1. Challenge

There are two research challenges, with potentially conflicting requirements that need to be investigated here:

1. **User experience:** Design user-interaction patterns that enable users to *teach* agents about their evolving trust assumptions. To develop the UX, we plan to perform co-design workshops including participants for each of the *interfaces* identified in our functional requirements for agents - including voice-only, text-based chat and web-application experiences. Separate studies may be required for **IDI** and **IDU** models respectively.
2. **Mediation Engines:** Design *integrity* and *usage control engines* to mediate between the conceptual models of trust, and the provenance and usage-control data semi-autonomous agents send between one another. The closest existing work to this is the ODRL formal semantics specification [26] from which ODRL evaluation engines can be implemented. We are actively investigating and designing the *integrity* engine using eye-qa.

We aim to find a solution on the Pareto Frontier [27] of providing the best *UX* for instantiating trust models - and having the best *instance data* with which the *mediation engines* can operate. We invite feedback on how to quantify and evaluate this tradeoff.

### 3.2.2. Context

In step 4 of section 2.1 semi-autonomous agents need to establish (1) what data they are going to share with the external entities established in step 2 and (2) which usage control policies need to be applied to that data. This means that semi-autonomous agents need to negotiate **EDU** instance data, while internally referring to **IDU** instance data to establish which terms they are willing to agree to (this is the *mediation*), and bringing users in-the-loop when the **IDU** model is too incomplete (this is where the *UX* designs are needed). As we expect to use structured data for these **EDU** negotiations, all conditions are unambiguous and can be executed by rule engines. This means that non-malicious systems can provide guarantees that they will comply with the agreed terms of use for the shared data, and it is unambiguous as to who is at fault when terms of use are violated. This also aids in supporting GDPR compliance requirements

that that business agents will have as agents can ensure that policies they use are valid against a set of compliance criteria.

Similarly in step 5 of section 2.1 there is mediation between the **EDI** and **IDI** instance data, enabling a semi-autonomous agent to establish which incoming claims it can *believe*. Considering again the example where Nigel’s agent receives a set of claims from Jun, which it wants to use, this mediation engine has the following responsibilities:

1. To determine whether there is pre-existing trust in Jun’s agent for the given task.
2. In the absence of pre-existing trust, to seek provenance from negotiating agents to support Jun’s claims. For example, if Jun is endorsed by an authority as an expert on the topics she is making claims about, and Nigel trusts endorsements from that authority, then Jun’s agent providing the proof of endorsement is sufficient.
3. In the absence of sufficient provenance, to prompt the user to enhance their **IDI** model by adding trust permissions arising from the current task. For example, by sending a notification to a user’s phone saying: “Do you trust Jun Zhao (Univ. Oxford) (verified) to provide correct information on her calendar for the purpose of booking an event?”.

### 3.3. Additional Challenges & Future Work

The following subsection identifies additional challenges in designing architectures for semi-autonomous agents and the communication protocol that they use. Noting the extensive literature in these areas, they are not the focus of our research.

1. **Entity recognition and selection:** How do we identify the entities that users are directly or indirectly referring to in their messages, and how do we select which ones need to be involved in given negotiations? In the LLM-based semi-autonomous agent architectures that we have created (see step 2a of section 2.1), there is a need for the agent to perform entity recognition [28] to identify the Web Identities of entities that users are describing using natural language. For example, when an Nigel asks his agent to “schedule a meeting with Jun”, it needs to identify that Jun is the entity with the identifier <http://example.org/jun>. In our initial prototypes, we use Retrieval Augmented Generation (RAG) with the prompt and the set of WebId contents with which a user has explicitly defined trust relationships. For this research challenge, we should create a benchmark to evaluate the performance of agent’s architectures on a range of entity recognition tasks. This should result in a resource paper with the benchmark.
2. **Relevant data selection & information retrieval:** How do we select which subgraphs or views of user data are required for a given task? This is necessary to determine which data autonomous agents may disclose in step 2b of section 2.1.

## 4. Acknowledgements

This work is supervised by Sir Nigel Shadbolt and Dr Jun Zhao within the Ethical Web and Data Infrastructure in the Age of AI (EWADA) project. EWADA is led by Sir Nigel Shadbolt and Sir Tim Berners-Lee at the University of Oxford. The student is fully funded by the Department of Computer Science Scholarship.

## References

- [1] O. Lassila, J. Hendler, T. Berners-Lee, The semantic web, *Scientific American* 284 (2001) 34–43.
- [2] S. Luke, L. Spector, D. Rager, J. Hendler, Ontology-based web agents, in: *Proceedings of the first international conference on Autonomous agents*, 1997, pp. 59–66.
- [3] S. Poslad, Specifying protocols for multi-agent systems interaction, *ACM Transactions on Autonomous and Adaptive Systems (TAAS)* 2 (2007) 15–es.
- [4] N. Shadbolt, T. Berners-Lee, W. Hall, The semantic web revisited, *IEEE Intelligent Systems* 21 (2006) 96–101. doi:10.1109/MIS.2006.62.
- [5] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, C. Wang, Autogen: Enabling next-gen llm applications via multi-agent conversation framework, *arXiv preprint arXiv:2308.08155* (2023).
- [6] Y. Deng, W. Zhang, W. Lam, S.-K. Ng, T.-S. Chua, Plug-and-play policy planner for large language model powered dialogue agents, in: *The Twelfth International Conference on Learning Representations*, 2023.
- [7] Y. Deng, A. Zhang, Y. Lin, X. Chen, J.-R. Wen, T.-S. Chua, Large language model powered agents in the web, *learning* 2 (2024) 20.
- [8] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, et al., Trustllm: Trustworthiness in large language models, *arXiv preprint arXiv:2401.05561* (2024).
- [9] M. Kang, J. M. Kwak, J. Baek, S. J. Hwang, Knowledge graph-augmented language models for knowledge-grounded dialogue generation, *arXiv preprint arXiv:2305.18846* (2023).
- [10] B. Esteves, H. J. Pandit, V. Rodríguez-Doncel, Odrl profile for expressing consent through granular access control policies in solid, in: *2021 IEEE European Symposium on Security and Privacy Workshops (EuroSPW)*, 2021, pp. 298–306. doi:10.1109/EuroSPW54576.2021.00038.
- [11] M. Richardson, R. Agrawal, P. Domingos, Trust management for the semantic web, in: *International semantic Web conference*, Springer, 2003, pp. 351–368.
- [12] S. Galizia, Wsto: A classification-based ontology for managing trust in semantic web services, in: *European semantic web conference*, Springer, 2006, pp. 697–711.
- [13] W. Sherchan, S. Nepal, J. Hunklinger, A. Bouguettaya, A trust ontology for semantic services, in: *2010 IEEE International Conference on Services Computing*, IEEE, 2010, pp. 313–320.
- [14] G. Amaral, T. P. Sales, G. Guizzardi, D. Porello, Towards a reference ontology of trust, in: *On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019*, Rhodes, Greece, October 21–25, 2019, *Proceedings*, Springer, 2019, pp. 3–21.
- [15] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, *The NeOn Methodology for Ontology Engineering*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 9–34. URL: [https://doi.org/10.1007/978-3-642-24794-1\\_2](https://doi.org/10.1007/978-3-642-24794-1_2). doi:10.1007/978-3-642-24794-1\_2.
- [16] R. Zhao, J. Zhao, Perennial semantic data terms of use for decentralized web (2024).
- [17] R. Iannella, S. Villata, Odrl information model 2.2, 2023. URL: <https://www.w3.org/TR/2018/REC-odrl-model-20180215/>.

- [18] B. Otto, M. ten Hompel, S. Wrobel, *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*, Springer Nature, 2022.
- [19] A. V. Sambra, E. Mansour, S. Hawke, M. Zereba, N. Greco, A. Ghanem, D. Zagidulin, A. Abounaga, T. Berners-Lee, *Solid: a platform for decentralized social applications based on linked data*, MIT CSAIL & Qatar Computing Research Institute, Tech. Rep. (2016).
- [20] A. Kim, J. Luo, M. Kang, *Security ontology for annotating resources*, in: R. Meersman, Z. Tari (Eds.), *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 1483–1499.
- [21] H. Story, *The cert ontology specification*, 2008. URL: <https://www.w3.org/ns/auth/cert>.
- [22] V. Charpenay, M. McCool, *Web of things (wot) security ontology*, 2023. URL: <https://www.w3.org/2019/wot/security>.
- [23] J. Sabater, C. Sierra, *Review on computational trust and reputation models*, *Artificial Intelligence Review* 24 (2005) 33–60. URL: <https://doi.org/10.1007/s10462-004-0041-5>. doi:10.1007/s10462-004-0041-5.
- [24] *Wot schema*, 2005. URL: <http://xmlns.com/wot/0.1/>.
- [25] M. Sporny, D. Longley, D. Chadwick, O. Steele, *Verifiable credentials data model v2.0*, 2024. URL: <https://www.w3.org/TR/2024/CRD-vc-data-model-2.0-20240416/>.
- [26] N. Fornara, V. Rodríguez-Doncel, B. Esteves, S. Steyskal, B. W. Smith, *OdrI formal semantics*, 2024. URL: <https://w3c.github.io/odrl/formal-semantics/>.
- [27] A. V. Lotov, K. Miettinen, *Visualizing the Pareto Frontier*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 213–243. URL: [https://doi.org/10.1007/978-3-540-88908-3\\_9](https://doi.org/10.1007/978-3-540-88908-3_9). doi:10.1007/978-3-540-88908-3\_9.
- [28] B. Jehangir, S. Radhakrishnan, R. Agarwal, *A survey on named entity recognition—datasets, tools, and methodologies*, *Natural Language Processing Journal* 3 (2023) 100017.