Here's Charlie! Realising the Semantic Web vision of Agents in the age of LLMs

Wright, Jesse[1]

[1]*Computer Science Department, University of Oxford, UK*

Abstract

This paper presents our research towards a near-term future in which *legal entities*, such as *individuals* and *organisations* can entrust semi-autonomous AI-driven agents to carry out online interactions on their behalf. The author's research concerns the development of semi-autonomous Web agents, which consult users if and only if the system does not have sufficient context or confidence to proceed working autonomously. This creates a user-agent dialogue that allows the user to teach the agent about the information sources they trust, their data-sharing preferences, and their decision-making preferences. Ultimately, this enables the user to maximise control over their data and decisions while retaining the convenience of using agents, including those driven by LLMs.

In view of developing near-term solutions, the research seeks to answer the question: "How do we build a trustworthy and reliable network of semi-autonomous agents which represent individuals and organisations on the Web?". After identifying key requirements, the paper presents a demo for a sample use case of a generic personal assistant. This is implemented using (Notation3) rules to enforce safety guarantees around *belief*, *data sharing* and *data usage* and LLMs to allow natural language interaction with users and *serendipitous* dialogues between software agents.

Keywords

Agent, Dialogue, LLM, Data Privacy, Trust, Semantic Web, Solid Reasoner, Inference, RDF, N3, Notation3, RDF Surfaces, Semantic Web, Proof, Proof Engine, Solid

1. Introduction

There exists a substantial body of research on communication protocols for multi-agent systems, and it is reflected in the vision of the Semantic Web itself [1, 2, 3] as shown by Charlie, the "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" [4] still rings true today. The growing use of LLMs raises a key challenge in building Trustworthy and Reliable Web Agents [5, 6]. This is heightened by growing interest among LLM researchers in building dialogues between multiple LLMs [7, 8]. Moreover, recent research indicates the strong potential of the Semantic Web to complement emerging LLM technologies [9]. For example, the use of Retrieval Augmented Generation (RAG) with Knowledge Graphs has shown to be effective in grounding LLM queries [10]. The universal semantics and proof mechanisms of the Semantic Web stack are therefore pertinent to the successful development of semi-autonomous Web agents using LLMs.

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

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)

2. Design Requirements

We identify the following non-functional requirements for an agent communication protocol. It must be possible for semi-autonomous agents to:

1. *Identify* legal entities, such as individuals or organisations, on the Web [11] so they can be referenced.
2. *Deterministically discover* other agents representing an entity from their Web identity [11]. This does *not* require all agents to be publicly advertised; some may be discovered from links to protected documents.
3. Describe, and agree to, any *usage controls* [12, 13, 14] associated with data they exchange. This allows sharing of protected data while articulating the recipient's legal or moral obligations [15].
4. Describe the *origin* and *provenance* of data they exchange. In an open world of agents that can "say anything about anything," systems can identify which external claims to believe for a given task, based on the agent's internal trust model.
5. *Unambiguously* describe *ground truths* they send, and *agreements* they make, using a formal representation. Consider the case where an individual's agent purchases a flight from an airline's agent. Structured ground truths eliminate an LLM's risk of hallucination or misinterpretation of key information, such as the flight time ("10 o'clock" could be 22:00 or 10:00). As agents represent entities in binding agreements, this approach also reduces the risk of legal disputes by limiting the subjectivity of agreed terms and thus the ability to reinterpret or rescind them [16]. Furthermore, agents can implement rule-based internal safeguards, such as user-defined daily spending limits. Truly generic agents may generate and communicate structured ontologies when encountering new tasks. In many cases we expect LLM-supported ontology construction [17] to facilitate generation; however, research is required to understand how (1) agents can align on conceptual models for use and (2) how human oversight can be maintained without disrupting user experience.
6. Contextualise a task which may be *ambiguous* or poorly defined, such that interacting agents can introduce new solution spaces or negotiating actors in a *serendipitous* manner.

3. Sample Use-Case and Implementation

We implemented the following flow where agents act as personal assistants for individual users:

1. Jun types into a chat "Schedule a meeting with Nigel next week";
2. Jun's agent identifies data to be shared with Nigel and requests relevant sharing permissions from Jun (where not already obtained);
3. Nigel's agent receives a request from Jun;
4. Nigel is prompted to confirm that he believes Jun is an authoritative source of truth for her calendar (where not already obtained);
5. Nigels agent proposes a meeting time to Nigel; and
6. the meeting is proposed to Jun's agent and automatically confirmed.

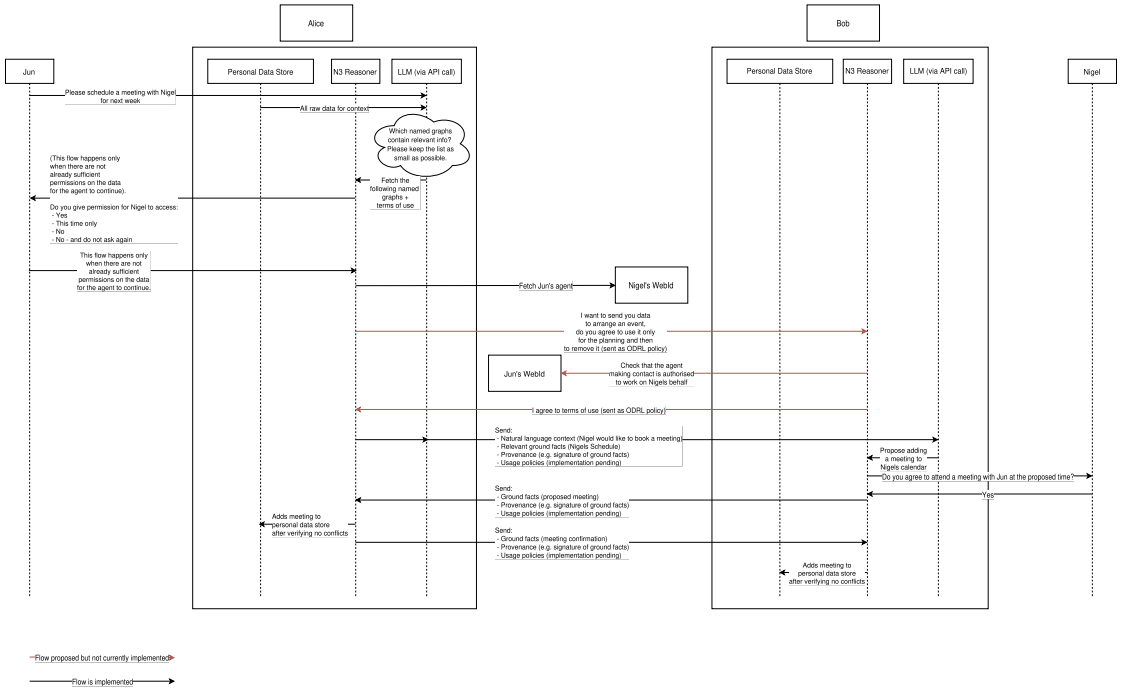

Figure 1: Flow diagram for the scheduling use case. Alice is Jun's agent and Bob is Nigel's agent.

We have created a running demo with a video, flow-diagrams (including Figure 1) and other resources for our codebase[1]. The implementation corresponds to the above use-case steps:

1. Given the user prompt and a set of known WebID profiles [11], an LLM called by Jun's agent identifies the relevant entities for the agent to negotiate with (Nigel), and the WebIDs of those entities. Given the user prompt, and the user's personal knowledge graph, an LLM called by Jun's agent identifies which subset (as a list of named graphs) of the user data are needed to fulfil the user's request.
2. Notation3 [18] reasoning is used to identify the policies applicable to the data subset. In the available demo recording, policies are encoded in ACP [12]; we are currently migrating to use ODRL [13] and DPV [14]. If these policies do not yet permit read access to Nigel, Jun is prompted to modify them. Jun's agent then dereferences Nigel's WebID [11] to discover information about his agent.
3. Jun's agent uses an LLM to construct a message for Nigel's agent, explaining the context of Jun's task: "Jun seeks to schedule a meeting for next week. Propose a time for Jun and Nigel to meet using their calendars." Jun's agent sends Nigel's agent this message along with the RDF description of Jun's calendar and any associated policies and provenance. With ACL, Nigel's agent does not need to agree to any policy obligations; this changes with ODRL. The provenance in this case is simply a signature of the canonicalised calendar dataset [19] using Jun's public key.

[1] https://github.com/jeswr/phd-language-dialogue-experiment

4. As Nigel has instructed his agent that Jun is an authoritative source of information on all topics, his agent *believes* (takes as ground truth) the signed RDF dataset sent by her agent. We are developing conceptual models for agentic trust; these extend existing trust vocabularies [20, 21, 22, 23] with a range of features including (1) qualifying whether sources are trusted for particular *types* of claims; for instance, most agents should trust certified airlines to present flight times and prices, but not medical data (2) qualifying the forms of provenance *secure* enough for a given task; for instance, an insurance provider may require provenance demonstrating a user was signed in with two-factor authentication when entering financial details to their knowledge base.
5. Nigel's agent proposes a meeting time, using the natural language context (*not* a ground truth) and the calendar dataset (ground truth). The LLM proposes a meeting time, then the N3 reasoner applies rules to (1) ensure no calendar conflicts and (2) check for user confirmation, before adding the proposed time to the knowledge base. In a future iteration, we plan to use the LLM to generate an N3 query that proposes a meeting time based on Nigel's Personal Data Store and Jun's calendar.
6. Upon meeting the above requirements, the reasoner sends to Jun's agent a meeting proposal, in the form of an RDF dataset with attached usage policies and provenance. Jun's agent confirms this dataset can be believed based on the internal trust model. The rules within Jun's agent validate that there are no conflicting events. Jun's personal knowledge base is updated with the event, and a confirmation is sent to Nigel's agent.

4. Conclusion and Future Research

We have implemented a generic personal assistant that communicates using a protocol satisfying the requirements of Section 2. Future work will make the design requirements more rigorous by (1) gathering requirements for personal agents through user studies, and (2) engaging with industry to develop specialised agents, including product sales agents. Concurrently, we shall formalise the vocabularies for exchanging *provenance* and *terms of use* between agents and modelling *trust* and *data policies* within agents, extending those vocabularies discussed in Section 3. Once these vocabularies mature, we will develop reasoning specifications to mediate between the internal representations and exchanged metadata. This enables agents to negotiate to obtain sufficient provenance to believe claims, and find agreeable data terms of use between agents - whilst concurrently updating their internal models via user interaction.

Acknowledgements

Jesse Wright is funded by the Department of Computer Science, University of Oxford.

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] 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.

[6] 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).

[7] 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).

[8] 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.

[9] J. Wright, The old and the new - using semantic web technologies to build better AI, 2024. URL: https://blog.jeswr.org/2024/04/18/better-ai.

[10] 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).

[11] H. Story, T. Berners-Lee, A. Sambra, R. Taelman, J. Scazzosi, Web Identity (WebID) 1.0, W3C Community Group Final Report, W3C, 2024. https://w3c.github.io/WebID/spec/identity/.

[12] M. Bosquet, Access Control Policy (ACP), Solid Editor's Draft, W3C, 2022. https://w3c.github.io/WebID/spec/identity/.

[13] R. Iannella, S. Villata, Odrl information model 2.2, 2023. URL: https://www.w3.org/TR/2018/REC-odrl-model-20180215/.

[14] H. J. Pandit, A. Polleres, B. Bos, R. Brennan, B. Bruegger, F. J. Ekaputra, J. D. Fernández, R. G. Hamed, E. Kiesling, M. Lizar, et al., Creating a vocabulary for data privacy: The first-year report of data privacy vocabularies and controls community group (dpvcg), 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. 714–730.

[15] J. Wright, B. Esteves, R. Zhao, Me want cookie! towards automated and transparent data governance on the web, 2024. URL: https://arxiv.org/abs/2408.09071. arXiv:2408.09071.

[16] M. Garcia, What Air Canada Lost In 'Remarkable' Lying AI Chatbot Case, https://www.forbes.com/sites/marisagarcia/2024/02/19/what-air-canada-lost-in-remarkable-lying-ai-chatbot-case/, 2024. [Accessed 05-07-2024].

[17] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An llm

supported approach to ontology and knowledge graph construction, 2024. URL: https://arxiv.org/abs/2403.08345. arXiv:2403.08345.
[18] T. Berners-Lee, Notation3, http://www.w3.org/DesignIssues/Notation3.html (1998).
[19] D. Longley, G. Kellogg, D. Yamamoto, M. Sporny, Access Control Policy (ACP), Solid Editor's Draft, W3C, 2022. https://w3c.github.io/WebID/spec/identity/.
[20] M. Richardson, R. Agrawal, P. Domingos, Trust management for the semantic web, in: International semantic Web conference, Springer, 2003, pp. 351–368.
[21] S. Galizia, Wsto: A classification-based ontology for managing trust in semantic web services, in: European semantic web conference, Springer, 2006, pp. 697–711.
[22] 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.
[23] 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.