Report from Dagstuhl Seminar 25051

Trust and Accountability in Knowledge Graph-Based AI for Self Determination

John Domingue[*1], Luis-Daniel Ibáñez[*2], Sabrina Kirrane[*3], Maria-Esther Vidal[*4], and Philipp D. Rohde[†5]

1 **The Open University – Milton Keynes, GB.** john.domingue@open.ac.uk
2 **University of Southampton, GB.** L.D.Ibanez@soton.ac.uk
3 **Vienna University of Economics and Business, AT.** sabrina.kirrane@wu.ac.at
4 **TIB – Hannover, DE.** vidal@l3s.de
5 **TIB – Hannover, DE.** philipp.rohde@tib.eu

Abstract

This report documents the program and results of the Dagstuhl Seminar 25051 "Trust and Accountability in Knowledge Graph-Based AI for Self Determination". The seminar focused on AI systems powered by Knowledge Graphs and their fundamental role in powering intelligent decision making. Knowledge Graphs complement Machine Learning algorithms by providing data context and semantics, enabling further inference and question answering capabilities, and their synergy with Large Language Models is being actively researched. Despite the numerous benefits that can be accomplished with KG-based AI, its growing ubiquity within online services may raise the loss of self-determination for citizens as a fundamental societal issue. The more we rely on these technologies, which are often centralised, the less citizens will be able to determine their own destiny. To counter this threat, AI regulation, such as the EU AI Act, is being proposed in certain regions. Regulation sets what technologists need to do, leading to questions concerning: How can the output of AI systems be trusted? What is needed to ensure that the data fueling and the inner workings of these artefacts are transparent? How can AI be made accountable for its decision-making?

Seminar January 26–31, 2025 – https://www.dagstuhl.de/25051

2012 ACM Subject Classification Information systems → Decision support systems; Theory of computation → Semantics and reasoning; Information systems → Graph-based database models; Computing methodologies → Artificial intelligence

Keywords and phrases access control and privacy, federated query processing, intelligent knowledge graph management, programming paradigms for knowledge graphs, semantic data integration

Digital Object Identifier 10.4230/DagRep.15.1.136

[*] Editor / Organizer
[†] Editorial Assistant / Collector

Trust and Accountability in Knowledge Graph-Based AI for Self Determination, *Dagstuhl Reports*, Vol. 15, Issue 1, pp. 136–200
Editors: John Domingue, Luis-Daniel Ibáñez, Sabrina Kirrane, Maria-Esther Vidal, and Philipp D. Rohde

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

John Domingue (The Open University – Milton Keynes, GB, john.domingue@open.ac.uk)
Luis-Daniel Ibáñez (University of Southampton, GB, L.D.Ibanez@soton.ac.uk)
Sabrina Kirrane (Vienna University of Economics and Business, AT, sabrina.kirrane@wu.ac.at)
Maria-Esther Vidal (TIB – Hannover, DE, vidal@l3s.de)

In just one minute in April 2022, there were 5,900,000 searches on Google, 1,700,000 pieces of content shared on Facebook, 1,000,000 hours streamed and 347,200 tweets shared on Twitter.[1] This content and data is linked to a plethora of AI services which have increasingly been based on Knowledge Graphs (KGs), i.e., machine-readable data and schema representations based on a web stack of standards. The term "Knowledge Graph" was first introduced by Google in 2012 and is strongly linked to the work of the Semantic Web community which first began in around 2001 based on the seminal paper Berners-Lee et al.[2]. The main types of areas covered by AI services include, for example, content recommendation, user input prediction, and large-scale search and discovery and form the basis for the business models of companies like Google, Netflix, Spotify, and Facebook.

Over a number of years, there has been a growing worry on how personal data can be abused and thus, how AI services impinge on citizens' rights. For example, the over centralisation of data and linked abuses led Sir Tim Berners-Lee to call the Web "anti-human" in an interview in 2018[3] and since 2016, hundreds of US Immigration and Customs Enforcement employees have faced investigations into abuse of confidential law enforcement databases including stalking and harassment, up to passing data to criminals.[4]

The subject of proposed legislation today is ensuring that digital platforms, including AI platforms, provide societal benefit. Within Europe, the proposed EU AI Act aims to support safe AI that respects fundamental human rights. The regulation sets what technologists need to do. Our seminar was structured around three pillar research topics – trust, accountability, and self-determination – that represent the desired goals, and four foundational research topics – Machine-readable Norms and Policies, Decentralised KG Management, Neuro-Symbolic AI, and Decentralised Applications – that constitute the necessary technical foundations to achieve the goals.

[1] https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/
[2] Berners-Lee, T., Hendler, J. and Lassila, O., The semantic web. Scientific American, 284(5), pp. 34-43.
[3] "I Was Devastated": Tim Berners-Lee, the Man Who Created the World Wide Web, Has Some Regrets | Vanity Fair
[4] https://www.wired.com/story/ice-agent-database-abuse-records/

2 Table of Contents

3 Invited Talks

3.1 Why are we still not there (The Semantic Web today)

James A. Hendler (Rensselaer Polytechnic Institute – Troy, US)

In 2001, Tim Berners-Lee, Ora Lassila and I wrote a vision paper in Scientific American outlining potential challenges on the Web and exploring how ontologies and (what are now called) knowledge graphs could improve interoperability among Web systems and allow agents on the Web to cooperatively solve problems. The paper opened with a scenario of two people coordinating their parent's medical treatment and involved a number of different web sites, but a relatively straightforward task. In this talk, I reminded people of that challenge, and pointed out that despite all that has happened in AI, in Semantic Web, and in "agentic systems", we still cannot do the simple task that was outlined in that paper nearly 25 years ago. This talk discussed why, and what we might think about going forward to solve such problems. The Semantic Web paper grew in part out of a 2000 Dagstuhl Seminar (00121) and led to a number of later meetings that have continued on the theme. The 2001 paper is the most academically cited paper in the history of the Scientific American publication, and a number of Dagstuhl meetings in the years since have focused on Semantic Web and related topics. Some of the most cited Dagstuhl Seminars have appeared in this series (such as 24061) and this most recent seminar was a wonderful opportunity to continue the series – so this talk also acknowledged the huge role Dagstuhl has had in creating and sustaining this community over the past 25 years.

4 4-Minute Talks

4.1 Neuro-symbolic, agentic AI for organizing scholarly knowledge

Sören Auer (TIB – Hannover, DE)

The exponential growth of scientific publications poses a significant challenge for researchers, policymakers, and automated systems alike: how to effectively access, structure, and reason over ever-expanding bodies of knowledge. We aim to leverage neuro-symbolic and agentic AI to transform the organization and consumption of scholarly knowledge. At the core of this vision is the Open Research Knowledge Graph (ORKG) – an infrastructure designed to capture, interlink, and semantically represent the core contributions of scientific articles. Unlike traditional bibliographic databases, the ORKG enables structured comparisons of research contributions, supports semantic search, and integrates knowledge across disciplines. Building on this foundation, ORKG ASK provides an intelligent assistant that can answer complex research questions by composing and contextualizing information across the graph, significantly enhancing discoverability and scientific insight. We explore how neuro-symbolic methods – combining deep learning with semantic representations over knowledge graphs – can be further augmented by agentic AI. These agents autonomously navigate, enrich, and reconcile scholarly knowledge by performing tasks such as hypothesis mapping, claim validation, and comparison synthesis. This synergy enables a new generation of interactive and explainable AI systems that actively support scientific discovery and meta-research.

4.2 Pillar #1: Meanings of Trust

Piero A. Bonatti (University of Naples, IT)

Trust in Knowledge Graphs needs to be supported with a wide range of methodologies and technologies that address complementary issues. First, KG are incresingly being used to encode confidential information and personal data that shall be appropriately protected; more generally, the use of such data shall be restricted, even after data disclosure. Second, KG contents – that is provided by manifold "agents" (both humans and AI) using diverse knowledge sources – should be reliable, and their integrity should be protected. The inferences that can be drawn from KG should be reliable as well. And the people in charge of maintaining the KG with all of its sensitive information should be trustable, too.

A closer look at the above points reveals several connections with the other pillars mentioned in this seminar's manifesto, for example:

- Accountability, like trust in KG content and KG management, may leverage logging, provenance, integrity preserving and non-repudiation techniques.
- Self determination involves – among other aspects – the protection of personal data and control on its usage.

Some of the techniques that may help in improving trust in KG, along the above lines, are well-established (such as those for integrity preservation and non-repudiation). Some challenges (such as making AI more reliable and explainable to trust its inferences, or enforcing usage restrictions after data disclosure) are probably not specific to KG. Thus, the list of strictly trust-related challenges is not obvious and needs to be carefully drawn.

For sure, access control to, and the anonymization of, KG are harder than their traditional counterparts for relational databases, both because there is no reference schema to rely upon, and because the many inference tools that operate on KG may reveal concealed confidential data. Accordingly, the intrinsic computational complexity of access control and anonymization is often harder than in the classic case, which poses an obvious challenge.

The research on access control and anonymization techniques for KG and knowledge bases does not always take into account the large body of experience developed in the area of computer security and privacy [1]. This is a major risk that calls for a more extensive exploration of security and privacy papers. Moreover, some machine-readable policy languages that are becoming increasingly popular in the KG world, lack formal semantics. Consequently, they are ambiguous and under-specified in several respects. This is likely to jeopardize trust in many natural distributed scenarios, where different parties and stakeholders should understand policies in the same way, in order to avoid violations and sanctions.

References

1 Piero A Bonatti. A false sense of security. *Artificial Intelligence*, 310:103741, 2022.

4.3 Challenging Scenarios for Trust and Accountability in KG-based AI

Irene Celino (CEFRIEL – Milan, IT)

While Knowledge Graphs and other AI technologies provide an irrefutable advantage in managing data and knowledge in a meaningful way, several application scenarios present difficult challenges in relation to the management of trust and accountability between people and between organizations. In my short talk, I presented three different scenarios coming from my experience with their specific issues.

In the context of *cognitive-behavioural therapy*, psychological/psychiatric patients compile the so-called cognitive diaries, i.e. stories about personal events, aimed to help them reflect on emotions, thoughts, feelings and behaviours. Those diaries are shared with therapists, to enable early identification of signals that could lead to psychotic episodes. A digital version of cognitive diaries can support their processing to help therapist to promptly intervene. In my experience of digitisation of cognitive diaries[5], several issues emerged:

(1) Do patients trust a system to collect their diaries? Who has the legal right to access them (e.g. legal guardians)?

(2) Cognitive diaries may contain mentions of real/imagined events involving real/imaginary people: how to deal with those reported "facts"? They are not the same as misinformation?

(3) How to anonymise diary content (mostly textual) in a responsible way? How to share anonymised content with the scientific community without adding specific context (which may be required to correctly understand)?

In the context of *industrial procedures*, I'm currently working[6] on enabling the construction of procedural knowledge graphs, to collect employees' knowledge about industrial procedures (i.e. how-to), often based on experience, possibly including tacit knowledge. The goal is to provide the industry workers with KG-powered AI tools [1] to support their compliance with the procedures and to reduce the possible mistakes. In this case, emerging challenges are:

(1) Are workers willing to share their procedural knowledge? Do they fear losing their professional value? Do they fear being monitored?

(2) Do AI tools always ensure accuracy and reliability when providing information to industry workers? Are they safe?

(3) Are industrial employees scared of being replaced by AI? What kind of human knowledge and abilities should be preserved as such (e.g. critical thinking)?

In the context of *mobility data spaces*[7], an ecosystem of mobility actors (service providers, authorities, etc.) may be willing to cooperate (e.g. Mobility-as-a-Service), but in the meantime they are in competition and they want to preserve their competitive advantage. Therefore, they need to regulate data and service sharing in a distributed, federated (e.g. European National Access Points) and possibly "untrusted" business environment. Apart from the research challenges specifically related to the data space topic, other issue may emerge: (1) What if there are multiple, incompatible data spaces? How can a mobility actor avoid duplication of work to connect to different digital ecosystems (especially when they

[5] DIPPS project, co-funded by the Italian Ministry of Enterprises and Made in Italy
[6] PERKS project, co-funded under the EU Horizon Europe Programme (`https://perks-project.eu/`)
[7] deployEMDS project, co-funded under the EU Digital Europe Programme (`https://deployemds.eu/`)

already entered some)? (2) Are mobility actors willing to share business-critical information (even if required by laws)? Can they preserve their right to retain their information and beliefs when entering a negotiation (e.g. non disclosure of disagreement, difference between ontological agreement and ontological commitment [2])?

References

1 Irene Celino, Valentina A. Carriero, Andrea Azzini, Ilaria Baroni, and Matteo Scrocca. Procedural knowledge management in industry 5.0: Challenges and opportunities for knowledge graphs. *Journal of Web Semantics*, 84:100850, 2025.
2 Emanuele Della Valle, Irene Celino, and Davide Cerizza. Agreeing while disagreeing, a best practice for business ontology development. In *Proceedings of the 11th International Conference on Business Information Systems*. Springer, 2008.

4.4 Policies in decentralised ecosystems and academic regulatory frameworks

Andrea Cimmino (Polytechnic University of Madrid, ES)

Knowledge graphs and decentralised data has become one of the pillars of the European data-driven infrastructures like data spaces or proposals to foster data sovereignty like SOLID Pods. In this context, specifying the circumstances under which data should be accessed has become a crucial task. On the one hand, one challenge is specifying in an unambiguous way the terms, conditions, and actions that constitute a policy under which such data can be exploited in a machine-readable format. On the other hand, evaluating and enforcing policies to check whether the usage of such data legit. In this context, my personal research interests go in two lines. The first, related to the former challenge, is to express different academic regulatory frameworks as policies or norms to analyse their interoperability, quality, or the conformance of physical lessons to such regulations. The second, related to the latter challenge, is to tackle challenges derived from the evaluation and enforcement of policies in a decentralised ecosystem. In addition, it is worth researching the usage of LLMs in the different steps of the life-cycle of policy management.

4.5 Improving trustworthiness of ML through evaluation and formal guarantees

Michael Cochez (VU Amsterdam, NL)

My research focuses on neuro-symbolic AI, particularly with knowledge graphs (KG). I investigate how graph neural networks can bridge the gap between discrete KGs and the statistical world of machine learning, enabling more robust and scalable solutions for tasks like structured and natural language question answering. My work often utilizes data from the medical and biomedical domains.

During the seminar, I hope to discuss two key challenges:

Improving ML Trustworthiness: In my view current evaluation metrics are inadequate, leading to overconfidence in models. Among possible solutions, I want to discuss alternative evaluation methods and the need for formal guarantees (e.g., error bounds) to enhance trust in graph ML models.

Addressing Bias in Multi-source KGs: I will discuss the challenges of learning on unbalanced KGs, where a larger, potentially biased graph overshadows smaller, more specific ones. This hinders the use of graph ML in decentralized settings where self-determination is crucial.

4.6 Decentralisation the key for Ensuring Trust, Privacy and Personalisation in KG-based AI systems

John Domingue (The Open University – Milton Keynes, GB)

The explosion in interest and take-up of AI based systems has grown enormously since the arrival of GenAI through ChatGPT at the end of 2022. At the Open University we have been exploring how machine learning and AI can aid our 200K+ students since 2011 when we created OU Analyse[8] – a learning analytics tool able to predict if a student is at risk of failing the next assignment or course overall with 95% accuracy in the best cases. In 2015 we began investigating how decentralising technologies, such as distributed ledgers and personal data stores, such as Solid, could enable students to be "Self Sovereign" with respect to their credentials[9]. Since the beginning of 2023 we have been developing and evaluating two GenAI based tools to support teaching and learning at the OU. Our AI Module Writing Assistant (AIMWA) [1] supports OU academics in the writing of new courses. A pilot is currently underway with a module writing team in the Faculty of Business and Law who are creating a new MBA course. The AI Digital Assistant (AIDA) [2] serves as a helper to students and is able to answer questions on course materials, generate quizzes and new learning activities and re-write the materials themselves to suit student need.

We are now beginning to bring the above together, building on [3] (also see subsection 5.2) and Tim Berners-Lee's note on 'Charlie Works'[10] to create a Lifelong Learning Coach. In particular combining GenAI, Personal Knowledge Graphs and Solid pods so that we can utilise personal student information to hyper-personalise the feedback that AIDA gives whilst preserving privacy.

References

1 Alexander Mikroyannidis, Nirwan Sharma, Audrey Ekuban, and John Domingue. Using generative ai and chatgpt for improving the production of distance learning materials. In *2024 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 188–192. IEEE, 2024.

2 Bart Rienties, John Domingue, Subby Duttaroy, Christothea Herodotou, Felipe Tessarolo, and Denise Whitelock. What distance learning students want from an ai digital assistant. *Distance Education*, pages 1–17, 2024.

[8] https://analyse.kmi.open.ac.uk/
[9] https://blockchain.open.ac.uk/
[10] https://www.w3.org/DesignIssues/Works.html

3 Piero A. Bonatti, John Domingue, Anna Lisa Gentile, Andreas Harth, Olaf Hartig, Aidan Hogan, Katja Hose, Ernesto Jimenez-Ruiz, Deborah L. McGuinness, Chang Sun, Ruben Verborgh, and Jesse Wright. Towards computer-using personal agents, 2025.

4.7 Robust explanations with Neurosymbolic AI

Michel Dumontier (Maastricht University, NL)

Language Models (LLMs) have created new opportunities to build trustworthy, transparent, and accountable AI systems. However, key challenges remain in ensuring that AI outputs can be trusted, explaining how decisions are reached, and integrating diverse sources of structured and unstructured knowledge. In our group, we explore how neurosymbolic agents that combine symbolic reasoning with machine learning techniques can be used with knowledge graphs and scientific literature to predict and explain unknown biomedical phenomena. Several projects in the group are focused on trustworthy generative AI. Our GENIUS Lab for Trustworthy Generative AI fosters academic–industrial collaboration to develop generative AI systems capable of supporting expert decision-making. Neuro-symbolic methods are used to produce interpretable reasoning capabilities, with a focus on creating open-source frameworks for conversational AI, leveraging FAIR (Findable, Accessible, Interoperable, Reusable) data and services. Complementary projects, such as REALM and CHARM, highlight collaborative efforts to harness data standards, blockchain technology, explainable AI, and advanced data management for healthcare applications. The overarching goal is to develop AI-driven solutions that not only yield accurate predictions but also provide transparent, scientifically grounded justifications. By integrating human expertise with AI-based reasoning, this research seeks to enhance the reliability, scalability, and accountability of AI systems across diverse, real-world domains.

4.8 Trustworthy Engineering of Neurosymbolic AI Systems

Fajar Ekaputra (Vienna University of Economics and Business, AT)

The rapid evolution of Neurosymbolic AI systems – particularly those that combine Knowledge Graphs (KGs) with machine learning – has opened a plethora of new possibilities for future development of AI systems. However, as these hybrid systems become more complex, they also present significant challenges. One of the most pressing concerns is the lack of standardized system representation for designing, engineering, and documenting such systems. This issue hampers the systematic characterization of these complex architectures, making them harder to analyze, compare, and trust.

To address these challenges, frameworks like boxology notation [1] have been proposed to visually simplify the representation of complex AI systems. By offering a clearer view of how different components of such systems interact, these approaches aim to improve understanding and foster greater trust. However, current solutions primarily focus on post-hoc analysis rather than geared towards supporting the entire engineering process.

In this seminar, I hope to explore ways to extend the existing AI system representations to make them more beneficial throughout the AI system development life cycle. We aim to support representation of diverse perspectives through a pattern-based engineering approach [2] – from interdisciplinary collaboration to the needs of various stakeholders and engineering processes. We also discuss how this approach could potentially enhance AI system auditability, particularly in the context of regulatory frameworks such as the EU AI Act.

References

1 Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. Modular design patterns for hybrid learning and reasoning systems. *Applied Intelligence*, 51:6528–6546, 2021.
2 Fajar J. Ekaputra. Pattern-based engineering of neurosymbolic ai systems. *Journal of Web Semantics*, 85:100855, 2025.

4.9 Using Semantic Web Technologies for Reasoning about Policies

Nicoletta Fornara (University of Lugano, CH)

Machine-readable rules and policies are fundamental to KG-based AI, as they can be used to formalize legal requirements, social norms, privacy preferences and licenses that govern the use and exchange of personal knowledge graphs between parties. For many years I have been studying systems for the formalization of norms and policies in the field of Agents and Multiagents Systems by using Semantic Web Technologies. We proposed models for representing and reasoning on obligations, by extending the ODRL language [1] and a model for representing and reasoning on norms able to generate at run-time deontic relationships [2]. Since 2021, I have been co-chair of the W3C ODRL (Open Digital Rights Language) Community Group. I coordinate the activities of the group that defines the semantics of ODRL. In this seminar I would like to investigate how languages for policy specification can be used in Knowledge Graph-based AI. I think it would be fundamental to study what types of explanations can be produced by the policies governing the statements in the various types of KG (personal KG and community KG) that are used by the ML algorithms. It is also crucial to study how to efficiently translate policies from natural language to machine readable formats and how to evaluate policies efficiently in real scenarios.

References

1 Nicoletta Fornara and Marco Colombetti. Using semantic web technologies and production rules for reasoning on obligations, permissions, and prohibitions1. *AI Commun.*, 32(4):319–334, January 2019.
2 Nicoletta Fornara, Soheil Roshankish, and Marco Colombetti. A framework for automatic monitoring of norms that regulate time constrained actions. In *Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIV: International Workshop, COINE 2021, London, UK, May 3, 2021, Revised Selected Papers*, page 9–27, Berlin, Heidelberg, 2021. Springer-Verlag.

4.10 KG-based AI in Industrial Data Ecosystems

Sandra Geisler (RWTH Aachen, DE)

In industrial settings, such as manufacturing, data sharing is still subject to distrust. Companies rarely share their data with academia or industry partners in fear of disadvantages for their business. Highly distributed settings require new methods to utilize meaningful and reliable information from several potentially disparated and contradictory sources to be useful for higher level AI services. Examples for such settings are e.g., cross-organizational stream processing in manufacturing or informing a Digital Product Passport by distributed data analytics at suppliers. Knowledge graphs as linked and semantically rich sources provide useful information to fuel AI methods and improve the quality of their outputs. However, to utilize information from KGs, their complete life cycle and ecosystem needs to be taken into account including their generation, evolution, and integration with other KGs. Especially differing ontologies as bases for the KGs, differing goals of actors in their ecosystems, quality and provenance as well as trust between stakeholders, are still big challenges which I would like to discuss in this seminar.

4.11 AI Accountability and Data Governance from a Pragmatic Perspective.

Anna Lisa Gentile (IBM Almaden Center – San Jose, US)

With the current magnitude of AI systems, a large part of accountability depends on the traceability and availability of the data used within the model. While governance of the data starts at acquisition time, keeping complete track of the vast amount of data used for model training can be unfeasible; therefore, tools that can effectively act and correct the outputs of the models are a must-have. These tools can be thin layers of detectors that can quickly identify sensitive topics and screen unwanted output, as well as detect and quantify inadvertent leakage of proprietary data and correct the model behaviour.

4.12 Infusing Large Language Models with Knowledge: A Round-trip Ticket

José Manuel Gómez-Pérez (Expert.ai – Madrid, ES)

Data is the fossil fuel of AI. While computational power is growing with better hardware and algorithms, data is not. Large language models (LLMs) need to be pre-trained using enormous amounts of data from the Internet. However, Internet data is limited, and it is estimated that LLMs will be trained on datasets equivalent in size to the available stock of public human text data between 2026 and 2032. Therefore, pre-training as we know it will change. Additionally, when fine-tuning pre-trained LLMs for specific domains, the required training data is often unavailable, locked in corporate or regulatory silos.

To address these challenges, I propose the notion of *Knowledge as the New Fuel of AI* and a new paradigm based on infusing existing resources, such as knowledge graphs (KG), into the parametric memory of an LLM. Conversely, the knowledge contained in an LLM can also be distilled and merged into a KG, iteratively producing richer structured and parametric representations. I also introduce the concept of *Problem-Solving Prompting* (PSP) in LLMs, which builds on knowledge-based methodologies such as *Problem-Solving Methods* (PSMs) to extend prompting approaches like *Chain-of-Thought (CoT)* by breaking down complex problems into simpler subtasks.

I conclude with a call to attention on the urgent need for new benchmarks and metrics. These will be instrumental to measure both the factuality of knowledge-infused LLMs and the amount and quality of the knowledge they have been infused with.

4.13 Agreements & Accountability

Paul Groth (University of Amsterdam, NL)

Since the emergence of Large Language Models (LLMs) with the ability to perform in-context learning, we have seen a large reconsideration of knowledge engineering methods and practice. In this talk, I provide an overview of recent discussions by the community and note how LLMs shift the focus from tasks such as knowledge acquisition from text and content to other knowledge engineering tasks. In particular, the ability to come to consensus becomes paramount. How do we get assist people in coming to agreement? How do we assist people and AI systems come to agreement? I argue that knowledge graphs provide mechanism for encoding such agreement. But for that agreement to be trusted, we need to document and be able to explain how that agreement was formulated. We need agents to be accountable for the agreements that they make. Hence, I argue that we need to develop new methods for consensus formulation, mechanisms for maintaining trust, and richer approaches to explanation.

4.14 Building the Future of Enterprise AI: From Neuro-Symbolic Intelligence to Agentic Systems

Peter Haase (Metaphacts – Walldorf, DE)

The integration of symbolic and neural approaches – often referred to as neuro-symbolic AI – is gaining significant momentum, particularly in enterprise contexts where explainability, domain specificity, and trust are paramount. While neural methods such as large language models (LLMs) excel at general language understanding and pattern recognition, they are limited by a lack of transparency, factual grounding, and access to enterprise-specific knowledge. This is where symbolic technologies like knowledge graphs come into play, enabling structured, logic-based reasoning, traceability, and contextual relevance.

At the forefront of this shift is metaphacts, a company developing advanced AI capabilities through its platform metaphactory. Embracing neuro-symbolic and agentic AI, metaphacts is building metis, an intelligent agent powered by an LLM that has direct access

to a knowledge graph. Metis combines cognitive capabilities – such as natural language understanding, semantic reasoning, and planning – with execution capabilities like SPARQL query generation, semantic modeling, and retrieval-augmented generation. This enables conversational interfaces that support use cases ranging from semantic search and discovery to knowledge graph construction and ontology engineering.

These innovations are particularly relevant as the industry moves toward more mature and responsible applications of AI. While generative AI is currently experiencing a period of recalibration – having passed the peak of inflated expectations and facing challenges like hallucinations and scaling – knowledge graphs are climbing the "Slope of Enlightenment" in Gartner's Hype Cycle. They are increasingly seen as critical infrastructure for grounding AI in enterprise semantics and delivering trustworthy, explainable solutions.

Metaphacts is actively contributing to this evolution by operationalizing neuro-symbolic AI in real-world settings. For example, metis can provide transparent answers rooted in enterprise knowledge, maintain provenance, and adapt to user context, making it a powerful tool for organizations navigating regulatory requirements and complex information landscapes. Beyond retrieval, metis also assists in automating parts of the knowledge engineering process, helping reduce the cost and effort of building and maintaining knowledge graphs.

4.15 The Age of Agents: A Tale of Two Traditions

Andreas Harth (Fraunhofer IIS – Nürnberg, DE)

Introduction and Motivation. The rise of large language models has sparked new interest in agent architectures by offering powerful ways to turn natural language into actions. Such development comes as web decentralisation technologies like ActivityPub, Solid and Web of Things are maturing. These decentralised technologies preserve individual self-determination and use knowledge graphs and standardised descriptions to create environments where autonomous agents can operate.

Earlier waves of excitement about software agents in the 1990s and early 2000s, for example around shopping bots and personal digital assistants, fell short of expectations. Now that language models provide powerful means to process natural language and the infrastructure begins to support machine-interpretable representations that would allow for automated interactions with content and services, it is time to re-examine both agent architectures and their underlying technical foundations.

Classification Framework. Agent research has historically followed two main schools of thought that differ in how they view agency. The two perspectives ask fundamentally different questions: "How do we build?" versus "What does agency mean?". These perspectives can be organised as follows:

Perspective	Single-Agent Focus	Multi-Agent Focus
Technical	Mapping inputs to actions	Rationality and game theory
Philosophical/social	Nature of goals and intentions	Social interaction (simulated)

Technical Tradition. The technical view focuses on building systems using formal methods and frameworks for rational decision-making. The computational approach, developed by Nilsson/Genesereth and Russell/Norvig, sees a single agent mainly as a complex function

that turns inputs into actions. When looking at multiple agents, the tradition often focused more on theoretical aspects of rationality and game theory rather than practical challenges of distributed systems.

Philosophical and Social Tradition. The philosophical tradition, exemplified by Cohen/-Levesque, looks at basic questions about the nature of agency itself. The work explores what goals and intentions mean for artificial agents. Wooldridge/Jennings expand the view to multi-agent systems by studying the social aspects of agency, highlighting properties like independence, taking initiative, and ability to interact socially. The tradition typically used centralised simulations to study how multiple agents interact, mostly avoiding the engineering challenges of distributed systems.

Back to Basics: A Minimal Notion of Agency. One approach to advancing the field considers stripping the concept of an "agent" down to fundamental elements: an independently executing process that can be created and terminated, can communicate with other processes and can fail independently. Whether these basic units are called processes, actors or agents may matter less than examining these fundamental capabilities as potential building blocks of any agent system. More sophisticated notions of agency, whether from the technical or philosophical traditions, might then emerge from the minimal foundation.

Web Architecture as Implementation Platform. Web architecture fundamentally operates on a client/server distinction, where the client initiates interactions and the server maintains state. In addition, web architecture assumes web resources, identified via a uniform name, that represent mappings from time to representations. Multiple concurrent processes, actors or agents can be built on such an architecture by combining resources with computation. These entities operate with dual client-server capabilities, requiring both addressability and state maintenance on the server side, along with the ability to initiate interactions as clients. The basic interaction patterns follow web architecture constraints, with all communication flowing through HTTP operations between clients and servers. Such patterns align with fundamental distributed system requirements, that is, maintaining state, message-based communication and independent failure.

Current web technologies demonstrate viable implementation approaches: ActivityPub defines federation protocols for structured interactions, Solid provides personal data spaces for process state storage and Web of Things offers standardised descriptions of interaction affordances. These concrete patterns support implementing minimal distributed components within web architectural constraints.

Conclusion. The combination of language models and decentralised web technologies presents opportunities for agent systems. Examining distributed and concurrent systems principles could inform the development of foundations that address historical challenges while enabling future developments. Clear accountability mechanisms, including user-controlled policy enforcement, transparent operation logs and verifiable computation trails, will be essential for ensuring these agent systems remain answerable to their users. The path forward may lie in rigorously evaluating which elements from both agent traditions can be meaningfully implemented within the architectural constraints of the web while maintaining these accountability guarantees.

4.16 Development of New Approaches for Federated Query Processing

Olaf Hartig (Linköping University, SE)

During the Dagstuhl Seminar I wanted to work on concrete approaches related to some of the goals laid out in the proposed research agenda of the article that was the basis of the seminar [1]. In particular, related to goal DKG6 (Federated Querying), my idea was to develop an approach to consider some form of policies (e.g., access control) during the query planning process of a query federation engine. Related to goal DKG2 (Alignment with Standardized and Community Ontologies), my idea was to apply our approach to consider ontology mappings during federated query processing [2] within a concrete use case. Related to goal MRP2 (Multi-Level Policy Evaluation), I had the idea to develop a policy evaluation algorithm, and in the context of goal DI1 (Comprehensive Recording), I was considering to develop a federated query processing approach that integrates a blockchain as a federation member.

References

1 Luis-Daniel Ibáñez, John Domingue, Sabrina Kirrane, Oshani Seneviratne, Aisling Third, and Maria-Esther Vidal. Trust, Accountability, and Autonomy in Knowledge Graph-Based AI for Self-Determination. *Transactions on Graph Data and Knowledge*, 1(1):9:1–9:32, 2023.
2 Sijin Cheng, Sebastián Ferrada, and Olaf Hartig. Considering vocabulary mappings in query plans for federations of rdf data sources. In *Cooperative Information Systems: 29th International Conference, CoopIS 2023, Groningen, The Netherlands, October 30–November 3, 2023, Proceedings*, page 21–40, Berlin, Heidelberg, 2023. Springer-Verlag.

4.17 "Fundamental Science" is discovering AI

James A. Hendler (Rensselaer Polytechnic Institute – Troy, US)

Semantic Web and Knowledge Graph technology is becoming more relevant again as we see that it can help solve some of the key challenges emerging with generative AI. In this short talk, I show some challenging problems in the traditional sciences and explore why advanced AI techniques, particularly focused on data integration (and its realization as ontologies and knowledge graphs on the distributed web) are critical to making progress in these areas.

4.18 Large Language Models, Knowledge Graphs and Search Engines: A Crossroads for Answering Users' Questions

Aidan Hogan (University of Chile – Santiago de Chile, CL)

Much has been discussed about how Large Language Models, Knowledge Graphs and Search Engines can be combined in a synergistic manner. A dimension largely absent from current academic discourse is the user perspective. In particular, there remain many open questions

regarding how best to address the diverse information needs of users, incorporating varying facets and levels of difficulty. This short talk introduces a taxonomy of user information needs, which guides us to study the pros, cons and possible synergies of Large Language Models, Knowledge Graphs and Search Engines. We further present a roadmap for future research.

4.19 Building Trust in Knowledge Graphs with Provenance and Data Quality

Katja Hose (TU Wien, AT)

As the volume of data continues to grow, ensuring the reliability of AI-driven systems depends on the quality, provenance, and interoperability of the knowledge they use. Knowledge Graphs play a crucial role in structuring, integrating, and providing access to factual knowledge, which is essential for both human understanding and AI applications such as LLMs. Without mechanisms to verify and track the origins of data, errors and inconsistencies can propagate, undermining trust in AI-generated results. This is particularly relevant for Large Language Models, which often suffer from hallucinations even without relying on incomplete or incorrect information [1].

To address these challenges, validation techniques based on SHACL and ShEx help enforce structural constraints and improve the consistency of knowledge graphs [3]. Provenance tracking further enhances transparency by documenting the origins and transformations of data and how answers to queries are derived from the data [5], allowing users to assess its reliability. Efficient access to evolving knowledge and scalable mechanisms for ensuring data integrity are also crucial for maintaining trustworthy AI applications.

In addition to data quality and provenance, interoperability across different knowledge graph data models is vital. Knowledge representations vary between RDF-based graphs, property graphs, and other emerging models, making it essential to establish common foundations that support seamless integration and querying [4, 2]. Standardized schemas and transformation techniques enable interoperability, ensuring that knowledge can be effectively shared and leveraged across different systems without loss of meaning or consistency.

By strengthening provenance tracking, validation mechanisms, and interoperability strategies, we can build more reliable, accountable, and explainable AI systems. Future developments in this area will shape the way knowledge is managed and utilized, ensuring that AI-driven insights are built on a foundation of high-quality, verifiable, and interoperable information.

References

1 Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. Knowledge graphs, large language models, and hallucinations: An nlp perspective. *Journal of Web Semantics*, 85:100844, 2025.

2 Kashif Rabbani, Matteo Lissandrini, Angela Bonifati, and Katja Hose. Transforming rdf graphs to property graphs using standardized schemas. *Proc. ACM Manag. Data*, 2(6), December 2024.

3 Kashif Rabbani, Matteo Lissandrini, and Katja Hose. Extraction of validating shapes from very large knowledge graphs. *Proc. VLDB Endow.*, 16(5):1023–1032, January 2023.

4 Shqiponja Ahmetaj, Iovka Boneva, Jan Hidders, Katja Hose, Maxime Jakubowski, José Emilio Labra-Gayo, Wim Martens, F. Mogavero, Filip Murlak, Cem Okulmus, Axel Polleres, Ognjen Savkovic, Mantas Simkus, and Dominik Tomaszuk. Common foundations for shacl, shex, and pg-schema. In *Proceedings of the ACM Web Conference 2025*, New York, NY, USA, 2025. Association for Computing Machinery.
5 Daniel Hernández, Luis Galárraga, and Katja Hose. Computing how-provenance for sparql queries via query rewriting. *Proc. VLDB Endow.*, 14(13):3389–3401, September 2021.

4.20 Remaining (Self-)Determined in a world of Agentic AI

Luis-Daniel Ibáñez (University of Southampton, GB)

In this talk I provide an overview of the challenges to self-determination that individuals face with the upcoming era of Agentic AI. AI agents could be misused to collect further data about individuals in order to nudge them towards a goal not necessarily aligned with their interests. In an extreme case, as described by Chaudhary and Penn, human motivations can be collected and sold to agents, an evolution of the economy of attention into the economy of intention. Then, I'll motivate the following research questions related to the construction of a counter-agent that help citizens navigate the economy of intention. Can an AI agent be designed to help a human retain their autonomy in a world of intent merchant agents? On what infrastructure? What is the minimum information required for such an agent to work? How much asymmetry in the intelligence of the competing agents is tolerable?

4.21 Trust and Accountability in Financial AI

Ryutaro Ichise (Institute of Science Tokyo, JP)

In this talk, I introduce two key projects related to trust and accountability.

Causality is essential for decision making in finance, yet existing knowledge graphs face challenges such as complex logical structures, inconsistencies, and limited reusability. To address these, we designed a pipeline for constructing a causal knowledge graph, FinCaKG-Onto, which integrates text, ontologies, and linked data. Our approach outperforms ChatGPT in capturing nuanced causality while avoiding generic concepts.

One application of FinCaKG is identifying dominant factors in financial causal chains. We developed a pattern mining strategy to extract dominant factors, validated through real market data. This work enhances explainability and accountability in financial decision-making.

The second project is about hallucination detection. Large Language models (LLMs) sometimes generate hallucinations – incorrect or misleading facts – which pose risks in high-stakes scenarios. Unlike LLMs, knowledge graphs offer structured reliable facts. We developed a text verification framework leveraging knowledge graphs to detect hallucinations and tested it using systematically generated hallucination data.

4.22 (Logics-aware) KG Alignment, KG Validation, KG Embeddings, Neurosymbolic AI

Ernesto Jiménez-Ruiz (City St George's, University of London, GB)

This presentation explores key aspects of Knowledge Graph (KG) technologies, focusing on: i) Knowledge Graph (KG) Alignment, ii) KG Validation, iii) KG Embeddings, and iv) Neurosymbolic AI. The talk highlights the importance of logical consistency, reasoning, and explainability in AI-driven knowledge graphs.

Knowledge Graph Alignment

Detecting Different Modeling Views. The integration of multiple models can lead to logical inconsistencies, known as unsatisfiabilities. These issues often arise due to different modeling perspectives or incorrect mappings. Algorithms for minimizing conservativity violations play a crucial role in ensuring alignment integrity.

Using Large Language Models (LLMs). Recent advancements in language models facilitate ontology subsumption inference, aiding the alignment process by identifying relationships between concepts.

Knowledge Graph Validation

Validation using Datalog Rules. Some OWL axioms are treated as integrity constraints. Missing information is captured through violation predicates, and reasoning with datalog rules ensures data consistency and completeness.

Hybrid AI: Ontology Embeddings

OWL2Vec* for Ontology Embeddings. Techniques like OWL2Vec* enable the embedding of OWL ontologies, enhancing machine learning applications by incorporating structured knowledge representations.

Learning with Knowledge Graph Embeddings. Knowledge Graph Embeddings (KGE) are utilized for classification tasks, such as predicting adverse biological effects of chemicals. These embeddings enhance explainability and facilitate reasoning over unseen entities.

Neurosymbolic AI

Learning with Prior Knowledge. Incorporating logical constraints into machine learning models helps maintain consistency by penalizing incorrect predictions. This approach bridges symbolic reasoning with data-driven AI.

Conclusion and Future Work

The presentation also provides an overview of research initiatives, including the work conducted at the AI Research Centre at City St Georges, University of London. It connects these efforts to the foundational topics of the Dagstuhl Seminar 25051, reinforcing the importance of robust validation, integration techniques, and explainability in AI-driven knowledge graphs. The research presented at the seminar emphasizes: i) The need for robust

alignment techniques to resolve inconsistencies in multi-source KGs. ii) The role of reasoning mechanisms in ensuring logical validity. iii) The integration of embeddings and symbolic AI for improved decision-making. Further exploration in neurosymbolic AI and adaptive KG alignment mechanisms will be crucial for advancing trust and accountability in AI systems.

4.23 Threats to Trust in Organizations' Operationalized Knowledge

Timotheus Kampik (SAP Berlin, DE & Umeå University, SE)

In my presentation, I discussed (classes of) threats affecting the trustworthiness of operationalized knowledge that is used to define how socio-technical systems run in and across organizations. I argued that threats relating to malicious behavior (attacks) and lack of compliance (e.g., privacy issues) are relatively well-understood. In contrast, there are more subtle classes of threats that are prevalent, severe, and poorly understood. Specifically, bullshit knowledge emerges when (human) agents are compelled to formalize knowledge they do not fully understand or care about, and that formalized knowledge is typically not generalizable, meaning that its trustworthiness depends on context. How these threats can be addressed in a systematic manner remains to be seen.

4.24 Compliance Technologies for Trust

George Konstantinidis (University of Southampton, GB)

In this talk I present recent advances in data compliance technologies. I discuss how the use of computational and machine processable policy languages can enable data usage control, encoding a range of rules from legislation to regulation and environmental or other preferences. I discuss how compliance algorithms implemented on policy engines can verify compliance or detect conflicts. I present our tools for managing, tracking and updating user consent and preferences on data and AI operations and pipelines. I present an approach for automated negotiation and execution of policies, contracts and agreements. Lastly, I present a framework and roadmap for trust and reputation management algorithms and systems.

4.25 Neurosymbolic GeoAI

Manolis Koubarakis (University of Athens, GR)

It has been shown by Tony Cohn and coauthors (COSIT 2024) that current large language models do not perform well on spatial reasoning problems. For example, they cannot answer questions such as "You are walking south along the east shore of a lake and then turn around to head back in the direction you came from, in which direction is the lake? Is it to your

left or to your right?" (Correct answer: Left; LeChat by Mistral does not answer correctly; ChatGPT answers correctly). My current research concentrates on evaluating the most recent large reasoning models on spatial reasoning tasks such as the above, and combining LLMs and spatial reasoners for solving such tasks effectively (hence, the neurosymbolic term in the title). In this way, we will be able to develop chatbots that will do better in geographic tasks and be more useful than current ones in geospatial applications (e.g., way-finding).

4.26 Knowledge Graph-Aware AI in the Evolving AI Landscape

Deborah L. McGuinness (Rensselaer Polytechnic Institute – Troy, US)

We are living in an age of rapidly advancing technology. History may view this period as one in which generative artificial intelligence is seen as reshaping the landscape and narrative of many technology-based fields of research and application. Times of disruptions often present both opportunities and challenges. I briefly introduce some areas for discussion about how and where knowledge graphs (both personal KGs and other KGs) may be positioned in emerging hybrid architectures that may provide value propositions that might impact adoption. I also provide some dimensions (such as explainability, interoperability, etc , through which we may view and evaluate the potential of knowledge graphs in today's landscape.

4.27 Machine processable policies for socio-technical systems

Julian Padget (University of Bath, GB)

Policies are one important source of trust for entities participating in socio-technical systems because they offer guarantees on accountability as well as expectations of behaviour and the achievability of goals. My past work on norm representation and reasoning has taken a formal approach using action languages, which was then combined with ODRL to provide an operational semantics for fragments of GDPR represented in ODRL. Other contributory contextualising factors for trust, although out of scope here, but still relevant are various standards and guidelines for process and for technologies, such as IEEE 7001-2021, IEEE 7003-2024, ISO 42001 and the UN Guide on Privacy-Enhancing Technologies. My goals for this seminar are to explore more effective ways to build and maintain machine processable policies, facilitated by large language models, but alongside formal approaches, to support the engineering of socio-technical systems.

4.28 Trusting Query Results

Philipp D. Rohde (TIB – Hannover, DE)

Federated query processing answers a query retrieving data from multiple sources as if they were a single source. This requires (semantic) source descriptions as well as query decomposition and planning with respect to the capabilities of the different sources. The data within a KG can be validated against (integrity) constraints using shape-based languages like SHACL or ShEx. When it comes to process-based data, e.g., cancer patients following the treatment guideline, a new shape-based validation language, PALADIN, is proposed. But when it comes to trust in query results, different perspectives need to be considered. A computer scientist might have a different view on what is trust than the average user. The quality of the data is only one dimension that contributes to trust. Other dimensions like access control and provenance are also discussed.

4.29 Trust-Based Decision Support Systems

Daniel Schwabe (Rio de Janeiro, BR)

We investigate how hybrid systems integrating Knowledge Graphs (KGs) and generative language models assist decision-making in various domains. Our goal is to explore how these systems can best support decision processes. The decision-making process should be understood as a reasoning mechanism that aligns with the intended goals (purpose) of a human agent executing a particular action, taking into account their personal preferences, characteristics, and values. To reach a decision, the agent assesses Its decision policies applied to trusted information. This includes both circumstantial information about the situation at hand and contextual information on relevant factors. To obtain trusted information, the agent accesses various sources, including potentially crowdsourced Knowledge Graphs (KGs) and Large Language Models (LLMs). The agent then applies its trust policies to the information retrieved from these sources to extract reliable information. To trust a particular piece of information, the user constructs (possibly recursively) a trust chain of claims, evidence, and supporting proofs to make a final trust decision. To decide if a particular claim is to be used as a fact for the specific intended action, there are three possible alternatives.

1. The agent already accepts that claim as a fact because it already knows it to be the case;
2. The claim is to be accepted as a fact because of social norms. For example, it was made by an agent with public faith, such as a notary public;
3. If all resources have been exhausted, e.g, time, computational resources, lack of additional information, etc., the agent makes an arbitrary decision which is locally referred to as a "leap of faith". In other words, accepting a claim as a fact without any kind of evidence.

Our research investigates the various architectures, representations, and functionalities of hybrid systems (also called neurosymbolic systems) that can support this decision-making. Specifically, we are looking on how to include explicit representations of context in knowledge graphs, and how to support justification dialogues, using both knowledge, graphs, and LLMs in supporting the decision process, for specific domains.

4.30 Bridging Resilient Accountable Intelligent Networked Systems (BRAINS)

Oshani Seneviratne (Rensselaer Polytechnic Institute – Troy, US)

This short talk introduces the research conducted at the BRAINS Lab at RPI, which focuses on enhancing trust and accountability in decentralized AI systems. Our work spans three high-level areas: decentralized privacy-preserving data infrastructures, smart contract innovations, and foundation model innovations with decentralized technologies. At its core, this research explores how decentralized knowledge graph ecosystems can empower individuals while ensuring safety, transparency, autonomy, and alignment with human values. In the context of this Dagstuhl Seminar, I am particularly interested in advancing several foundational topics. For machine-readable policies and norms, I aim to explore how knowledge graphs can accurately represent policies and how these policies can remain adaptable and enforceable in decentralized systems. For decentralized infrastructures, I focus on how best to ingest knowledge into vertical and hybrid federated learning systems and the design of knowledge-infused architectures for LLMs. For decentralized knowledge graph management, I seek to address the challenges of sustainably managing decentralized ecosystems by keeping knowledge up-to-date, incentivizing contributions and verification, and handling contradictions or diverse viewpoints. Finally, for explainable neuro-symbolic AI, I am investigating how to ensure AI safety and protect personal data when integrating LLMs with personal knowledge graphs while providing clear and effective explanations.

4.31 Trust, Accountability, and Autonomy in Generative Health AI

Chang Sun (Maastricht University, NL)

This talk addresses the dimensions of trust, accountability, and autonomy in the development and deployment of generative AI systems for health data. It introduces a novel methodology for generating synthetic patient data for rare or previously unseen diseases using ontology-enhanced generative adversarial networks (Onto-CGAN). By integrating biomedical ontologies into the training process, the approach improves the quality and relevance of synthetic data, enabling machine learning models to generalize more effectively in scenarios where real data is scarce. While synthetic data does not fully match the performance of real-world data, it significantly outperforms models trained on limited or no data.

In addition to unimodal synthetic data generation, the talk explores the application of multimodal language models to radiological visual-linguistic tasks, highlighting the need for interpretability and task-specific evaluation in clinical settings. The presentation also introduces the ciTIzen-centric DAta pLatform (TIDAL), a privacy-preserving infrastructure designed to support dynamic and fine-grained digital consent management. Built on Solid (SOcial LInked Data) principles and employing the Data Privacy Vocabulary, TIDAL enables secure, decentralized storage and governance of personal health data, with consent-aware federated learning capabilities. All data and consent artifacts are represented in RDF, allowing for semantic interoperability and standards-based integration.

4.32 You can't pin down trust, but you can still do something

Aisling Third (The Open University – Milton Keynes, GB)

Concepts like trust get defined in various fields, without necessarily representing the same phenomenon. This carries the risk of serious problems where, e.g., technical systems with different definitions interact. This is by no means unique to trust, of course, but we can observe that its nature as a foundational concept of social interaction makes it easier for unconscious assumptions to come into play. It is clear that we need to handle these concepts of trust in a flexible way. This sort of problem is what the Semantic Web can be useful for: interoperability by making concepts explicit. But it is equally unlikely that formal languages can be used to capture these concepts either. We argue therefore that it would be more fruitful to model instead the factors which go into making trust decisions, e.g., user ethical and social values, any requirements of secrecy, etc., including how these relate to trust, so that operationalising trust decisions can still be handled by relevant actors with the required information to do so.

4.33 The value of trust that surrounds data

Ruben Verborgh (Ghent University, BE)

For the longest time, we assumed that Linked Data – public or private – was about technologies that facilitate the transfer of data. Maybe we had it all wrong. When we download RDF from DBpedia, we essentially get back a list of triples. Like any series of bytes, these map losslessly to a list of natural numbers and back. With the natural numbers being a known – albeit infinite – set, DBpedia cannot possibly send us any new data points. Hence, it must be sending us something else. The story becomes very different when we realize that the value of what DBpedia sends us, is not in the numbers themselves, but in the trust assessment that DBpedia is implicitly making when sending them. Namely: this is a list of triples to which DBpedia attaches some truth value. And while there similarly exist an infinite number of such lists, we attach value to this particular one, because DBpedia endorses it. Unfortunately, we as a community are not very explicit at all about the semantics of that trust, which makes it hard to capture and discuss value. Let's talk about trust.

4.34 Hybrid AI Systems with Knowledge Graphs: Enabling Trust, Accountability, and Autonomy

Maria-Esther Vidal (TIB – Hannover, DE)

Artificial Intelligence (AI) is transforming science and medicine by enabling powerful predictive and decision-support systems. However, ensuring trust, accountability, and autonomy in AI remains a challenge, particularly when models operate as black boxes. Hybrid AI

systems, combining symbolic reasoning with machine learning, offer a promising approach to overcoming these challenges. This presentation discusses the integration of Knowledge Graphs (KGs) into neuro-symbolic AI systems, emphasizing their role in enhancing interpretability and ensuring robust decision-making. Knowledge Graph (KG) ecosystems provide structured, semantic representations of knowledge, supporting data integration, constraint validation, and symbolic reasoning. A KG ecosystem is defined by various components, including data sources, ontologies, mappings, and constraints, facilitating the construction of AI systems that are both explainable and trustworthy. The life cycle of KG-based AI systems involves services, actors, roles, constraints, and requirements that ensure sustainable and transparent AI-driven decision-making. Hybrid AI leverages both symbolic and neural components. Symbolic methods, such as constraint validation and rule-based reasoning, ensure valid and explainable link prediction and counterfactual inference. Meanwhile, neural components, including numerical learning, KG embedding models, and large language models (LLMs), enhance learning from unstructured data while preserving logical consistency. The synergy between these components enables the development of AI systems capable of self-determination, improving autonomy in critical applications. The discussion will explore principled vs. integrated neuro-symbolic systems, highlighting the need for AI architectures that ensure reliability without sacrificing efficiency. This principled approach fosters trust in AI-driven applications, particularly in domains requiring high levels of interpretability, such as medicine and scientific research.

4.35 Towards neuro-symbolic agents that represent legal entities on the Web

Jesse Wright (Open Data Institute – London, GB)

The notion of agentic AI is seeing resurgent popularity in the age of LLMs-based AI. We pose a research agenda towards building hybrid agents which use LLMs to provide a human interface for agents and to support the negotiation capability of agents – whilst query and reasoning is used to provide operational safeguards to data "belief" and "sharing".

To support the maintenance of proof and provenance over derived data that agents receive we propose directions including zero knowledge proofs or e.g. SPARQL query correctness, to enable agents to have models for establishing whether the provenance they receive is sufficient to take that data to be true. We propose personalised trust modelling so agents can learn what sources their users are willing to take as authoritative for particular tasks. We also propose personalised privacy preference modelling to enable agents to automate data sharing.

5 Breakout Groups

5.1 Machine Readable Norms and Policies

Piero A. Bonatti (University of Naples, IT)
Irene Celino (CEFRIEL – Milan, IT)
Andrea Cimmino (Polytechnic University of Madrid, ES)
Nicoletta Fornara (University of Lugano, CH)
Andreas Harth (Fraunhofer IIS – Nürnberg, DE)
Luis-Daniel Ibáñez (University of Southampton, GB)
Timotheus Kampik (SAP Berlin, DE & Umeå University, SE)
George Konstantinidis (University of Southampton, GB)
Julian Padget (University of Bath, GB)
Oshani Seneviratne (Rensselaer Polytechnic Institute – Troy, US)

Abstract. As AI systems increasingly mediate complex interactions in socio-technical ecosystems, the need for formal, machine-readable representations of norms and policies becomes critical. This report introduces the concept of Computational Policy Languages and formalizes core policy reasoning tasks: generation, activation, evaluation, and enforcement. We define a Policy Engine as a computational artifact capable of supporting these tasks against a dynamic, knowledge-graph-based representation of the world. To ground this framework, we explore four operational scenarios – intending, attempting, monitoring, and auditing – that structure the temporal and procedural dimensions of policy application. A set of diverse use cases, including business process compliance, financial contract management, industrial safety, organizational governance, and energy data sharing, illustrates the breadth of challenges and requirements such systems must address. This report consolidates a research agenda for formal, interoperable, and context-aware policy systems, identifying open problems at the intersection of logic, semantics, system design, and regulatory alignment.

5.1.1 Introduction

Complex ecosystems for Knowledge-Graph based AI include multiple interactions between their participants. In several scenarios, a participant's action may hurt the self-determination of another (human) actor. For example, an AI agent actor deciding to deny a benefit, potentially in an unlawful manner, without providing the right to recourse. Another example is a human actor sharing some data that was considered private or sensitive by another actor.

To remediate, it is essential to equip these ecosystems with the means to define and enforce norms, policies, and rules so they can express 1. Global norms that all actors are expected to follow, or rules that their actions must not break. This could serve, for example, to encode principles that protect the self-determination of all actors. 2. Individual rules, constraints, or preferences of actors establishing boundaries with respect to actions from other actors.

A desirable characteristic of these policy languages is being machine-readable and machine-processable. The reason is two-fold: first, it facilitates the checking of when a policy has been violated by an action, or if an individual preference is incompatible with a global norm; second, when AI-agents are actors in the ecosystem, allow them to read the policies.

- Trust: Clear definitions of rules that every actor can evaluate and compare following a deterministic algorithm improve trust in the whole ecosystem.
- Accountability: In combination with the appropriate log and trace systems, determine which actor has violated a rule and what the consequences or repair actions are.
- Autonomy: On the one hand, rules can be designed to protect the autonomy of certain actors, while on the other, their existence allows AI agents to become more autonomous, in the sense they can perform more actions with confidence a rule is not violated.

In this short paper, we provide a review of the state of the art, our proposed approach based on the definition of computational policy languages, a list of use cases from the practice of the participants' group discussion, including desiderata and challenges for computational policy languages to support them.

5.1.2 State of the Art

Policy Languages were initially designed to tackle the problem of *Access Control* and its generalisation to *Role-based access control* (RBAC): An organisation defines a number of roles, to be fulfilled, an actor or agent playing a role needs access to data and resources whose access is governed by policies. An RBAC model and implementation must guarantee that users can access the required data and resources in accordance with organisational policies [4].

The notion of Access Control was further generalised to *Usage Control* [13]. Access Control can be considered as a problem of Authorisation or Permission to perform an action on a target object, and usage control adds the concepts of obligations and conditions. Obligations are requirements that have to be fulfilled for usage allowance. Conditions are environmental or system requirements that are independent of individual subjects and objects. Usage control also emphasizes the continuity of enforcement. Policies are enforced not only before access but also during and after the agent is acting upon the target object. If attributes change during access and the policy is no longer satisfied, usage control systems may revoke the granted access and terminate the usage.

Ibáñez et al. [8] classifies policy languages as general and specific. General languages cater to a diverse range of functional requirements (e.g., access control, query answering, service discovery, negotiation), whereas specific languages focus on a single functional requirement. A number of general languages were developed, but none of them achieved mainstream adoption. On the specific languages front, the Open Digital Rights Language (ODRL), which is a W3C recommendation, has gained a lot of traction in recent years thanks to its use to express intellectual property rights management. Additionally, the ODRL model and vocabularies have been extended to model contracts, personal data processing consent, and data protection regulatory requirements.

Akaichi and Kirrane [2] defines a usage control framework as a complete solution that allows for the specification of usage control policies, the enforcement of said policies, and the realization of policies and enforcement mechanisms via a usage control system. They provide a taxonomy of requirements for a usage control framework. The three top categories are (i) Specification, representing requirements relating to policy expressiveness, defined semantics, as well as flexibility and extensibility of the policy language; (ii) Enforcement, or mechanisms used to enforce and manage usage policies throughout the usage process, which consists of three phases: before usage, ongoing usage, and after usage; and (iii) System, that refers to non-functional requirements such as usability and performance.

A problem of additional interest is the application of Usage Control in a decentralised system, or a complex scenario with multi-agent systems.

Kampik et al. [9] discusses the relevance of norms, policies, and preferences for governing complex sociotechnical multiagent systems on the Web. The key challenge they identify is the integration of normative concepts with WoT abstractions and systematic evaluation of the practical usefulness of the integration results. They propose a conceptual framework that serves to define the role played by various norms, policies, and preferences when it comes to complex sociotechnical systems on the Web and demonstrate it via a simple but realistic scenario.

[11] consider the general distributed case; they propose a generic and formal model that allows for the explicit distinction of different systems, their individual behaviors, as well as their interplay, enabling reasoning about the distributed system they form. The first model and implementation that transparently and generically tracks dataflows and policies across systems.

5.1.3 Proposed Approach: Computational Policy Languages

We propose the following definition of Computational Policy Languages:

▶ **Definition 1** (Computational Policy Language)**.** A Computational Policy Language is a formal language used to define, reason with, and enforce policies, including permissions, prohibitions, and obligations, to control and govern the usage of resources and behavior of actors in socio-technical systems.

▶ **Definition 2** (State of the World)**.** A data structure that holds knowledge about the state of the socio-technical systems on which Computational Policies apply. In the spirit of the seminar, we assume the State of the World is encoded in a Knowledge Graph. Depending on context of application, the state of the world may be a snapshot of the system a ta given time, or contain information about states across a time continuum.

Based on these definitions, we identify the following scientific problems associated *Policy Computational Languages*

▶ **Definition 3** (Policy Generation)**.** In general, this problem refers to the translation of Natural Language to a Computational Language. We recognise the following variations of the problem:

- Policy creation/authoring/specification: given a description of a policy in natural language, generate a policy in a computational policy language. It is also possible to consider the problem of creating a User Interface to generate policies.
- Policy legitimation: given a machine-readable or machine-processable format, generate a natural language version expressed with domain-specific terms, such that it is admissible in a given legal framework.
- Policy explanation: given a machine-readable, machine-processable, or natural language policy, generate a natural language description of what the policy is about.

▶ **Definition 4** (Policy Activation)**.** Given a state of the world and a list of policies, select the subset of policies that are relevant to be evaluated against the state of the world. For example, if a policy states that the Actor must be inside the library on Tuesdays between 09:00 and 18:00 and the state of the world is: Today is Wednesday, 10:00, and the actor is in the Library, then the policy is not 'active' in this state of the world.

▶ **Definition 5** (Policy Evaluation). Given a Computational Language Policy and a state of the world, decide if the state of the socio-technical system violates the policy or not. For example, if a policy states that Actor must be inside the library on Tuesdays between 09:00 and 18:00, then for the following states of the world:

- Is Tuesday 10:00 and the actor is in the Library: The state does not violate the policy.
- Is Tuesday 10:00 and the actor is in the Kitchen. The state violates the policy.

▶ **Definition 6** (Policy Enforcement). Given a state of the world and a set of violated policies, compute a set of updates to the state of the world such that the updated state of the world does not violate the input set of policies.

Finally, for practical reasons, we consider the definition of an artefact that encapsulates algorithms for each of the problems, that we dub *Policy Engine*.

▶ **Definition 7** (Policy Engine). A system that integrates algorithms that solve each of the scientific problems of a Computational Policy Language.

5.1.3.1 Operational Scenarios

The operationalisation of the solutions to the problems described in section 5.1.3 depends on the contextual scenario in which they are invoked. We identify four *operational scenarios* for the application of policies.

- *Intending* refers to when an agent intends to execute an action on the state of the world, before proceeding, the agent asks a policy engine to activate and evaluate policies to understand if any violation would occur.
- *Attempting* refers to when an agent attempts to execute an action on the state of the world, before letting the action affect the state of the world, a policy engine activates and evaluates policies to decide if any violation would occur.
- *Monitoring* refers to a process or meta-system that monitors the actions multiple agents execute upon the state of the world, continuously activating and evaluating policies upon those actions.
- *Auditing* refers to the post-facto analysis of a temporal trace of actions upon the state of the world with the purpose of determining if any action on the trace violated any of the policies active at the moment it was executed.

5.1.4 Use Cases and Challenges

In this section, we describe five use cases that would benefit from Computational Policy Languages, highlighting their desiderata and what challenges they pose to the computational problems defined in section 5.1.3.

5.1.4.1 Business Process Conformance Checking

Process conformance evaluates whether an executed business process conforms to certain policies or norms. For example, in an internal audit, an organization may want to check to what extent employees engage in so-called *maverick buying*, *i.e.*, the issuing of a purchase requisition before the approval of the purchase. Maverick buying cannot generally be prohibited. For example, employees may need to be able to purchase work equipment even when their line management is not available and hence cannot issue approvals before the fact. However, if maverick buying is rampant, this indicates that employees are exploiting the system, undermining cost control.

Operational scenario/life-cycle steps. Process conformance checking is a well-established method in business process management, describing the assessment of how well traces of real-world process execution conform to expected behavior [5]. Such expected behavior can be derived, for example, from external regulations, internal best practices, or knowledge about system fundamentals (for data quality checks). In the life cycle of business process management, which in its simplest form consists of process design, execution, and analysis, conformance checking is typically assigned to the analysis step. Accordingly (and in accordance with the maverick buying example), the most common operational scenario is *auditing*, followed by *monitoring*, when applied to cases of a process that have not yet terminated.

Computational problems. The primary computational problem in conformance checking is to determine whether a trace of a process execution is conformant or not. This corresponds to *policy evaluation* in an auditing scenario. However, other computational problems are relevant as well. For example, comparing (sets of) conformance rules helps assess to what extent different organizational conformance requirements agree or, on a more fundamental level, are logically consistent with each other.

Desiderata/challenges. A suitable computational policy language needs to fulfill the following high-level requirements.

- **Logical time as first-class abstraction.** In business process management, logical time is generally considered a crucial first-class abstraction and is commonly formalized using Petri nets [16] or abstractions utilizing finite-trace linear temporal logic [6]. Accordingly, a policy language that can be used or designed for conformance checking must feature a notion of logical time.
- **Operationalization on (Big) Symbolic Data.** In practice, conformance checks are typically executed on large amounts of data that is often stored and queried using special-purpose database systems [10, 17]. Accordingly, policies for process conformance must have an operational semantics that ultimately allows for execution in the context of the aforementioned systems, or by mainstream (e.g., SQL-based) query engines.
- **Models of agents and their roles.** In business processes, (human and software) agents interact, executing activities in order to achieve an organizational objective. Accordingly, the notion of an *agent* is important. In business process management, agents are traditionally called *resources* that have specific *roles* when executing activities [18]. More recent work features the notion of an *agent*, more specifically in a meta-model for process traces that can be utilized for agent-oriented process conformance checking [14].
- **Meta-level Meaning/Labels of Policy Rules.** When a process trace violates a compliance rule, the nature of the rule impacts the implication that non-conformance has. For example, rule violation may imply (on the meta-level) that the process trace does not comply with a specific regulation, is likely to negatively impact organizational performance, or even that the underlying data most be incorrect (e.g., if according to the trace, a message is received before it is sent) [1]. Accordingly, a policy language for process conformance must support the labeling of policy rules according to their meta-level meaning.
- **Support for Deontic Notions.** Some academic work on process conformance checking applies deontic logic [7]. Still, it is not clear how exactly mainstream process conformance checking relates to classical deontic notions of permission, prohibition, and obligation. However, conformance checking typically distinguishes between *imperative* (*How does the process have to behave?*) and *declarative* (*How is the process permitted to behave?*) approaches, thus reflecting deontic ideas.

5.1.4.2 Over the Counter Financial Derivatives Contracts

Over-the-counter (OTC) financial derivatives contracts are customized agreements between counterparties, and their terms must be explicitly defined in a machine-readable format. The contract terms – such as payment schedules, interest rate adjustments, margin requirements, and credit events – are encoded as rule-based policies. OTC contracts can have customized tenors (durations) ranging from days to decades. This flexibility requires a computable policy that governs time-sensitive aspects, including contract lifecycle management, expiration triggers, and time-dependent risk adjustments. OTC derivatives contracts are inherently flexible, allowing for bilateral negotiations and secondary market trading. They can be transferred to new counterparties through assignment or novation while retaining the original policy conditions or incorporating additional terms.

The OTC contracts consist of multiple phases from trade initiation to termination, with several operational steps (Depicted in Figure 1).

1. The terms of the contract (e.g., notional amount, tenor, strike price, underlying asset) are agreed upon bilaterally for the *trade initiation*.
2. This is followed by *trade confirmation* where the contract details are matched, and signed by the counterparties.
3. *Collateral exchange* then takes place to mitigate the counterparty risk.
4. During the period the OTC contract takes place, several *trade lifecycle events*, such as interest payments, happen.
5. When the contract reaches the expiration date *trade termination* happens.

In the case of a transfer of the OTC contract between `Party A` and `Party B`, by `Party A` to another counterparty `Party C`, we may encounter the following set of activities, given the nature of the transfer.

1. *Assignment (Partial Transfer):* The original counterparty `Party A` assigns its rights and obligations to a new counterparty `Party C`. Party A is still legally responsible unless explicitly released by the original agreement.
2. *Novation (Full Transfer)*: The entire contract is legally replaced with a new agreement. The original party (`Party A`) is fully discharged, and the new party (`Party C`) takes full responsibility.

When transferring an OTC derivative, the core contract structure remains unchanged unless explicitly renegotiated. In other words, obligations, risk exposure, and collateralization remain as originally defined. In some cases, counter-parties may wish to modify the contract when trading it to a new party. Additional conditions can include credit enhancement clauses (e.g., requiring a third-party guarantor for counter-parties with a lower credit rating), trigger-based clauses (e.g., automatic termination if market conditions exceed predefined risk thresholds), or regulatory compliance adjustments (e.g., reporting structure adjustments due to jurisdiction changes).

Desiderata/challenges. A suitable computational policy language needs to fulfill the following high-level requirements.

- Possibility to define obligations, risk, exposure, and definition of collaterals.
- Ability to model policy validity and expiration as a function of time.
- The concept of a contract as a first-class citizen. An interesting question is whether Smart Contract languages can be classified as computational policies or if they should be regarded as an extension with additional problems.
- Support for a single policy involving multiple actors with different obligations.

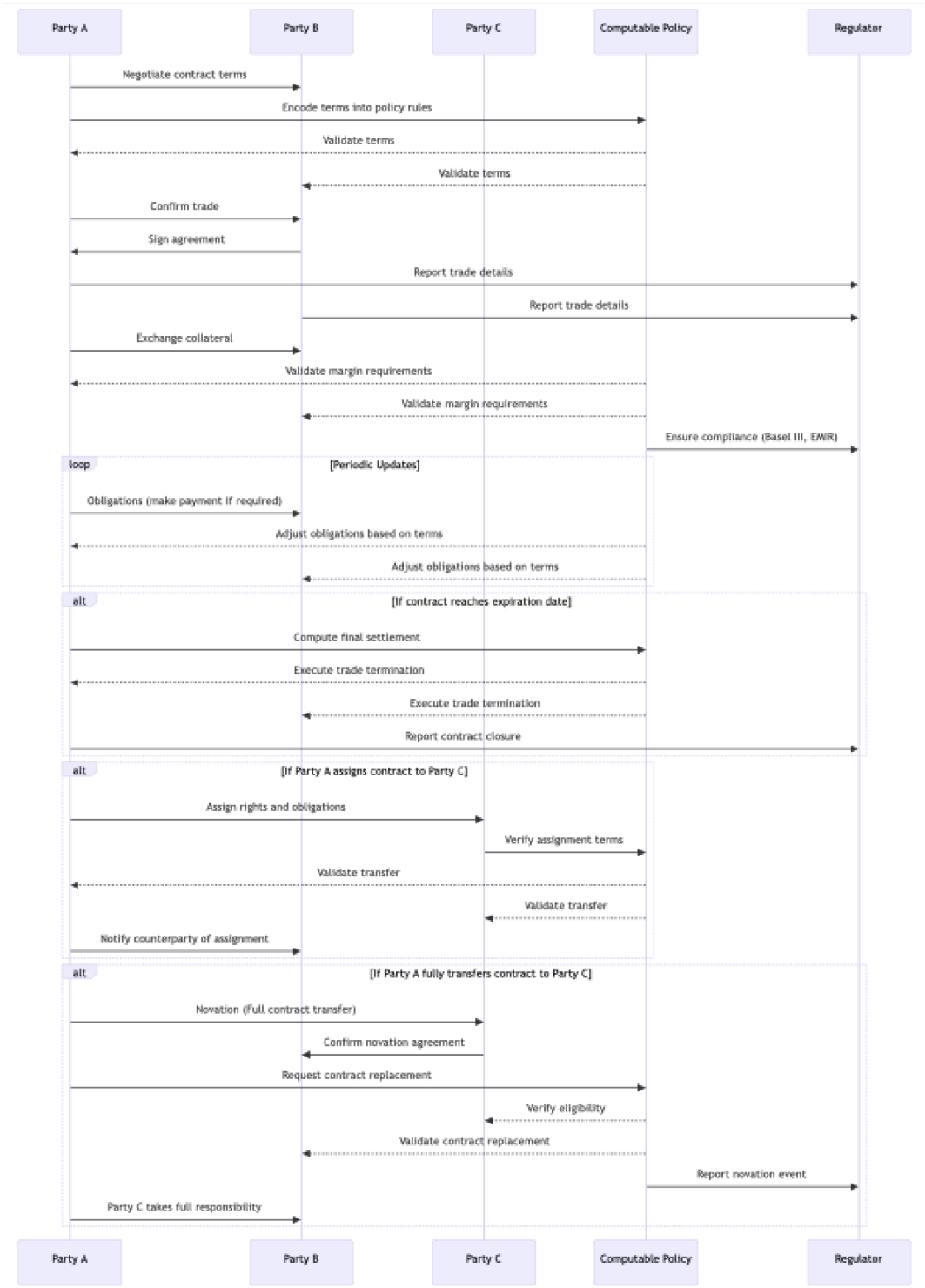

Figure 1 Sequence Diagram of Policy Decisions in OTC Financial Derivatives Contracts.

- Embed compliance rules from regulations such as Basel III, Dodd-Frank, or European Market Infrastructure Regulation (EMIR) to ensure legal adherence in an auditing operational scenario.
- The problem of enforcement usually involves fund transfer, imposing additional security challenges.

5.1.4.3 Industrial Safety Policies

In a company, especially but not limited to manufacturing companies, employees are required to follow *safety procedures* when operating on a production line, e.g., to make a maintenance intervention. Procedures exist to guide expert users to follow the expected process and ensure compliance. In parallel to operational procedures, *policies* and *guidelines* are usually provided to regulate the expected behaviours.

In this scenario, the concept of policy is interpreted as the commitment of one party to behave to match an expectation of another party: expectations may concern future states of affairs or future behaviour of the parties involved; the policy is used to track fulfillment and violation, rather than specific actions or steps to be executed. Therefore, by industrial safety policies we refer to the strategic aspect of safety (opposed to the operational aspect of safety procedures); in other words, within the same company the same policy (strategic safety) can be valid across different procedures (operational safety): for example, different lines within the same factory may need different operational steps due to the differences in the machines, but they must all be compliant to the same expected (strategic) safety behaviours.

Policy Management Life-cycle Steps. All steps, but especially intention, monitoring and auditing. For *intention*, *employees* are interested in checking if a specific behaviour is expected, mandatory, or prohibited, e.g., wearing protective equipment in specific contexts. For *execution/monitoring*, *operators* want to make sure that the production line is constantly in a safe state, while *controllers* are interested in checking that all employees behave safely to ensure the smooth operations of the line. For *auditing*, violations of safety behaviours must be collected to intervene and prevent future misbehaviours.

Computational problems. Mainly activation/evaluation and enforcement. When checking the policy to get informed about the allowed/prohibited behaviours, *activation/evaluation* takes place, to select all policies that apply to the specific state of the world (e.g., the specific activities going on in the factory) and to verify that the expected behaviours match the expectations and, in case they don't, indicate that there is a violation. When such a violation is identified, *enforcement* is needed to understand the consequences (e.g., corrective action, formal/official warning notice, etc.). Across those problems, some opportunities and challenges may emerge to apply *inference* and provide *explanations*: the factory "state of the world" may be multi-faceted, making it complex to identify the applicability of a policy (e.g., employee role, production line, type of intervention on the machinery, temporal and spatial context, etc.); at auditing time, repeated cases of safety policy violation by an employee may lead to specific sanctions or to the need to identify and apply corrective actions, e.g., training/retraining courses.

Desiderata/challenges. A policy language to fulfill the above scenario should take into account the following requirements:

- Ability to correctly formulate and check expected behaviours, especially in the case of concepts like safety that may be overlooked and considered "common sense" knowledge.

- Possibility to precisely define roles and their accountability/responsibility w.r.t. the strategic policy: in the case of safety, there are clear social and legal implications.
- Alignment between company-specific policies and legislation in force: policy specification needs to make sure that "local" permissions/prohibitions are not in contrast with laws.
- Interplay between (strategic) policies and (operational) procedures: if strict compliance with specific processes is crucial, policies should also include the procedural knowledge (see previous scenario); otherwise, policies must ensure that the expected behaviors are in line with operational indications.

5.1.4.4 Business Travel Expenses Reimbursement

Description. In any company where employees travel to perform part of their work, the need arises to manage and reimburse the travel expenses incurred by each person. While there are policies specific to each organization, there are usually documents explaining the rules for reimbursement, which may include a definition of eligible expenses, spending limits, and temporal deadlines for the reimbursement process management. An *employee* is interested in understanding if an expense can be reimbursed, sometimes even before making a purchase, while *responsible managers* and *financial departments* are interested in checking the correctness of reimbursement claims and acting upon them. Whenever a problem arises during the reimbursement process, the involved actors are also interested in understanding what happened and why, to perform remedial actions (policy enforcement).

Policy Management Life-cycle Steps. All steps, but especially intention, monitoring and auditing. For *intention*, employees are interested in checking if an expense will be eligible for reimbursement before making the payment. For *attempting/execution*, employees submit their claims for expense reimbursement and wait for their processing. For *monitoring*, the finance department may wish to supervise the various requests for reimbursement and have a list of actions to be carried out within specific time constraints. Similarly, employees would like to know the list of reimbursements they have to submit and the corresponding time scales. For *auditing*, all actors want an explanation on the expenses that were not approved and understand how to act about it, and managers specifically may also be keen to check the behaviours of their employees to identify the need or obligation of sanctions, as well as recognition for constant fulfillment.

Computational problems. Activation, evaluation and enforcement. (i) Checking the policies to get informed on their state of *activation*, to select all policies that regulate expense reimbursement and potentially those that apply to the specific state of the world (e.g., the specific type of expense a given employee incurred in). (ii) When checking for eligibility of reimbursement, the issue of *evaluation* arises to verify if the expenses match the expectations and, in case they do not, indicate that there is a violation. (iii) When such a violation is identified, *enforcement* is needed to understand what the steps to be followed are (e.g., reject an expense that overcame the spending limit, communicate to the user, etc.). In addressing these issues, some opportunities and challenges may arise for the application of *inference* techniques and for providing *explanations* for the reasoning performed: the user "state of the world" may be articulated and precise to understand the applicability of a policy and correctly "instantiate" it (e.g., employee level, type of expense, temporal and spatial context, etc.) especially in presence of a high number of exceptions and corner cases; with specific reference to trust, repeated cases of policy violation by a reimbursement requester may lead to specific sanctions or to the decrease of the trust other actors have w.r.t. them.

Desiderata/challenges. A policy language to fulfill the above scenario should take into account the following requirements:

- Bridge between the natural language version of the policies and potentially automated systems to process the computational version.
- Consider the dimension of trust, as trust can change between the parties when policies are evaluated and enforced.
- Consider the distinction between the deontic part of the policy (*regulative rules*, e.g., permissions/prohibitions/obligations/rights of expenses) and the descriptive/definition part of the policy (*constitutive rules*, e.g., what means that an expense is business-related, what means that an expense is excessive, etc.). There is, from the representative prospect, a continuum between those two parts (as policies may mix "schema and instances"), and Knowledge Graphs seem to be a fitting solution to cover their representation.
- Management of exceptions, especially those usually managed "by hand" on a case-by-case basis.
- Support the policy dynamics over time, as this kind of regulation may change unpredictably due to business, organizational, or legislative reasons.

5.1.4.5 Energy System Data Sharing

Description. In 2023, the UK Government commissioned a report to examine the case for a data-sharing infrastructure (DSI) for the UK energy system [3]. A follow-up assessment [12] makes the case for the urgent development of an MVP to explore the needs of influencing and impacted stakeholders in a diversified energy system for which a whole-system approach becomes essential for its effective management, in contrast to the current siloed, hierarchical structure. We consider operational issues such as data cleaning, etc., out of scope for this discussion, although data quality – and hence the policies governing it – pervade such a system and clearly embody risk to system function. However, here we focus on the many machine-processable policies that can facilitate the function and the evolution of the data-sharing infrastructure, in contrast to fundamentally fragile solutions that rely on top-down, regimented control with mandated representations.

Policy Management Life cycle Steps. All steps are critical for this scenario. *Intending* and *attempting* matter for a participant in the DSI so that both they and the DSI can be assured ahead of time that an action is compliant with the policies active at the time and that an action will not affect – as can best be determined – system integrity. *Monitoring* is a logical follow-on that observes (sequences of) actions to assess the continuing performance of the system as a whole and how individual actions are contributing (or not) to the achievement and maintenance of system goals. System operating actors may then, in line with additional policies, step in to alter the system trajectory and keep it within the desired behavioural envelope. Lastly *auditing* serves to process the record of participants' actions to contextualise individual actions against the bigger picture at the time, such as in the case of retrospective analysis of incidents and accidents, but also for the system operator to uncover behavioural patterns at scale that may indicate the need for policy revisions or additional policies to maintain system performance, and to carry out functions for the daily balancing market.

Computational problems. These are largely the same as in the other use cases, but in contrast to the definitions in Section 5.1.3 the state of the world is likely to be decentralised rather than a single data structure and hence partially observable for participants, while the system operator may have a complete but not necessarily up-to-date representation

of the state of the world. Policy generation will primarily be under the control of the system operator, although some features of the activation of a policy may subject to the requirements of the parties governed by a particular instantiation of a policy, for example, in the case of what features are shared and the privacy enhancing technology [15] to be used for sharing. Enforcement will be the responsibility of the system operator, using the monitoring and auditing mechanisms to obtain evidence of what happened when. Some aspects of enforcement may be enacted by system software actors, but others may transfer to human actors representing organisations participating in the DSI for resolution by human governance mechanisms.

Desiderata/challenges. Identified technical challenges [12] include a lack of common standards adoption by organizations in the sector, a lack of scalable infrastructure, and a prevalence of inflexible legacy systems. Associated cultural challenges [12] include perceived value of private data that inhibits sharing, concerns over the data quality of others, and potential embarrassment over an organization's own data quality.

At first sight, a regimented solution appears to offer simplified governance, but in reality, it pushes the governance burden on to

(a) data producers

(b) data consumers

(c) system maintenance.

The last is a hidden cost and a threat to evolution: over time, possibly even quite rapidly, the *one-size-governs-all* approach will meet an incompatible use case. The possible resolutions are to reject the use case, force the use case into the existing framework, or change the framework, but inertia will generally work against the last option. Thus, the challenge here is to embed sufficient flexibility in the policy framework such that is captures a space of acceptable policy solutions, which in turn implies the existence of over-arching meta-policies – these could be formal and represented in a policy language, or in natural language interpreted by humans, or a mix of both – that constrain actual policies, while themselves also being changeable to account for shifts in system requirements and participant values over time.

5.1.5 Conclusion

This report outlines a structured foundation for advancing the study of machine-readable norms and policies within knowledge graph-based AI systems. By formalizing the concept of Computational Policy Languages and introducing precise definitions for policy-related tasks – generation, activation, evaluation, and enforcement – we aim to enable rigorous reasoning about normative systems in complex, multi-agent socio-technical environments.

The proposed framework situates policy reasoning within a dynamic socio-technical system or *state of the world*, modeled as a knowledge graph. It also introduces the Policy Engine as a computational artifact to operationalize core reasoning tasks. These tasks are contextualized through four operational scenarios: *intending*, where agents assess policy compliance before acting; *attempting*, where actions are evaluated just prior to state changes; *monitoring*, which continuously evaluates ongoing behavior; and *auditing*, which retroactively analyzes actions against historical policy states.

The use cases examined in this report – spanning process conformance, financial contract execution, safety compliance, organizational governance, and energy data sharing – highlight the interdisciplinary nature of the problem space. They raise critical research challenges, such as handling temporal and deontic logic, reconciling declarative and procedural representations, modeling multi-party obligations, ensuring semantic alignment between natural language and formal policies, and integrating domain-specific ontologies and regulatory constraints.

Addressing these challenges requires a combination of formal methods, natural language understanding, semantic web technologies, and socio-legal modeling. Future research must also investigate the trade-offs between expressiveness and tractability, the integration of explanation and accountability mechanisms, and the adaptation of policy reasoning to decentralized and federated systems.

By consolidating key problems and illustrating their practical implications, this report invites further investigation into principled, interoperable, and context-aware policy languages – paving the way for a new generation of AI systems respectful of the self-determination of their actors.

References

1 Greta Adamo, Stefano Borgo, Chiara Di Francescomarino, Chiara Ghidini, Nicola Guarino, and Emilio M. Sanfilippo. Business process activity relationships: Is there anything beyond arrows? In Mathias Weske, Marco Montali, Ingo Weber, and Jan vom Brocke, editors, *Business Process Management Forum – BPM Forum 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings*, volume 329 of *Lecture Notes in Business Information Processing*, pages 53–70. Springer, 2018.

2 Ines Akaichi and Sabrina Kirrane. A comprehensive review of usage control frameworks. *Computer Science Review*, 56, 2025.

3 Arup, Energy Systems Catapult, and University of Bath. Digital spine feasibility study: exploring a data sharing infrastructure for the energy system, 8 2024.

4 Elisa Bertino, Piero Andrea Bonatti, and Elena Ferrari. Trbac: a temporal role-based access control model. In *Proceedings of the Fifth ACM Workshop on Role-Based Access Control*, RBAC '00, page 21–30, New York, NY, USA, 2000. Association for Computing Machinery.

5 Josep Carmona, Boudewijn F. van Dongen, Andreas Solti, and Matthias Weidlich. *Conformance Checking – Relating Processes and Models*. Springer, 2018.

6 Claudio Di Ciccio and Marco Montali. Declarative process specifications: Reasoning, discovery, monitoring. In Wil M. P. van der Aalst and Josep Carmona, editors, *Process Mining Handbook*, volume 448 of *Lecture Notes in Business Information Processing*, pages 108–152. Springer, 2022.

7 Laura Giordano, Alberto Martelli, and Daniele Theseider Dupré. Temporal deontic action logic for the verification of compliance to norms in asp. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, ICAIL '13, page 53–62, New York, NY, USA, 2013. Association for Computing Machinery.

8 Luis-Daniel Ibáñez, John Domingue, Sabrina Kirrane, Oshani Seneviratne, Aisling Third, and Maria-Esther Vidal. Trust, accountability, and autonomy in knowledge graph-based ai for self-determination. *Transactions on Graph Data and Knowledge*, 1(1), 2024.

9 Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian Padget, Terry R. Payne, Munindar P. Singh, Valentina Tamma, and Antoine Zimmermann. Governance of autonomous agents on the web: Challenges and opportunities. *ACM Transactions on Internet Technology*, 22(4):1–31, 2022.

10 Timotheus Kampik and Cem Okulmus. Expressive power and complexity results for signal, an industry-scale process query language. In Andrea Marrella, Manuel Resinas, Mieke Jans, and Michael Rosemann, editors, *Business Process Management Forum – BPM 2024 Forum, Krakow, Poland, September 1-6, 2024, Proceedings*, volume 526 of *Lecture Notes in Business Information Processing*, pages 3–19. Springer, 2024.

11 Florian Kelbert and Alexander Pretschner. Data usage control for distributed systems. *ACM Transactions on Privacy and Security*, 21(3):1–32, 2018.

12 Furong Li, Nigel Turvey, Lewis Dale, John Scott, Julian Padget, Isaac Flower, Jennifer R. Fitzpatrick, Nico Ostler, Rob Oldaker, and Simon Yeo. Do we need a data sharing infrastructure for the energy sector? *IET Smart Grid*, n/a(n/a).

13 Jaehong Park and Ravi Sandhu. The ucon-abc usage control model. *ACM Transactions on Information and System Security*, 7(1):128–174, 2004.

14 Qingtan Shen, Artem Polyvyanyy, Nir Lipovetzky, and Timotheus Kampik. Agent system event data: Concepts, dimensions, applications. In Wolfgang Maass, Hyoil Han, Hasan Yasar, and Nicholas J. Multari, editors, *Conceptual Modeling – 43rd International Conference, ER 2024, Pittsburgh, PA, USA, October 28-31, 2024, Proceedings*, volume 15238 of *Lecture Notes in Computer Science*, pages 56–72. Springer, 2024.

15 United Nations (Statistics Division). *The United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*, 2.

16 Wil M. P. van der Aalst, Marlon Dumas, Chun Ouyang, Anne Rozinat, and Eric Verbeek. Conformance checking of service behavior. *ACM Trans. Internet Techn.*, 8(3):13:1–13:30, 2008.

17 Thomas Vogelgesang, Jessica Ambrosy, David Becher, Robert Seilbeck, Jerome Geyer-Klingeberg, and Martin Klenk. Celonis PQL: A query language for process mining. In Artem Polyvyanyy, editor, *Process Querying Methods*, pages 377–408. Springer, 2022.

18 Petia Wohed, Wil M. P. van der Aalst, Marlon Dumas, Arthur H. M. ter Hofstede, and Nick Russell. On the suitability of BPMN for business process modelling. In Schahram Dustdar, José Luiz Fiadeiro, and Amit P. Sheth, editors, *Business Process Management, 4th International Conference, BPM 2006, Vienna, Austria, September 5-7, 2006, Proceedings*, volume 4102 of *Lecture Notes in Computer Science*, pages 161–176. Springer, 2006.

5.2 Towards Computer-Using Personal Agents

Piero A. Bonatti (University of Naples, IT)
John Domingue (The Open University – Milton Keynes, GB)
Anna Lisa Gentile (IBM Almaden Center – San Jose, US)
Andreas Harth (Fraunhofer IIS – Nürnberg, DE)
Olaf Hartig (Linköping University, SE)
Aidan Hogan (University of Chile – Santiago de Chile, CL)
Katja Hose (TU Wien, AT)
Ernesto Jiménez-Ruiz (City St George's, University of London, GB)
Deborah L. McGuinness (Rensselaer Polytechnic Institute – Troy, US)
Chang Sun (Maastricht University, NL)
Ruben Verborgh (Ghent University, BE)
Jesse Wright (Open Data Institute – London, GB)

Abstract. Computer-Using Agents (CUA) enable users to automate increasingly-complex tasks using graphical interfaces such as browsers. As many potential tasks require personal data, we propose Computer-Using Personal Agents (CUPAs) that have access to an external repository of the user's personal data. Compared with CUAs, CUPAs offer users better control of their personal data, the potential to automate more tasks involving personal data, better interoperability with external sources of data, and better capabilities to coordinate with other CUPAs in order to solve collaborative tasks involving the personal data of multiple users.

5.2.1 Introduction

Advances in Generative AI, and particularly Large Language Models (LLMs), have led to the recent release of various *Computer-Using Agents* (*CUAs*) that automatically operate a user's computer on their behalf. These agents use multimodal capabilities to interact with graphical interfaces via simulated mouse and keyboard inputs. Prominent commercial examples of CUAs include OpenAI's Operator, Google's Jarvis, and new functionalities in Anthropic's Claude.

Potential use cases for CUAs involve personal and often sensitive data, such as credit card details for purchases, passport numbers for flight booking, addresses for deliveries, and allergy information for dinner reservations. While modern browsers sometimes store personal data to autocomplete web forms, CUAs could additionally take context into account (e.g., selecting between a home or work address, depending on the purchase) and go beyond simple autocompletion.

Passing personal data to CUAs raises valid concerns about how such data might be (mis)used. Currently, OpenAI's Operator invokes a *takeover mode* for tasks involving sensitive data (e.g., log-in or payment details): the user is required to fill the details in manually [25]. Such measures target users' concerns about how their personal information will be used by CUAs. OpenAI themselves state that Operator is *"still learning, evolving and may make mistakes"* [25]. There are thus many open questions relating to the use of personal user data by CUAs.

Conversely, there are many potential benefits to users if CUAs are empowered with personal data. CUAs could autofill forms with personal data for users in a context-aware and potentially generative manner, automating a tedious task. CUAs could potentially enrich personal data with public data to better solve tasks. The CUAs of multiple users could negotiate to achieve a mutually beneficial result based on their users' personal context and preferences.

Towards providing users more oversight over their personal data while enabling higher levels of automation for complex tasks, we propose **Computer-Using Personal Agents** (**CUPAs**): *a Computer-Using Agent (CUA) that has controlled access to a structured repository of private information relating to a user.* This concept is illustrated in Figure 2. Specifically, we propose to instantiate the repository as a *Personal Knowledge Graph* (*PKG*) representing the user's personal data, which would facilitate the specification by users on how the CUA can access and use these data. This PKG can collect more personal data over time, with policies also evolving to reflect the user's fluctuating trust in the system [2]. Looking further forward, one can then imagine a scenario where CUPAs interact with websites and services via the underlying Web APIs instead of through a vision model, where CUPAs can assist in recommendations and negotiations based also on interactions with similar users and/or users' CUPAs.

We provide a road-map towards realising this vision of CUPAs, discussing what is achievable now with current technology, and what gaps must be addressed via further research and development.

5.2.2 User Scenario

Sam is expecting Jane over for dinner at 8pm, and is thinking about preparing Thai food. Sam is pre-diabetic, while Jane has a shellfish allergy. Sam requests that his CUPA to generates some suggestions of Thai recipes for the occasion. Consulting Sam's schedule, the CUPA recommends to filter recipes requiring more than an hour to prepare based on

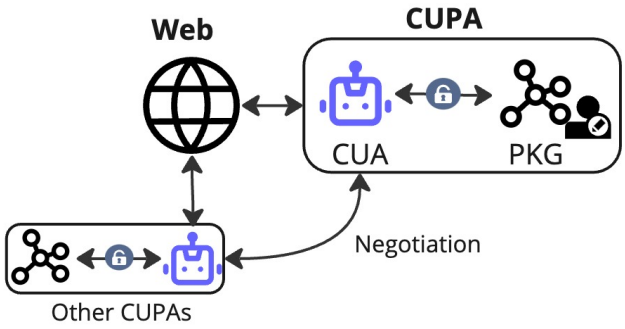

Figure 2 Computer-Using Personal Agent.

when he finishes work and his commute time. Sam agrees, and the CUPA starts to retrieve and present shellfish-free recipes of Thai food that are quick to prepare. Upon consulting external sources of nutritional information and recipes on the Web, the CUPA flags some recipes as being above the postprandial glucose threshold recommended by Sam's doctor (<180 mg/dL), or as having high glycaemic indices (>70).

Sam asks his CUPA to find out what recipe Jane might like. As Sam and Jane are friends, Sam's CUPA can send the candidate recipes to Jane's CUPA to see what she might like. Jane's CUPA suggests to avoid some recipes that include coriander (listed in some recipes as cilantro), which Jane hates. Sam's agent enforces his glucose thresholds and flags ingredients with high glycaemic indices, using external food and recipe knowledge graphs (e.g., the FoodKG [16]) to find the alternative ingredients. Of the remaining options, Sam's agent suggests a tofu green curry recipe that catches Sam's eye. Since the recipe is flagged for having a high glycaemic index (78), the agent asks Sam if he might consider replacing jasmine rice with cauliflower rice as a healthier option. Sam refuses the substitution as it is a special occasion.

Sam requests his CUPA to order the ingredients from a local supermarket. Since green bell peppers are out of stock, the CUPA suggests to replace them with yellow bell peppers. Sam agrees, and the CUPA prepares the order for delivery to Sam's home address, soliciting Sam's confirmation. Later that night, Sam and Jane enjoy their dinner of Thai green curry. After Jane leaves, Sam suffers some slight heartburn. He requests his CUPA to order antacids and additionally registers the fact that green curry dishes may cause Sam heartburn for future reference.

5.2.3 State of the Art

Personal data play an increasingly important role in modern life [24, 6]. Early works [18] characterise such data based on the *concept of six senses*: owned by me, about me, directed to me, sent by me, already experienced by me, and useful to me. More restrictive definitions only include data created by the individual [3], or that the individual cares for/about [11, 12].

Much literature has been dedicated to Personal Information Management systems (PIMs), which deal with the acquisition, organisation, maintenance, retrieval and sharing of personal data [19]. Notable PIM technologies include blockchain systems [36, 14], systems capturing user behaviour on multiple user devices [22, 26], and end-user prototypes [9, 24, 20]. Personal Knowledge Graphs (PKGs) [7, 8, 27] further apply a graph abstraction to personal data, opening up possibilities for declarative access policies, deductive inference, and integration with external Knowledge Graphs.

Towards taking fuller advantage of such data, AI-powered agents show much promise, particularly those that can automate tasks currently performed by the user. Robotic Process Automation (RPA) [29, 13] automates interactions with human interfaces. However, such approaches are hard-coded, brittle to changes in the interface, and incapable of generalising to unseen interfaces. Conversely, AI-based agents are capable of learning and generalising. LLM-based agents have been proposed to operate in diverse environments using recursion, feedback, and careful prompt engineering [33]. Such LLM-based agents are capable of solving computer tasks – despite the limited reasoning capabilities in LLMs [21] – paving the way for CUAs such as Operator [25].

Regarding works unifying LLM-based agents with PKGs, AGENTIGraph [34] heads in this direction, but rather focuses on question answering. Closer to the idea of CUPAs is Charlie: a brief proposal by Berners-Lee [4] on combining LLM-based agents with PKGs instantiated by Solid pods using Semantic Web standards. This proposal, and the user scenario presented previously, echo the (yet unrealised) vision laid out by Berners-Lee et al. [5] for the Semantic Web 24 years ago. Wright [31] presents a "discuss then transact" model of LLM-interaction in support of this vision for LLM-based personal agents that represent legal entities.

5.2.4 Added Value

Societal and legal debates on personal data emphasise *protection* from the harm that they could inflict, and understandably so. Yet people voluntarily exchange personal data with others in their every-day lives in the pursuit of mutual benefit. People can decide to leverage more personal data, or different kinds of personal data, to achieve a desired outcome. For instance, patients might prefer to share fitness-tracker data with their doctor if this improves their treatment, or consumers might want to divulge allergies and dietary needs to streamline online shopping and avoid nasty surprises.

A dangerous assumption is that companies are more capable of distilling value from people's personal data than the people the data describe. A company certainly has advantages over individuals in this respect, such as the ability to aggregate over a great many users. But personal data about a particular individual in isolation has much greater potential to empower that individual than a company they interact with, especially when the individual is coached by an agent such as a CUPA. CUPAs representing different parties could even negotiate a better outcome for *all* parties involved.

Considering the added value of CUPAs, and more generally of providing AI-based agents access to personal data, we highlight:

Multi-dimensional negotiation. CUPAs can help users to strike sweet-spots between multiple dimensions, such as the cost and duration of multi-hop flights, the deliciousness and healthiness of meal options, etc.

Increased granularity. Humans struggle to negotiate on a fine-grained level, and may thus prefer broad policies that reduce cognitive load (e.g., to always accept all cookies) [32]. CUPAs can help to reach fine-grained agreements that improve outcomes and honour party preferences.

Improved risk/reward assessment. CUPAs can help users simulate and analyse a variety of hypothetical data exchange scenarios, and warn users of a particular risk, for example that the supermarket – if informed of a condition of a severe allergy – could sell this information to third parties, leading to an increase in life assurance premiums.

Auditing and follow-up. CUPAs could automatically perform audits to assess whether the data were treated as agreed during the negotiation process, evaluate the benefit to the user, and improve for future interactions.

Such added value is, of course, dependent on the value outweighing the potential harms caused. This can be addressed via AI alignment, which ensures that artificial intelligence systems act in accordance with human intentions, values, and societal norms. It involves *outer alignment*, where an AI's objectives accurately reflect human goals, and *inner alignment*, ensuring learned behaviours remain aligned in novel scenarios. Machine-readable policies on how personal data from the PKG can or should be used by the AI-based agent can also help to avoid harm. Representing personal data as PKGs allows standards such as the Open Digital Rights Language (ODRL) [17] and policy engines implementing formal semantics [15] to specify and automate the processing of policies about how personal data are used, in what contexts, and under what conditions.

5.2.5 CUPA Capabilities

Computer-using personal agents must be able to *interact with diverse websites and APIs*. This allows them to book flights and hotels, search for job openings, and even schedule appointments. Moreover, they must possess the ability to *interact with other such agents*, such as coordinating travel arrangements with a travel agent or collaborating with a financial agent to manage expenses.

In addition to being able to *generate and adapt content* (e.g., personalised summaries and creative text), a computer-using personal agent must be able to *combine private data* from the user's personal knowledge graph (PKG) with external information. For example, when searching for a new apartment, the agent should combine the user's preferred neighbourhood from their PKG with data from real estate websites and local amenities databases to find the most suitable options. When utilising the knowledge stored within the PKG, the agent must also be able to *adapt the knowledge from the PKG for the current task*. For instance, when filling out a job application form, the agent should selectively use information from the user's CV and work history stored in the PKG, tailoring the presentation to the specific requirements of each application. This adaptability is crucial for ensuring that agent actions are relevant and effective in the given context.

CUPAs must continuously *collect and enrich user information* to effectively assist them. This involves gathering data from various sources, including interactions with websites and APIs, user inputs, and external sources. By continuously *learning about user preferences*, these agents can personalise their assistance, such as recommending travel options that align with the user's preferences or suggesting recipes that cater to specific dietary restrictions or tastes. However, it is also crucial for such agents to *avoid learning one-off or irrelevant patterns*, for example, to assume that Sam will always suffer heartburn after eating Thai food and should thus avoid it.

Computer-using personal agents must exhibit a high degree of autonomy. They should ideally *act maximally autonomously*, including the ability to *proactively anticipate and address user needs*. For example, an agent could proactively remind users of upcoming appointments or suggest relevant articles based on their recent reading history. However, this autonomy must always be balanced with the ability to *be guided and controlled by the user*, allowing users to provide instructions, adjust preferences, and maintain control over agentic actions.

While acting largely autonomously, it is crucial that a computer-using personal agent *acts in alignment with the user*, ensuring that tasks are completed as desired. This is essential in scenarios like recipe searches where the agent must accurately reflect dietary restrictions and preferences. Moreover, such an agent should always act in the user's interests, even when *dealing with potentially conflicting goals*. For example, an agent helping a user plan a trip should consider factors like budget, travel time, and personal preferences, even if these factors

may lead to a slightly more expensive or less convenient option. The agent should avoid *acting in an unethical or illegal manner* even if it potentially maximises a users immediate interests, e.g., via tax evasion.

To maintain user trust and ensure responsible behaviour, it is also crucial that agents *do not overstep bounds*, respecting user privacy and only acting within explicitly granted permissions. Finally, the repeated offering of *clear explanations of all actions* will aid in the fostering of trust and allow users to understand and verify agent behaviour.

5.2.6 Technical Challenges

The aforementioned desired capabilities for CUPAs, based on our vision of a trusted, accountable and largely autonomous agent acting with personal data for user benefit, raises a number of technical challenges.

Accountability and Liability In the case of undesired, illegal, or unethical acts involving CUPAs, it is important to determine who – or what – is responsible, who should be held accountable, and where the liability lies.

Explainability, Traceability, and Provenance Provenance techniques are required to trace and explain how personal and external data led to specific answers or actions being derived or carried out by the CUPA. These provenance techniques would need to support diverse data models, machine learning processes, user inputs and policies.

Data Interoperability Data interoperability is a key challenge towards implementing CUPAs. Being able to draw on and integrate more sources of data will improve the CUPAs performance. This is particularly challenging for new sources discovered on the fly.

Inter-Agent Communication, Negotiation and Coordination Agents must communicate effectively in the context of multi-agent systems to achieve shared goals, requiring both a shared conceptual understanding and a means of encoding and decoding messages [30]. The same challenge applies to networks of CUPAs who coordinate to solve a particular set of goals for users.

Security, Privacy, and Policies The sensitive nature of data processed by a CUPA calls for security, privacy, and usage control mechanisms, and the ability of the CUPA to understand and correctly apply the access/usage/action control policies of the user. In some countries, this would even be a legal requirement (e.g., under GDPR in the E.U.).

Trust, Delegation, and Action Control Achieving agent autonomy requires trust modelling, delegation mechanisms, and structured action control policies [28]. Trust models must be adaptable to different contexts, from rigid policies applicable in government agencies to more flexible, reputation-based approaches for personal agents [10].

User-in-the-Loop CUPAs will require input, guidance, permission and confirmation from the user. But to increase automation, the CUPA must avoid unnecessary interactions with the user. This creates the challenge of *when* to call upon the user, and how.

Self-Improvement The CUPA should leverage its experience with the user in order to improve the services it provides over time, leading to greater automation, and actions/results that better benefit the user. This raises questions about how such a history can be captured, represented, stored and leveraged.

Self-Determination and Alignment

5.2.7 Roadmap

We envisage that moving from the current state of the art to fully addressing the above technical challenges will occur in three stages. These levels represent varying degrees of trust, accountability and autonomy.

CUAs enhanced with personal data In the first instance, we foresee extensions of CUAs – in the style of OpenAI's Operator [25] in a commercial setting and Agent-E [1] in a research setting – such that they use a PKG in order to access knowledge personal to the user. This would safely enable higher levels of automation, obviating the need to pass control back to the user in scenarios of the user's choosing that involve personal data.

Web-aware CUPAs CUAs currently rely on existing browser implementations to render an HTML page and then make use of vision models to interact with the page. An agent could rather observe HTTP requests made to a particular website, as well as the HTML forms present on a page, to invoke requests and actions directly via HTTP.

Networks of CUPAs We envision networks of CUPAs interacting in order to complete tasks involving multiple users. This may involve structured service descriptions [23], or a mix of natural language and structured communication per a "discuss then transact" model [31] whereby agents use natural language to first negotiate about a transaction they wish to perform, and then confirm this transaction using structured data.

5.2.8 Conclusion

Computer-Using Agents (CUAs) have the potential to transform how users interact with their computers, their browsers and amongst themselves. Not having access to personal data limits such interactions. Giving CUAs unfettered access to the personal (and most sensitive) data of a user seems unwise, as does providing CUAs no access to personal data. We thus argue for CUPAs as a configurable middle-ground, where a Personal Knowledge Graph (PKG) is used to represent, store and control access to the personal data of the user. As a starting point, the data that a user fills into web forms can be captured in the PKG, and enriched by an AI-based agent. These data can then be used, if the user so wishes, by CUAs to automate further tasks. In a next step, CUPAs can learn to interact with websites via HTTP APIs rather than though visual interfaces. Finally, we envisage further into the future a network of CUPAs collaborating to address users' tasks.

References

1 Tamer Abuelsaad, Deepak Akkil, Prasenjit Dey, Ashish Jagmohan, Aditya Vempaty, and Ravi Kokku. Agent-e: From autonomous web navigation to foundational design principles in agentic systems. *arXiv preprint arXiv:2407.13032*, 2024.

2 Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1–30, 2024.

3 Ofer Bergman and Steve Whittaker. *The science of managing our digital stuff*. MIT Press, 2016.

4 Tim Berners-Lee. Charlie Works. Design Issues, `https://www.w3.org/DesignIssues/Works.html`, 2025.

5 Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

6 Kean Birch, D. T. Cochrane, and Callum Ward. Data as asset? the measurement, governance, and valuation of digital personal data by big tech. *Big Data and Society*, 8, 2021.

7 Prantika Chakraborty, Sudakshina Dutta, and Debarshi Kumar Sanyal. Personal research knowledge graphs. In *WWW 2022 – Companion Proceedings of the Web Conference 2022*, pages 763–768. Association for Computing Machinery, Inc, 4 2022.

8 Prantika Chakraborty and Debarshi Kumar Sanyal. A comprehensive survey of personal knowledge graphs, 11 2023.

9 Amir Chaudhry, Jon Crowcroft, Heidi Howard, Anil Madhavapeddy, Richard Mortier, Hamed Haddadi, and Derek McAuley. Personal data: Thinking inside the box. *Aarhus Series on Human Centered Computing*, 1:4, 2015.

10 Ray Chen, Fenye Bao, and Jia Guo. Trust-based service management for social internet of things systems. *IEEE transactions on dependable and secure computing*, 13(6):684–696, 2015.

11 Amber L. Cushing. PIM as a caring: using ethics of care to explore personal information management as a caring process. *Journal of the Association for Information Science and Technology*, 74(11):1282–1292, 2023.

12 Amber L. Cushing and Páraic Kerrigan. Personal information management burden: A framework for describing nonwork personal information management in the context of inequality. *Journal of the Association for Information Science and Technology*, 73:1543–1558, 11 2022.

13 Diogo António da Silva Costa, Henrique São Mamede, and Miguel Mira da Silva. Robotic Process Automation (RPA) adoption: a systematic literature review, 6 2022.

14 Benedict Faber, Georg Michelet, Niklas Weidmann, Raghava Rao Mukkamala, and Ravi Vatrapu. Bpdims:a blockchain-based personal data and identity management system. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2019-Janua:6855–6864, 2019.

15 Nicoletta Fornara, Víctor Rodríguez-Doncel, Beatriz Esteves, Simon Steyskal, and Benedict Whittam Smith. ODRL Formal Semantics, May 2024.

16 Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness, and Mohammed J. Zaki. FoodKG: A semantics-driven knowledge graph for food recommendation. In Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois, and Fabien Gandon, editors, *The Semantic Web – ISWC 2019*, pages 146–162, Cham, 2019. Springer International Publishing.

17 Renato Iannella and Serena Villata. ODRL Information Model 2.2, Feb 2023.

18 William Jones. *The future of personal information management, part 1: Our information, always and forever*. Morgan & Claypool Publishers, 2012.

19 William P Jones and Jaime Teevan. *Personal information management*, volume 14. University of Washington Press Seattle, WA, 2007.

20 Varvara Kalokyri, Alexander Borgida, and Am lie Marian. YourDigitalSelf: a personal digital trace integration tool. *International Conference on Information and Knowledge Management, Proceedings*, pages 1963–1966, 2018.

21 Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language Models can Solve Computer Tasks. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 – 16, 2023*, 2023.

22 Jiangxu Lin and Meng Wang. PKG : A Personal Knowledge Graph for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11â • fi15, 2022, Madrid, Spain*, volume 1, pages 3334–3338. Association for Computing Machinery, 2022.

23 David Martin, Massimo Paolucci, Sheila McIlraith, Mark Burstein, Drew McDermott, Deborah McGuinness, Bijan Parsia, Terry Payne, Marta Sabou, Monika Solanki, Naveen Srinivasan, and Katia Sycara. Bringing semantics to web services: The OWL-S approach. In Jorge Cardoso and Amit Sheth, editors, *Semantic Web Services and Web Process Composition*, pages 26–42, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

24 Richard Mortier, Jianxin Zhao, Jon Crowcroft, Liang Wang, Qi Li, Hamed Haddadi, Yousef Amar, Andy Crabtree, James Colley, Tom Lodge, Tosh Brown, Derek McAuley, and Chris Greenhalgh. Personal data management with the databox: What's inside the box? In *Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking*, CAN '16, page 49–54, New York, NY, USA, 2016. Association for Computing Machinery.

25 OpenAI Team. Introducing Operator. OpenAI Blog `https://openai.com/index/introducing-operator/`, published 2025-01-23, accessed 2025-01-29, 2025.

26 Markus Schröder, Christian Jilek, and Andreas Dengel. A Human-in-the-Loop Approach for Personal Knowledge Graph Construction from File Names. In *Knowledge Graph Construction*, volume 3141. CEUR Workshop Proceedings, 2022.

27 Martin G Skjæveland, Krisztian Balog, Nolwenn Bernard, Weronika Łajewska, and Trond Linjordet. An ecosystem for personal knowledge graphs: A survey and research roadmap. *AI Open*, 5:55–69, 2024.

28 Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. Authenticated delegation and authorized ai agents. *arXiv preprint arXiv:2501.09674*, 2025.

29 Wil M.P. van der Aalst, Martin Bichler, and Armin Heinzl. Robotic process automation. *Business and Information Systems Engineering*, 60:269–272, 8 2018.

30 Michael Wooldridge. *An introduction to multiagent systems*. Wiley, 2009.

31 Jesse Wright. Here's Charlie! Realising the semantic web vision of agents in the age of LLMs. *CoRR*, abs/2409.04465, 2024.

32 Jesse Wright, Beatriz Esteves, and Rui Zhao. Me want cookie! Towards automated and transparent data governance on the Web, 2024.

33 John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 – 16, 2023*, 2023.

34 Xinjie Zhao, Moritz Blum, Rui Yang, Boming Yang, Luis Márquez Carpintero, Mónica Pina-Navarro, Tony Wang, Xin Li, Huitao Li, Yanran Fu, Rongrong Wang, Juntao Zhang, and Irene Li. AGENTiGraph: an interactive knowledge graph platform for LLM-based chatbots utilizing private data, 2024.

35 Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment. *Philosophical Studies*, November 2024.

36 Guy Zyskind, Oz Nathan, and Alex Sandy Pentland. Decentralizing privacy: Using blockchain to protect personal data. *Proceedings – 2015 IEEE Security and Privacy Workshops, SPW 2015*, pages 180–184, 2015.

5.3 Evaluation of AI Systems

Paul Groth (University of Amsterdam, NL)
Michel Dumontier (Maastricht University, NL)
Michael Cochez (VU Amsterdam, NL))
Fajar Ekaputra (Vienna University of Economics and Business, AT)
Monica Palmirani (University of Bologna, IT)

Abstract. The evaluation of AI systems is central to fostering trust and ensuring accountability. A systematic assessment of such systems offers insights into a model's strengths and limitations regarding its performance on specific tasks. Furthermore, rigorous evaluation reveals how well the model generalizes across different scenarios, handles uncertainty, and adheres to ethical standards. Traditional AI systems evaluation focuses on the benchmarking approach using task-based evaluation metrics concerning ground truth [2].

However, such evaluations are fraught with challenges due to the lack of benchmark datasets, difficulties in creating gold standards, and the complexity of assessing new problems and domains. In this working group, we outline these challenges and propose innovative strategies for evaluating AI systems.

5.3.1 Discussed Problems

Evaluating complex AI systems presents unique challenges. These challenges have been highlighted in recent meta-reviews of evaluation failures across machine learning systems, such as those discussed by Liao et al. [1], emphasizing issues like implementation variations, overfitting, and metrics misalignment. Some of the key challenges are in the following:

1. Lack of Benchmark Datasets: Creating gold standards and benchmark datasets for new problems and domains is a difficult task. Even harder is to create a benchmark that is such that when a systems performs well on it, then the system can be applied on a context broader than the one the benchmark was derived from.
2. Dynamic Contexts: For example, in the legal domain, changing legislation and environments can render benchmark datasets obsolete.
3. Complexity of Domains: Each domain, such as public policy or clinical trials, has unique contexts that complicate standard evaluation approaches.

In addition to these challenges, several common pitfalls can lead to misleading results when evaluating AI systems. Some issues are data-related: *Incomplete training and test splits* can cause unexpected distribution shifts between training and deployment, leading to poor generalization. *Missing or incorrect ground truth data* can result in inconsistent or unbalanced datasets, often due to insufficient annotators or domain expertise. Another critical problem is *data leakage*, where unintended overlap between training and test data skews evaluation results, giving an inflated sense of model performance.

Beyond data, process-related issues also pose a threat to evaluations. *Confirmation bias* occurs when researchers selectively use data or metrics favouring their AI system, leading to overly optimistic assessments. Some systems may even be designed to exploit specific evaluation metrics or datasets – known as *gaming the system* – rather than demonstrating true generalization. Moreover, *incomplete comparisons* arise when only favourable metrics are highlighted, ignoring aspects where the system may underperform.

5.3.2 Possible Approaches

Based on our experience with the field and our discussion during the seminar, we identified the following initial strategies to address the issues presented in the previous section.

1. **Backtesting** involves training AI systems on data from one specific time period or geographic region and testing them on a different time period or region. For example, a model trained on European data may be evaluated on data from the U.S. to assess its adaptability across different contexts. Commonly used in causal discovery, this method evaluates robustness across temporal or spatial shifts.
2. **Lifelong Benchmarking.** Instead of relying solely on static benchmarks, this evaluation approach continuously updates benchmarks with new datasets and annotations while reusing previously validated models to test these new datasets. This dynamic approach ensures evaluations remain relevant over time.
3. **Reproducibility Testing** assesses whether an AI system's outcomes can be consistently replicated across different domains, datasets, or implementation approaches, highlighting the generalizability of the system.
4. **Sandboxing and Red Teaming**
   - **Sandboxing** implies deploying the AI system in a controlled environment, recording results and user interactions. It helps assess how the system's behaviour in different scenarios, records performance metrics, and gathers user feedback.
   - **Red Teaming** involves a dedicated team of experts to intentionally "break" the system, identifying weaknesses that standard evaluation might overlook and logging qualitative feedback.
5. **Coherence Testing** evaluates whether the AI system produces consistent and logically coherent results across similar queries, emphasizing internal reliability. Inconsistent results may indicate issues, such as inadequate training or overfitting to specific datasets/tasks.
6. **Evidence-Based Evaluation** measures the system's ability to provide well-supported, documented evidence for its outputs. The quality and amount of evidence are proxies for performance and key indicators of the system's reliability and trustworthiness.
7. **Explanation-Based Evaluation.** Explanations bridge an AI system's internal reasoning and human understanding, providing a basis for trust. High-quality explanations are clear, logical, and relevant, while poor-quality explanations can erode trust and perpetuate biases.

5.3.3 Connection to overarching topics

The seminar centred on four central aspects of KG-based AI systems: transparency, trust, accountability, and self-determination. One question is posed for each of these aspects; here, we reflect on these questions and illustrate how evaluation approaches can help address them.

Transparency *What is required to ensure that the data fueling and the inner workings of AI artefacts are transparent?*

Evidence-based Evaluation supports transparency by evaluating the capability of AI systems to provide well-supported, verifiable, and documented evidence for their outputs. The evaluation approach allows users to understand the basis of the system's decisions and verify the sources of information used. *Reproducibility Testing* also contributes to transparency by ensuring that results can be replicated across different domains and implementations.

Trust *What are the key requirements for an AI system to produce trustable results?*
Coherence Testing is essential for trust to demonstrate an AI system's ability to provide consistent and logically sound outputs. Other approaches can also improve the trust in an AI system, e.g., *Sandboxing and Red Teaming* by stress-testing the system under controlled conditions, identifying vulnerabilities, and ensuring robustness before full deployment, and *Lifelong Benchmarking* through dynamic updates of evaluation benchmark, making it relevant over time.

Accountability *How can AI be made accountable for its decision-making?*
Evidence-Based Evaluation enforces accountability by requiring AI systems to justify their decisions with verifiable data. Other approaches enhance accountability, e.g., *Explanation-Based Evaluation* that allows stakeholders to review how decisions are made, and *Red Teaming*, which helps identify weaknesses and biases in AI systems. Note that the latter delegates accountability to the stakeholders.

Self-Determination *How can users and citizens maintain self-determination when using or being the subject of KG-based AI systems?*
Transparency-focused methods such as *Evidence-Based Evaluation* empower users by giving them insight into how AI systems work, helping them to reason and make an informed decision whether and how to use AI system results. Other approaches also contribute to self-determination, e.g., *Lifelong Benchmarking*, which ensures models remain aligned with evolving user needs, societal norms, and ethical standards.

5.3.4 Conclusion

Traditional benchmarking approaches to AI evaluation have significant limitations, particularly for complex and dynamic domains. By integrating innovative strategies such as explanation-based evaluation, backtesting, and lifelong benchmarking, we can overcome these challenges and develop more robust, accountable neurosymbolic AI systems. As next steps, we see a need for:

- **Exploring evaluation strategies.** Implement and evaluate possible evaluation approaches, such as backtesting protocols for knowledge-graph-based AI systems and develop standardized protocols for red teaming and sandboxing.
- **Relating evaluation to architectures.** Investigation of evaluation strategies in the context of neurosymbolic AI design pattern architectures [3] to enhance their interpretability and robustness.

References

1 Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable AI. 10(1):147–159, 2022.
2 Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086, 2024.
3 Michael van Bekkum, Maaike de Boer, Frank van Harmelen, André Meyer-Vitali, and Annette ten Teije. Modular design patterns for hybrid learning and reasoning systems. *Appl. Intell.*, 51(9):6528–6546, 2021.

5.4 Trust and Accountability in Knowledge Graph-Based AI for Self Determination: Building World Models in Formal Representations using LLMs

José Manuel Gómez-Pérez (Expert.ai – Madrid, ES)
Marko Grobelnik ((Jozef Stefan Institute – Ljubljana, SI)
Ryutaro Ichise (Institute of Science Tokyo, JP)
Manolis Koubarakis (University of Athens, GR)
Heiko Paulheim (University of Mannheim, DE)
Daniel Schwabe (Rio de Janeiro, BR)

5.4.1 Motivation

Large Language Models (LLMs) have demonstrated impressive capabilities in understanding and generating natural language, including extracting structure and meaning from unstructured text. However, their internal world knowledge and reasoning processes remain largely opaque. In this working group, we investigated the hypothesis that it is possible to construct explicit, formal world models from the latent knowledge encoded in LLMs and textual sources – models that could serve as stable, inspectable, and verifiable representations of a domain.

This endeavor aligns with long-standing goals in artificial intelligence and knowledge engineering, particularly the knowledge acquisition bottleneck, which has historically hampered the development of large-scale, formalized knowledge bases[2]. Inspired by systems such as Cyc, which utilized microtheories to manage and modularize world knowledge, we explored whether modern LLMs could support a similar, but largely automated, process of world model construction – potentially generating formal representations in languages such as OWL, SWRL, or Prolog directly from text.

The core motivation rests on several anticipated benefits of this approach:

- **Improved reasoning:** Formal models constrain the solution space and enable more accurate, consistent, and explainable inferences than directly querying an LLM.
- **Explainability and transparency:** Representing knowledge symbolically allows inspection of both structure and content, addressing key concerns in trustworthy AI.
- **Reproducibility and auditability:** Formalized ontologies can be versioned, verified, and reused in a way that black-box LLM behavior cannot.
- **Human-AI collaboration:** LLMs can support domain experts and knowledge engineers in structuring, extending, and refining ontologies and rule-based systems.

As a working hypothesis, we posited that such LLM-bootstrapped world models – constructed through a combination of natural language interpretation and formal reasoning – could eventually lead to systems that are both expressive and interpretable, supporting multi-hop reasoning, comparative validation, and context-sensitive explanation.

This investigation was grounded in a concrete domain: AI regulation and compliance, specifically centered on the EU AI Act[1]. This domain combines rich natural language source texts, evolving legal definitions, and complex reasoning requirements, making it an ideal setting to evaluate the feasibility and value of constructing formal world models with LLM assistance.

5.4.2 Method and Experiments

Our approach combined classical knowledge engineering (KE) methodologies with the affordances of modern Large Language Models (LLMs), using the latter as active participants throughout the modeling pipeline. The LLMs used during our experimentation spanned across the spectrum of models in the Google Gemini and Open AI families. However, soon we settled for models with capabilities involving inference-time scaling[3], such as o1 and Gemini-2.5, which demonstrated to be better suited for complex reasoning tasks like this.

Rather than relying on LLMs solely as generators of content or answers, we treated them as collaborators in an iterative, human-in-the-loop process of constructing a formal world model from regulatory text. The process was inspired by standard KE lifecycles and followed an adapted knowledge engineering lifecycle, leveraging LLM capabilities at each stage:

1. **Domain and Scope Definition** We selected the EU AI Act as our primary source, focusing on its provisions for high-risk AI systems. This emerging legislation offers a structured yet complex body of natural language, with deep implications for the classification, deployment, and oversight of AI systems. This domain was chosen based on group expertise, real-world relevance, and the presence of rich, non-trivial natural language source material. We narrowed our focus to Title III and Annex III of the AI Act, which provide definitional content, classification rules, technical requirements, and compliance procedures. The textual structure of the act itself is quite complex. We first asked the LLM to analyze its structure and identify the portions relevant to high-risk systems and compliance, using these sections to drive world model extraction.
2. **Competency Question Generation** To frame the modeling effort, we generated a set of competency questions – queries that the resulting world model should be able to answer. These ranged from high-level regulatory assessments (e.g., "Is this product considered high-risk under the AI Act?") to accountability and governance inquiries (e.g., "Who is responsible for ensuring compliance?"). We created approximately 25 such questions, some crafted manually, and others generated by prompting LLMs such as ChatGPT or Gemini with high-level modeling intents. These questions helped clarify modeling scope and purpose, and provided a foundation for validating the emerging models.
3. **Requirements Gathering and Ontology Design** Using LLMs, we identified relevant regulatory texts, conceptual elements, and ontologies. This included mining the AI Act for definitions, roles, obligations, and processes, as well as exploring existing semantic vocabularies (e.g., LegalRuleML, FOAF, PROV-O) for potential reuse. Prompts were used to generate OWL ontologies in Turtle syntax, as well as Prolog-style FOL rules. LLMs effectively recommended languages (e.g., OWL, SWRL, Prolog), modeling strategies, and external ontologies, showcasing multi-representational flexibility by shifting between different formalisms as needed.
4. **Ontology Construction and Extension** Initial top-down ontologies were produced by feeding full or partial text of the AI Act to LLMs. Later iterations involved incremental refinements based on new examples or competency questions. We explored structuring models into reusable fragments, akin to microtheories, to enhance modularity. However, incremental modeling proved brittle: LLMs struggled to maintain consistency across iterations, often introducing duplication, inconsistency, or reference drift. Prompt sensitivity remained high, requiring careful tuning.
5. **Instance-Level Annotation** We selected Waymo One, an autonomous driving platform, as a representative AI system. LLMs were used to annotate product descriptions with formal instance data. Prompts requested annotation using previously generated

ontological classes. This exercise revealed key challenges: LLMs frequently introduced new namespaces or redundant concepts instead of reusing prior structure, underscoring limitations in consistency and reuse.

6. **Validation and Demonstration** Although full reasoning tests were out of scope during the seminar, we emphasized competency-query-based evaluation – ensuring that the formal model could correctly classify, explain, or reject assertions about specific AI systems based on regulatory criteria. Initial validations were performed using both OWL reasoners and Prolog engines. These early-stage demonstrations illustrated how symbolic reasoning could support competency question answering, such as "Why is this AI system considered high-risk?" or "Who holds compliance responsibility in this scenario?"
7. **Reflection and Iteration** Throughout the process, we embraced an iterative, bi-directional workflow – moving from text to formalism and back again, with LLMs supporting each translation. This round-trip modeling paradigm allowed for rapid prototyping and exploration of design alternatives. LLMs played multiple roles: advisor, translator, editor, explainer, and generator. Their ability to contextualize text, propose structured models, and convert between representational forms enabled fluid movement between informal and formal levels of abstraction.

5.4.3 Lessons Learned

Our experiments yielded a range of insights into the strengths, weaknesses, and emerging design patterns associated with using LLMs for world model construction. These lessons fall broadly into two categories: capabilities and affordances, and challenges and limitations.

5.4.3.1 Capabilities of LLMs in Knowledge Engineering

LLMs proved valuable collaborators in a variety of modeling tasks, particularly in the early phases of formalization:

- **Bootstrapping structured representations:** LLMs effectively proposed taxonomies, relations, and axioms from unstructured legal text. With the right prompting, they could output OWL in Turtle syntax or Prolog-style logic programs.
- **Generating competency questions:** When guided appropriately, LLMs generated high-quality, domain-relevant competency questions, helping to clarify modeling scope and purpose.
- **Advising on tools and methodologies:** LLMs could recommend languages (e.g., OWL, SWRL, Prolog), suggest modeling strategies, and identify relevant external ontologies.
- **Multi-representational flexibility:** LLMs easily shifted between different formalisms, such as transforming an OWL ontology into SWRL rules, or converting plain text into logic-based representations.
- **Prototyping and iteration:** LLMs supported rapid prototyping of models and allowed quick exploration of design alternatives – essential for an exploratory setting like a Dagstuhl Seminar.

A key observation was that LLMs were most effective when deployed in a hybrid approach, mixing: **top-down** modeling from source texts and domain concepts, and **bottom-up** enrichment via instance-level annotations and use-case reasoning. This interplay allowed for richer and more grounded models – though not without difficulty.

5.4.3.2 Challenges and Limitations

Despite their impressive capabilities, LLMs exhibited several recurring limitations that hindered more robust modeling workflows:

- **Incremental modeling is brittle:** LLMs struggled to maintain consistency across iterations. Changes to a previously generated ontology often led to duplication, inconsistency, or regression.
- **Reference drift:** LLMs frequently failed to reuse previously defined classes, properties, or namespaces, even when explicitly instructed to do so.
- **Prompt sensitivity:** Small changes in prompt structure often led to significantly different outputs. Prompt engineering required careful tuning and could not be reliably abstracted.
- **Implicit vs. explicit knowledge:** When asked to extract axioms from text, LLMs frequently filled in gaps with inferred or assumed knowledge, blurring the boundary between source-derived content and background knowledge.
- **Provenance and traceability:** It was difficult to track which parts of the generated model were based on the original source text versus LLM inference or hallucination.
- **Toolchain fragility:** Integration with formal modeling tools (e.g., OWL editors, Prolog reasoners) exposed limitations in both syntax fidelity and model completeness.

These challenges underscored the need for better support for human-in-the-loop verification, robust referencing, and systematic iteration when using LLMs in formal modeling tasks.

5.4.4 Open Questions and Future Work

While our initial experiments confirmed the potential of LLMs to assist in formal world model construction, they also raised a number of open questions and future research directions. These span technical, methodological, and conceptual dimensions.

5.4.4.1 Methodological Open Questions

- **How should human–AI collaboration be structured in formal modeling?** The process we followed was ad hoc but promising. More systematic methodologies are needed to orchestrate interactions between domain experts, knowledge engineers, and LLMs – particularly in maintaining consistency across iterations.
- **What does "correctness" mean in this context?** Unlike traditional logic programs or ontologies, LLM-generated models may reflect probabilistic or context-dependent interpretations. Determining when a world model is "good enough" or "trustworthy" remains an open question.
- **What role should competency questions play in validation?** Competency questions helped frame the modeling task, but their use as a systematic evaluation mechanism is underdeveloped. Future work could define benchmarks or test suites to assess model coverage and consistency.

5.4.4.2 Technical Challenges and Research Directions

- **Consistency and Reuse** A key technical challenge is ensuring that LLMs consistently reuse previously defined structures – classes, properties, namespaces – across sessions and modeling phases. This requires better prompt design, memory management, and possibly external constraint injection.

- **Iterative, verifiable modeling** Formal models need to evolve incrementally without introducing contradictions or drift. Future research could explore interfaces where LLMs propose edits that are checked by reasoners or validated against typed assertions.
- **Toolchain integration** Bridging the gap between LLMs and formal reasoning systems remains a challenge. Improved APIs, modeling environments, and reasoning-aware LLM prompts could help operationalize neurosymbolic modeling pipelines.
- **Round-trip modeling and explanation** One of the most exciting possibilities is the idea of a round-trip loop between symbolic models and LLMs: models extracted from LLMs are refined by humans and used in turn to improve LLM outputs or explanations. Supporting this requires mechanisms for traceability, grounding, and goal-sensitive reasoning.

5.4.4.3 Broader Implications

- **From world models to world views** If LLMs can generate multiple, diverging models from the same text (as our experiments suggest), this opens up the possibility of comparing different stakeholder perspectives – e.g., between a regulator, a provider, and an affected user. This introduces new opportunities for accountability, argumentation, and normative reasoning.
- **Explainability as emergent behavior** Formal models on their own do not constitute explanations – but they provide the scaffolding for tailored, stakeholder-specific narratives. Future work could explore how formal world models and LLM-based natural language generation can jointly support explainable AI in high-stakes domains like regulation.

References

1 European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024, 2024.
2 Cogan Shimizu and Pascal Hitzler. Accelerating knowledge graph and ontology engineering with large language models. *Journal of Web Semantics*, 85:100862, 2025.
3 Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv*, arXiv:2408.03314, 2024.

5.5 Knowledge Graph Ecosystems

Sandra Geisler (RWTH Aachen, DE)
James A. Hendler (Rensselaer Polytechnic Institute – Troy, US)
Philipp D. Rohde (TIB – Hannover, DE)
Aisling Third (The Open University – Milton Keynes, GB)
Maria-Esther Vidal (TIB – Hannover, DE)

5.5.1 Introduction

In this working group, we based our initial discussions on the concepts of Knowledge Graph Ecosystems (KGEs) inspired by the paper by Geisler et al.[1] In the paper, knowledge graph ecosystems describe the context and crucial aspects impacting the creation, updates, and usage of knowledge graphs by a sextuple. This includes the data sources populating the KG,

the ontologies to enrich and structure the data in the KG, mappings between the ontology and the data, constraints, which enable consistency and quality checks, the KG which is "living" in the KGE and subject to changes and usage, and finally a log, tracing the operations executed on the KG and corresponding events. Further, the paper delineates life cycles and life cycle steps which structure and order the operations on the KG. A life cycle step can be, e.g., the creation or the update of a KG, or a query or analysis on the data of the KG. Formally, it is defined also as a sextuple with a service executing the step, actors and roles involved in the step, as well as requirements, constraints, and needs which impact the life cycle step. While this work is a crucial step towards the formalization and therefore automatization and verifiability of operations on and around KGs, it only targets individual KGs. However, today many applications require sharing data, i.e., the querying and analysis of multiple distributed KGs across organizations, domains, and even countries. Hence, we need to not only take into account contexts and life cycles of single KGEs, but also networks or federations of KGEs.

In exploring the concepts of federated and decentralized Knowledge Graph Ecosystems (KGEs), this work inherently ties into the foundational themes of trust, accountability, and autonomy discussed in [2]. Note that each of these themes are very broad concepts with significant variation in their concrete application; for a fully flexible model of KGEs, it would be a mistake to attempt to narrow trust, accountability, and autonomy down to very specific or technical definitions only covering a subset of their broader meanings. We rather leave the framework open to accommodate different concrete understandings of these terms in different scenarios or use cases. Trust in federated KGEs is established through shared ontologies and robust data verification processes, similar to the way blockchain's transparency and immutability enhance trust by ensuring the traceability and provenance of data. This aligns with decentralized KGEs where trust mechanisms, like reputation systems, are vital for ensuring knowledge integrity. Accountability is addressed through verifiable contributions and collaborative data management in federated networks, paralleling the mechanisms in [2], such as compliance-checking and provenance tracking that uphold data integrity through blockchain technology. Autonomy is a defining feature of decentralized KGEs, where stakeholders maintain control over their data and operations, consistent with the use of decentralized infrastructures in [2], which empower individuals through self-sovereign identities and personalized data management. By integrating these technologies, both papers emphasize the potential to advance KGEs in ways that are reliable, user-centric, and conducive to collaborative innovation.

In the following, we will present the examples for such networks and federations of KGs discussed in the seminar. Based on these examples, we will delineate the different types of KGEs we derived from the examples and discuss challenges and requirements for the different types. These challenges and requirements led to a reformulation and extension of the definitions for KGEs and the corresponding life cycles, which we will sketch subsequently. Finally, we will give an outlook on future work sparked by the results of the seminar.

5.5.1.1 Use Case 1: Digital Product Passports

Transparency of product constitution, production processes, and supply chains, especially in the light of sustainability, energy consumption, and circular economies, motivates the need for a Digital Product Passport (DPP). A DPP serves as a comprehensive digital record that comprises data about a product's components and life cycle, including its materials, manufacturing processes, usage, and recycling potential. DPPs are promising transparency and traceability, ensuring that all stakeholders – from manufacturers to consumers – can

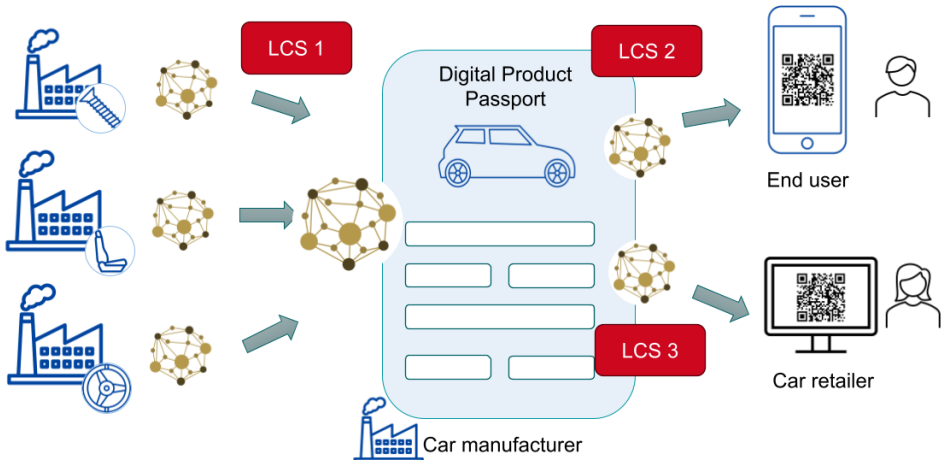

Figure 3 A KGE of a Digital Product Passport.

make informed decisions while fostering accountability. Additionally, a DPP can facilitate compliance with regulatory requirements, improve waste management, and support efficient product recalls or repairs. However, the data for the DPP needs to be retrieved from the involved stakeholders. Suppliers, logistics, as well as manufacturers, can provide information regarding a product or parts of it. Furthermore, usage behavior, repairs, and other events related to the product user and impacting the product quality and integrity can also be part of a DPP.

Figure 3 shows the example of a DPP infrastructure for a car. Suppliers of screws, seats, or steering wheels all may maintain data about the part they are providing. It is assumed that product data for the DPP is stored in KGEs located at the suppliers. To enable a DPP, the suppliers extract and provide a part of this data to the car manufacturer who in our example is assumed to maintain the DPP of the car, i.e., integrate the data from the suppliers into a bigger KGE. We assume further that different stakeholders may have different views on the KGE of the product maintained by the manufacturer. Thereby, subsets in the form of views can be extracted from the KG of the DPP and are provided to consuming applications, such as mobile applications for end users or web applications for retailers. In this scenario, the stakeholders follow a common goal, namely to provide a DPP. In terms of trust, the car manufacturer trusts that the suppliers provide the DPP information of their part, i.e., they are accountable for the provision of correct, accurate, and complete data. They do not have a high autonomy as they are dependent on the car manufacturer as their customer and follow its rules.

5.5.1.2 Use Case 2: A Dagstuhl Seminar

Opposed to the example of a DPP, in a Dagstuhl Seminar multiple stakeholders (the participants and organizers) attend with one or more individual goals for the overall seminar, but do not necessarily share a common goal. They all have their own "internal KGE" and may update their knowledge in interaction with other participants. Further, they build smaller working groups, where the group follows common group goals and at the same time can contribute to the overall goal of the seminar. Hence, there are multiple KGEs (one for each participant, one for a group, and one for the seminar), which have their own life cycles, "interact" with each other and exchange knowledge between them. Between the participants,

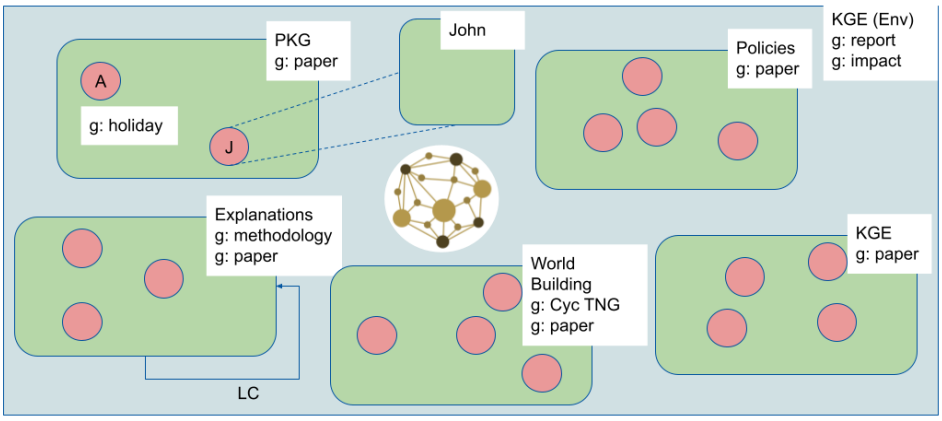

Figure 4 The KGEs of a Dagstuhl Seminar.

trust needs to be established as a basis to update their KGEs in the interaction. There is a very high autonomy as no participant is dependent on another and in consequence also no accountability.

5.5.2 Types of KGE Networks

We define a KGE network as a set of KGEs, which are interconnected by relationships.

We have considered each of the examples above in terms of how they demonstrate different assumptions about the three pillars of trust, accountability, and autonomy, and how they relate to each other. By configuring KGE networks with these assumptions, we can see that network phenomena such as *federation*, decentralization, and so on, emerge. In the DPP example, low autonomy of stakeholders, represented by the common specification set by the car manufacturer, and trust ensured via contractual accountability, looks very like a *federated* scenario; a Dagstuhl Seminar's high autonomy of participants, with low accountability and high trust, resembles more a *decentralized* scenario.

A relationship is established between two KGEs to enable knowledge exchange between them. Based on the observations from the above two and also other examples, we distinguish the following two types of KGE networks:

1. **Federated KGE Networks:** In a federated KGE network all stakeholders have a KGE and share a common goal towards which they exchange knowledge.
2. **Decentralized KGE Networks:** In this type of networks stakeholders also have a KGE, but may have one or more individual goals. Additionally, they can also share common goals with one or more other stakeholders.

These definitions are in terms of entities formally defined in the KGE framework of [1]: goals, knowledge graphs, etc. We show below how these can nonetheless provide models of these trust, accountability, and autonomy scenarios. This provides an insight into how to model scenarios involving the foundational pillars of [2] without overcommitting to narrow definitions of them, and we argue that this flexibility coupled with the representation formalism of KGEs is essential for AI ecosystems which work for such fundamentally human concepts.

5.5.3 Requirements

Understanding the requirements inherent in different use cases of Knowledge Graph Ecosystems (KGEs) provides insight into the diversity of challenges and solutions necessary to support various applications. This section describes the requirements for the two primary types of KGE networks – Federated and Decentralized – each characterized by unique aspects of collaboration, autonomy, trust, and lifecycle management.

5.5.3.1 Federated KGE Networks

The case of Digital Product Passports (DPP) serves as an archetype for federated KGEs where stakeholders share a collective objective and are inherently dependent on each other's data integrity and contributions. The requirements for such setups include:

1. **Shared Semantics and Ontologies:** Stakeholders must align on a common ontology to ensure interoperability across KGEs. This includes shared vocabularies and data models that describe products consistently.
2. **Data Integration and Consistency:** The integration of data from multiple suppliers entails rigorous consistency checks and validation mechanisms to maintain data integrity across the ecosystem.
3. **Accountability and Verifiable Contributions:** Federated systems necessitate mechanisms for traceable contributions to ensure that each actor can be held accountable for the data they provide, thus fostering trust in the overall system.
4. **Role-Based Access and Governance:** Clearly defined (e.g., suppliers, manufacturers, regulators) must be supported, with correspondingly specific access rights and responsibilities. Governance models ensure appropriate access and usage of the ecosystem by different stakeholders.
5. **Lifecycle Synchronization:** Operations across constituent KGEs must be coordinated, especially for updates and validations, to maintain a consistent state and ensure collaborative workflows proceed smoothly.

5.5.3.2 Decentralized KGE Networks

In contrast, decentralized KGE networks as exemplified by collaborative scenarios like scientific research or Dagstuhl Seminars, emphasize autonomy and less rigid interactions. The requirements for decentralized structures include:

1. **Alignment Mechanisms:** Decentralized networks require lightweight ontology alignment methods to facilitate knowledge exchange without enforcing uniformity, thus respecting participants' autonomy.
2. **Trust-Based Interactions:** Establishing trust through provenance and reputation mechanisms is critical in decentralized setups where actors operate independently but collaborate based on shared interests or goals.
3. **Asynchronous and Independent Lifecycles:** Supports for asynchronous updates and independent lifecycle management are needed, allowing participants to operate under varied goals and timelines without disrupting mutual interactions.
4. **Autonomy Preserving Policies:** Each participant should have the autonomy to apply local policies and operate independently, with knowledge exchange only occurring in mutually beneficial circumstances.
5. **Conflict Resolution and Negotiation Frameworks:** Mechanisms to resolve conflicts or inconsistencies are essential where differing goals or data interpretations may arise from the diverse participants.

5.5.3.3 Challenges and Opportunities

The interplay between these requirements in both federated and decentralized networks highlights several overarching challenges:

1. **Interoperability:** Striking a balance between shared understanding and local autonomy requires robust semantic alignment techniques applicable across diverse contexts.
2. **Trust and Accountability:** Crafting ecosystems where trust and accountability are clear yet flexible enough to accommodate decentralized decision-making and governance remains a pivotal challenge.
3. **Efficiency and Scalability:** Efficient data management, processing, and query execution across distributed KGEs need to support scalability while preserving data quality and provenance.

Addressing these requirements not only involves technical innovation in areas like ontology matching, knowledge integration, and lifecycle management but also necessitates careful consideration of legal, ethical, and organizational factors to foster successful KGE deployments. As such, continued research in synchronizing lifecycles, implementing robust alignment frameworks, and enhancing-based governance will pave the way for more adaptable and scalable KGE networks.

5.5.4 Extension of the KGE Concept

The original concept of a Knowledge Graph Ecosystem (KGE), as introduced by Geisler et al. [1], models the fundamental operational components required to manage the creation, evolution, and analysis of a knowledge graph. Formally, a KGE is defined as a sextuple: $KGE = (D, O, M, DC, KG, L)$, where:

- D is a set of **data sources**, each with schema $\theta(ds)$ and instances $\alpha(ds)$.
- O is the **ontology**, a logical theory describing the domain using a structured vocabulary.
- M is a set of **mappings** linking D to O through semantic assertions.
- DC is a set of **domain constraints** ensuring data consistency and quality.
- KG is the resulting **knowledge graph**, built from D using M under O.
- L is a **log** of lifecycle steps, tracking changes and ensuring traceability.

Limitations. While this formalization provides a solid foundation for the structured management of a single evolving knowledge graph, it assumes:

1. A single, self-contained knowledge graph (KG) under centralized control.
2. A fixed and isolated lifecycle of operations for one ecosystem context.
3. No explicit support for collaboration or interactions across multiple independent KGEs.

However, real-world applications increasingly rely on *networks of interconnected KGEs* operated by diverse stakeholders. These scenarios require knowledge to be collaboratively constructed, exchanged, and reused across institutional, geographical, or organizational boundaries. Consider the previously defined use cases:

- **Digital Product Passports (DPPs)**: A manufacturer aggregates data from multiple suppliers, each maintaining their own KGE to describe components or materials. The manufacturer composes a global KGE that integrates these subgraphs. All actors are accountable for their data and contribute toward a shared objective.
- **Collaborative Scientific Research**: Researchers maintain local KGEs to organize data and hypotheses. As part of a collaborative research initiative, they align concepts and exchange selected data views, but retain autonomy in their lifecycle and modeling.

To enable such scenarios, we extend the original definition of KGE by introducing a recursive structure where ecosystems can consist of multiple interrelated KGEs – each with its own lifecycle, goals, and context.

Extended Definition. A Knowledge Graph Ecosystem is now defined as:
$KGE = (KGM, \{KGE_1, \ldots, KGE_n\}, LC, G, C)$, where:
- $KGM = (D, O, M, DC, KG, L)$ is the underlying **knowledge graph model**.
- $\{KGE_1, \ldots, KGE_n\}$ is a (possibly empty) set of **nested or interacting KGEs**.
- LC is a **lifecycle**, formally modeled as a partially ordered structure of lifecycle steps, each representing operations or analyses on the KG.
- G is a set of **goals** pursued by the ecosystem.
- C is the **context**, capturing domain-specific and regulatory constraints or assumptions.

This extended definition allows the modeling of composite or hierarchical KGEs, enabling both *federated* and *decentralized* structures.

Federated KGEs. A KGE is **federated** if it aggregates multiple KGEs that share:
- **Common goals and context**: $G \subseteq \bigcap_{i=1}^{n} G_i$ and $C \subseteq \bigcap_{i=1}^{n} C_i$.
- **Shared concepts**: $O \models \bigcup_{i=1}^{n} O_i$ and $KG \subseteq \bigcup_{i=1}^{n} KG_i$.
- **Shared data**: $D \subseteq \bigcup_{i=1}^{n} D_i$.

This setup supports alignment and accountability among actors collaborating toward a unified goal, such as regulatory compliance or industry standards.

Use Case 1 (DPP) is an example of a federated KGE, where suppliers and manufacturers contribute to a shared infrastructure, follow a common goal, and are held accountable for the correctness of their data.

Decentralized KGEs. A KGE is **decentralized** if it includes autonomous KGEs that:
- Maintain **individual goals and contexts**, i.e., $G \cap \bigcup G_i \neq \emptyset$ or $C \cap \bigcup C_i \neq \emptyset$.
- Communicate via **alignments**: for each pair (i, j), there exists an ontology O_{ij} such that $O_i \models_a O_{ij}$ and $O_j \models_a O_{ij}$, where $\models_a$ denotes entailment up to alignment.
- Possibly share some **data elements**.

Such configurations reflect looser, trust-based networks – e.g., research collaborations or international knowledge exchanges – where autonomy is preserved, and accountability is local. **Use Case 2 (Dagstuhl Seminar)** exemplifies this structure, with researchers individually maintaining their KGEs while participating in collaborative groups with overlapping knowledge and trust-based data sharing.

Summary. Extending the KGE model from single knowledge graphs to federated and decentralized ecosystems enables:
- Formal modeling of complex, distributed knowledge infrastructures.
- Explicit lifecycle tracking across ecosystem components.
- Representation of autonomy, trust, and alignment across diverse actors.

These extensions are essential to support knowledge-centric systems that operate across organizational, geographic, and technical boundaries.

5.5.5 Challenges and Future Work

Extending the KGE concept to support federated and decentralized ecosystems introduces new research challenges that go beyond those of traditional or standalone knowledge graphs. These challenges stem from the need to enable collaboration, preserve autonomy, ensure accountability, and foster trust across heterogeneous, interlinked systems.

1. Semantic Alignment in Heterogeneous Ecosystems. While classical KGEs already face challenges in integrating heterogeneous data, these are further exacerbated in federated and decentralized settings. Each participating KGE may use distinct ontologies, vocabularies, and constraints, making alignment a prerequisite for knowledge exchange. Federated settings assume shared semantics; decentralized ones require ontology alignments that preserve local autonomy while supporting partial interoperability. Ontology matching, cross-KGE mappings, and alignment reasoning are required to maintain interoperability and trust.

2. Trust and Accountability in Distributed Architectures. As KGEs become interlinked, accountability must be clearly defined across organizational and jurisdictional boundaries. Federated KGEs require all participants to commit to shared goals and contexts, making it possible to define collective accountability [2]. Decentralized KGEs, by contrast, emphasize autonomy – participants operate independently but must still be trusted sources. Establishing trust involves formalizing provenance, defining verifiable contributions, and enabling transparency in lifecycle actions. Ecosystem-wide logging mechanisms and policy-compliant data usage protocols are needed to support accountability and traceable decisions.

3. Modeling Autonomy and Interactions. In decentralized KGEs, each participant may pursue different goals or operate under distinct regulatory or ethical frameworks. Maintaining autonomy requires the ability to define local policies, ontologies, and lifecycle models – while still enabling interaction through lightweight semantics and alignments. Ensuring that knowledge exchange does not infringe on local autonomy calls for formal contracts, soft alignments, and partial knowledge views. Such mechanisms allow for trust-based collaboration while upholding the principle of self-determination [2].

4. Lifecycle Across KGEs. Federated and decentralized KGEs introduce the challenge of coordinating lifecycles across independently evolving components. In federated settings, lifecycle steps (e.g., updates, validations) must be synchronized to ensure consistent outcomes. In decentralized settings, looser coordination must support asynchronous updates and versioning. Future work should explore distributed lifecycle management models, including temporal consistency, event-driven propagation, and local override mechanisms.

5. KGE Validation and Provenance. Validation across KGEs requires understanding how constraints defined in one ecosystem apply to others. This is particularly challenging when KGE components evolve independently or when only partial views are exchanged. Provenance models must capture not only source alignment but also how knowledge was transformed or aligned across systems. Explainability – especially for outputs from KG-based AI systems – requires tracing decisions back to source KGEs and validating their integrity.

6. Explainability and Repeatability in KG-Based AI. As KGEs serve as the backbone for AI systems, the ability to explain inferences and reproduce results is crucial for building trust. This requires detailed provenance, clear alignment semantics, and interpretable reasoning paths – especially when hybrid approaches (e.g., involving LLMs) are used. Repeatability in decentralized settings must account for potential disappearance or evolution of external data sources. Mitigation mechanisms, such as fallback reasoning strategies or versioned snapshots, are essential to uphold the trustworthiness and reliability of AI outcomes.

7. Defining Role-Based Trust and Governance Models. With multiple stakeholders, each with distinct roles (e.g., data provider, knowledge builder, auditor, consumer), KGEs must support differentiated views, permissions, and responsibilities. Role-specific governance models and access policies must ensure that users interact with the ecosystem in ways that are transparent, justified, and traceable. This includes defining what actions are permissible, who is accountable for data changes, and how conflicts or inconsistencies are resolved.

8. Modeling Roles, Personas, and Task-Specific Interactions. KGEs serve a diverse set of stakeholders – including data stewards, domain experts, developers, auditors, and end users – each with distinct objectives, capabilities, and responsibilities. The extension to federated and decentralized KGEs further amplifies this diversity, requiring systems to explicitly model and support *role-specific interactions*. These roles influence how users contribute to, consume from, or govern the KGE. For example, a data provider must ensure schema and content quality, while a consumer needs trustable and explainable insights. Supporting personas involves:

- Designing user interfaces and APIs tailored to task-specific workflows.
- Implementing role-based access control and provenance-based accountability.
- Supporting transparency through lifecycle-aware logs and explainable outputs aligned with each persona's mental model.

Future work must focus on formalizing persona definitions, mapping tasks to lifecycle stages, and capturing the responsibilities of each actor. These models are essential to foster trust and usability in multi-actor KGEs, and to ensure that autonomy and accountability are respected across stakeholder boundaries.

Summary. Federated and decentralized extensions of KGEs offer a powerful abstraction for supporting complex, multi-actor knowledge infrastructures. However, realizing these systems at scale requires addressing a new set of challenges around interoperability, lifecycle synchronization, role-specific trust, and accountability. Future research must focus on:

- *Trust*: Establishing mechanisms for verifiable contributions, alignment justifications, and explainable inferences.
- *Accountability*: Formalizing roles, logs, and lifecycle provenance to assign responsibility and enable redress.
- *Autonomy*: Supporting local lifecycles, policies, and ontologies while enabling coherent interactions across KGEs.

These directions are essential to align the evolution of KGEs with principles outlined in the proposed EU AI Act [11], including transparency, reliability, and respect for autonomy.

References

1 Sandra Geisler, Cinzia Cappiello, Irene Celino, David Chaves-Fraga, Anastasia Dimou, Ana Iglesias-Molina, Maurizio Lenzerini, Anisa Rula, Dylan Van Assche, Sascha Welten, et al. From genesis to maturity: Managing knowledge graph ecosystems through life cycles. *Proceedings of the VLDB Endowment*, 2025.

2 Luis-Daniel Ibáñez, John Domingue, Sabrina Kirrane, Oshani Seneviratne, Aisling Third, and Maria-Esther Vidal. Trust, accountability, and autonomy in knowledge graph-based AI for self-determination. *TGDK*, 1(1):9:1–9:32, 2023.

[11] https://artificialintelligenceact.eu/the-act/

6 Conclusions

John Domingue (The Open University – Milton Keynes, GB, john.domingue@open.ac.uk)
Luis-Daniel Ibáñez (University of Southampton, GB, L.D.Ibanez@soton.ac.uk)
Sabrina Kirrane (Vienna University of Economics and Business, AT, sabrina.kirrane@wu.ac.at)
Maria-Esther Vidal (TIB – Hannover, DE, vidal@l3s.de)

In conclusion, this seminar stands as a pivotal gathering that convened researchers and industry partners from diverse backgrounds. Together, we explored the complexities, challenges, and advancements inherent in managing and leveraging knowledge graphs within real-world contexts. Spanning from technical considerations to social dimensions, we identified essential requirements, imperatives, and actionable strategies necessary to foster the development of a new generation of knowledge graph ecosystems.

Given the advent of generative AI and its demonstrated benefits when integrated with intricate data structures such as knowledge graphs, ensuring readiness across all facets of knowledge graph implementation is paramount. The convergence of knowledge graphs with emerging technologies presents novel avenues for advancing knowledge representation, reasoning, and applications. Our discussions underscored the significance of robust quality assessment mechanisms and stressed the importance of integrating human expertise and feedback loops throughout the knowledge graph lifecycle. From an educational standpoint, it is imperative for experts to disseminate their knowledge through educational programs tailored to different levels of learning and professional training. However, standardizing competencies across all levels is essential to ensure a uniform understanding of fundamental concepts among potential knowledge graph practitioners.

Participants

- Sören Auer
TIB – Hannover, DE
- Piero A. Bonatti
University of Naples, IT
- Irene Celino
CEFRIEL – Milan, IT
- Andrea Cimmino
Polytechnic University of Madrid, ES
- Michael Cochez
VU Amsterdam, NL
- John Domingue
The Open University – Milton Keynes, GB
- Michel Dumontier
Maastricht University, NL
- Fajar Ekaputra
Vienna University of Economics and Business, AT
- Nicoletta Fornara
University of Lugano, CH
- Sandra Geisler
RWTH Aachen, DE
- Anna Lisa Gentile
IBM Almaden Center – San Jose, US
- José Manuel Gómez-Pérez
Expert.ai – Madrid, ES
- Marko Grobelnik
Jozef Stefan Institute – Ljubljana, SI
- Paul Groth
University of Amsterdam, NL
- Peter Haase
Metaphacts – Walldorf, DE
- Andreas Harth
Fraunhofer IIS – Nürnberg, DE
- Olaf Hartig
Linköping University, SE
- James A. Hendler
Rensselaer Polytechnic Institute – Troy, US
- Aidan Hogan
University of Chile – Santiago de Chile, CL
- Katja Hose
TU Wien, AT
- Luis-Daniel Ibáñez
University of Southampton, GB
- Ryutaro Ichise
Institute of Science Tokyo, JP
- Ernesto Jiménez-Ruiz
City St George's, University of London, GB
- Timotheus Kampik
SAP Berlin, DE & Umeå University, SE
- George Konstantinidis
University of Southampton, GB
- Manolis Koubarakis
University of Athens, GR
- Deborah L. McGuinness
Rensselaer Polytechnic Institute – Troy, US
- Julian Padget
University of Bath, GB
- Monica Palmirani
University of Bologna, IT
- Heiko Paulheim
University of Mannheim, DE
- Philipp D. Rohde
TIB – Hannover, DE
- Daniel Schwabe
Rio de Janeiro, BR
- Oshani Seneviratne
Rensselaer Polytechnic Institute – Troy, US
- Chang Sun
Maastricht University, NL
- Aisling Third
The Open University – Milton Keynes, GB
- Ruben Verborgh
Ghent University, BE
- Maria-Esther Vidal
TIB – Hannover, DE
- Jesse Wright
Open Data Institute – London, GB