Report from Dagstuhl Seminar 25271

Policy Modeling and Reasoning in Sociotechnical Systems

Marina De Vos[*1], Nicoletta Fornara[*2], Munindar P. Singh[*3], Leon van der Torre[*4], and Jessica Woodgate[†5]

1 **University of Bath, GB.** cssmdv@bath.ac.uk
2 **USI – Lugano, CH.** nicoletta.fornara@usi.ch
3 **North Carolina State University – Raleigh, US.** mpsingh@ncsu.edu
4 **University of Luxembourg, LU.** leon.vandertorre@uni.lu
5 **University of Bristol, GB.** jessica.woodgate@bristol.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 25271 "Policy Modeling and Reasoning in Sociotechnical Systems". This seminar brought together researchers from academia and industry who are interested in studying the intersection between computer science, philosophy, logic, ethics, and law to discuss policy modelling and reasoning in a world where computers and humans need to work together. After lightning talks, two invited talks, and an open space topic gathering activity, we settled on four topics for deeper discussion in working groups, interspersed by primer talks from the various communities. The four topics were: 1) Concepts: What are the underlying aspects of this interdisciplinary field, and can they be defined consistently? 2) Agentic AI: How can we enable agents to interact and reason with human users through large language models? 3) Standardisation: How can we facilitate data sharing and compliance in international work with competing business interests? 4) Coevolution: How can we make sure that sociotechnical systems evolve with the societies they operate in? This report provides the abstracts of the talks, including participants' lightning talks, the two invited talks, and four primers, along with short reports from each working group detailing their discussions, including challenges and future opportunities.

Seminar June 29 – July 4, 2025 – https://www.dagstuhl.de/25271

2012 ACM Subject Classification Computing methodologies → Multi-agent systems; Computing methodologies → Philosophical/theoretical foundations of artificial intelligence; Information systems → World Wide Web

Keywords and phrases Multi-agent Systems, Norms and Values, Policy Modelling, Standardisation

Digital Object Identifier 10.4230/DagRep.15.6.132

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

Policy Modeling and Reasoning in Sociotechnical Systems, *Dagstuhl Reports*, Vol. 15, Issue 6, pp. 132–188
Editors: Marina De Vos, Nicoletta Fornara, Munindar P. Singh, Leon van der Torre, and Jessica Woodgate

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

1 Executive Summary

Marina De Vos (University of Bath, GB)
Nicoletta Fornara (USI – Lugano, CH)
Munindar P. Singh (North Carolina State University – Raleigh, US)
Leon van der Torre (University of Luxembourg, LU)

Introduction

A policy is a declarative basis for a decision by an individual or organisation. In computing, policies occupy the fruitful space between (social or legal) norms and (architectural or algorithmic) mechanisms. Specifically, a policy is an explicit knowledge-based machine-processable representation that guides the decision-making of an autonomous party. Previously, policies have been studied in computing from a primarily technical standpoint, e.g., in characterising access control in a database management system.

This Dagstuhl Seminar takes a fresh and comprehensive perspective on policies by viewing them as part of a sociotechnical system (STS) comprising intelligent agents and people. What makes policies particularly attractive in the modern milieu, with the rise of Artificial Intelligence (AI), is that they promise transparency, interpretability, and support for effective explanations as well as opportunities for making improvements to individual agents or the system. Yet major challenges remain in understanding how to (1) model policies (so they express stakeholder needs in emerging problem domains precisely, clearly, and succinctly), (2) provide a formal semantics for them that respects the autonomy of and interactions between individual agents, and (3) reason about them efficiently. Among these challenges are handling conflicts, allowing agents to deviate from policies, and understanding the trade-offs between system architectures that accommodate different levels of autonomy and efficiency.

Organisation of the Seminar

The seminar was held over the week of June 29 to July 04, 2025 (Monday to Friday with arrival on Sunday). We had 36 on-site participants. We started the first day with a two-minute lightning talk of each of the participants. Each participant was asked to answer the following questions: "what I'm doing, what I want to discuss here, and what input I would like". Before the Monday lunch, we had our two invited vision talks:

Matthew Arrott presented his vision from an industrial perspective in "Adoption of Interaction Protocols in Multi-Party Financial Transaction Platforms".

Pınar Yolum gave her view from an academic perspective in "Sociotechnical Thinking of Privacy".

After lunch, we brainstormed in random groups on topics for discussion for the week, followed by an open space session to determine the working groups for the week. We settled on

- Concepts: What are the underlying aspects of this interdisciplinary field and can be defined consistently?
- Agentic AI: How can we make agents interact and reason with human users through large language models (LLMs)?

- Standardisation: How can we facilitate global data sharing and compliance in a world with competing business interests?
- Coevolution: How can we make sure that STS evolve with the societies they operate in?

For the remainder of the seminar, we focused on discussing these topics, with the intention of coming up with plans for research publications in each of them. This was done through breakout groups and plenary discussions for cross-fertilisation between groups. Participants were encouraged to move between groups. At the end of each day, we had one or two primers of one of the participating disciplines. These primers addressed

- Judith Simon, "Ethics".
- Rigo Wenning, "Standardisation".
- Amit K. Chopra, "Programming with Norms".
- Harko Verhagen and Julian Padget, "The Value of Values".
- Beishui Liao, "Logic for New Generation AI: An Argumentation Based Methodology".
- Joris Hulstijn, "Regulatory Supervision".

In keeping with Dagstuhl tradition, we organised a social event. While normally this would be an excursion or a walk, we decided, due to the excessive heat that week, to stay at the Castle and have a local wine tasting. Julian Padget and Harko Verhagen kindly hosted the event. The final session of the seminar was dedicated to looking ahead and discussing plans for future collaboration and publications as a result of working group discussions. We also discussed the possibility of a follow-up Dagstuhl Seminar. Overall, the seminar was highly engaging, intellectually stimulating, and a great success.

Outcome of the Seminar

Scientific content

Open problems: Each of the four working groups selected by the participants focused on advancing understanding of open problems in the field. The working group reports at the end of this document detail the context of each working group, the progress they made during the seminar, which further challenges were identified, and how the group will take the topic forward.

New connections: The seminar brought together researchers who are working in the broad area of policy modelling but who approach that subject from different angles. Through primers, discussion in the streams, and in social time, we saw interesting exchanges of ideas among the subcommunities. We hope that these interactions will be fostered into new collaborations.

Extensive collaborations: Participants extensively discussed problems and challenges with colleagues, establishing new avenues for collaboration or reinforcing existing ones.

2 Table of Contents

3 Research area

Marina De Vos (University of Bath, GB)
Nicoletta Fornara (USI – Lugano, CH)
Munindar P. Singh (North Carolina State University – Raleigh, US)
Leon van der Torre (University of Luxembourg, LU)

Research Challenges

This seminar synthesised research perspectives from computing with insights from the law, public administration, and the social sciences. In particular, the relevant communities in computing include Semantic Web, Knowledge Representation and Reasoning, Logic Programming, Multi-agent Systems, Privacy and Security, and Legal Informatics, which we introduce below.

This seminar investigated the entire lifecycle of problems in policy dealing with STS. The design of a policy in an STS faces a fundamental trade-off between the *autonomy* accorded to member agents and the *control* exercised over those agents to guide them toward stakeholder objectives.

- Architecture. Architecture here encompasses the social and technical tiers of an STS. It captures what assumptions member agents can make about each other and what guarantees they can expect from the social and technical tiers. These guarantees can be expressed as policies and motivate an interest in an expanded view of policies. These guarantees may include organisational controls, such as sanctions applied for deviation from a policy.
- Models. Models concern how policies are conceived, including the languages in which they are expressed and how they relate to other parts of the relevant information systems. In a formal sense, the models reflect the architecture in an information model along with the needs of the domain. Models include considerations of the formal semantics, e.g., in terms of the computations that can be realised from a system and a determination of which computations are compatible with a given set of policies.
- Reasoning. Reasoning concerns how decisions can be derived from policies, given the facts and reasoning about policies, such as whether they conflict or one subsumes another. It incorporates monitoring (for ease of exposition) to enable reasoning on specific instances as well as determining if a particular deviation was legitimate.
- Methodology. Methodology concerns ways in which policies may be specified for an STS, given stakeholder requirements. It incorporates making changes in light of observed decisions, whether deviations took place, and whether the outcomes and the deviations (if any) were deemed legitimate.

The objective of this seminar was to provide a platform for researchers from different fields to form a new community. Specifically, we sought to motivate participants to define new research problems along with promising ways of tackling them.

Contributing Research Fields

Here is an overview of the various fields and communities that study policies in various forms.

Semantic Web. addresses languages and methods for encoding information formally so that intelligent agents can use that information without the risk of ambiguity that attaches to informal notations such as natural language. The Semantic Web is focused on ontologies – a form of expressive metadata – along with algorithms for formal reasoning on the ontologies. A famous realisation of the Semantic Web is Linked Data, wherein data are mutually linked to enhance their meaning and usefulness.

Knowledge representation and reasoning. (KRR) addresses ways to represent information so it can be used by an intelligent agent in its reasoning and planning. KRR incorporates findings from folk psychology to design formalisms that facilitate solving complex tasks. KRR traditionally emphasises computational logic to automate reasoning and includes studies of rules. Here, logic programming concerns models that automatically generate solutions to formally represented problems, thus obviating the need for procedural algorithms. KRR also includes deontic logic, which focuses on reasoning about obligations, permissions, and rights and thus relates closely to policies.

Multi-agent systems. (MAS) addresses developing systems of autonomous agents that are logically decentralised. A MAS is characterised by how its member agents interact, which leads to research into formal communication languages (described by the information they convey and the social relationships they affect) as well as intelligent decision-making about whether and when to perform a communicative act and whether and how to respond to a communicative act by another agent. These topics are thus well-aligned with policies. The connection is stronger in the Normative MAS (NorMAS) subfield, which focuses on social and legal norms and on organisational architectures.

Privacy, security and policies. for their proper handling of specific data are crucial for many areas of research. Namely, the Semantic Web, because it is focused on enabling data-sharing, for the legal domain, because privacy is regulated by data protection regulations, like GDPR, and the database community for its access control studies. Various privacy issues may arise at different stages of information management, from its collection to its processing and dissemination.

Legal informatics. addresses the formal modelling of laws concerning the usage of AI (with respect to AI provider or platform, e.g., with respect to privacy) and any domain where AI is used (e.g., with respect to the liability of a robot or a robot operator). This field relates well to the above areas of computing; it presents them with challenging problems and benefits from their solutions.

Seminar Topics

Legal knowledge representation and reasoning. Laws can be understood as high-level norms on behaviour and interaction in a society. Policies can be understood as operationalisations of laws. Policies in the legal sense are still high-level in that they may not be readily computed with, partly because they are expressed in natural language and partly because they reference information that may not be readily computationally characterised.

This seminar will study formal policy models and concomitant methodologies through which legal nuance can be reliably represented and reasoned about. Important concerns

include (1) conflicts between policies (e.g., due to jurisdiction or other attributes), (2) the (constrained) freedom of an agent to violate a policy, and (3) revisions.

Reasoning about correctness. Policies occupy the space between (legal and social) norms and agent behaviour. In our STS framework, this exposes important challenges concerning (1) validation: whether a policy represents stakeholder needs as evidenced in the applicable norms, (2) verification: whether agent interactions as designed respect the applicable policies, and (3) compliance: whether agent interactions as realised deviate from the applicable policies. Responses to these challenges determine how an STS and its member agents can be improved through continual revision – e.g., deviations may be justified by an "upstream" argument that the computational policy omitted a possibility allowed by the underlying informal policy.

This seminar explored not only policies as artefacts (and how to represent and reason about them computationally) but also the human-driven processes through which they are developed and revised.

Sociotechnical architecture. The above vision calls for new thinking about the architecture of STS. That is, we need to capture what an agent can expect from other agents and from the STS. Specifically, these expectations concern how information and control are distributed: are some policies enforced through technical artefacts (and difficult to violate without circumventing those artefacts)? Are there compliance checks when onboarding new agents into the system? Are interactions monitored? Are there social controls in place, e.g., reputation or eviction? Can sanctions arising from deviations be negotiated? This seminar studied alternative architectures as devised in informal real practice (such as the law and organisations), semiformal practice (such as access control and break-the-glass scenarios in healthcare, and more formal models (such as in organisational models in MAS).

Applications. Policy technologies are a case where engineering has gotten ahead of science. For example, Open Digital Rights Language (ODRL) is a W3C Recommendation, it is a policy expression language that provides an information model and a vocabulary for policies about the usage of digital assets and services. Even though ODRL is gaining traction (it's now in version 2.2), it lacks a formal model and semantics. We anticipate that use cases from ODRL, albeit limited to data policies, could be interesting real-life challenges for our discussions of the policy lifecycle, and especially on the formal models. This seminar will study practical use cases of policies in practice and identify research to give practice a robust foundation so that it can proceed with greater rigor and generality.

4 Overview of Talks

4.1 Prosociality and Ethics in Sociotechnical Systems

Nirav Ajmeri (University of Bristol, GB)

Prosociality refers to voluntary actions or behaviours intended to benefit an individual or society at large. Normative ethical principles are philosophical guidelines that define acceptable and unacceptable behaviours, offering a foundation for evaluating actions based on their ethical implications. As AI systems increasingly influence decisions with societal impact, their ability to act in prosocial and ethically aligned ways becomes critical. AI

Agents designed today often prioritise the goals and preferences of their primary users – risking outcomes that reinforce existing privileges and disadvantage vulnerable individuals or marginalised communities. Even agents designed for multi-stakeholder contexts may inadvertently overlook broader societal implications. At this Dagstuhl Seminar, I explore how normative ethics can inform the design of AI systems that account for the well-being of others, and discuss recent methods for embedding ethical norms and prosocial behaviours in agents through interaction and social learning. These methods enable more equitable, fair, and responsible STS – better aligned with the values of all stakeholders.

4.2 Introducing exceptions, accountability, commitments, information protocols into JaCaMo and Jason

Matteo Baldoni (University of Turin, IT) and Cristina Baroglio (University of Turin, IT)

JaCaMo and JADE + 2COMM: we show the benefits of explicitly representing social relationships between agents. This approach improves code modularity and interaction flexibility. Additionally, treating commitments as manipulable resources allows agents to reason about their interactions and strategically decide how and when to engage with others to pursue their own goals, enhancing the overall system's effectiveness.

JaCaMo extended with exceptions and accountability: we aimed to enhance the robustness of MAS. The first extension to JaCaMo introduces an exception handling mechanism tailored for MAS, while the second uses accountability to establish feedback chains among agents. Both extensions offer high-level abstractions and follow a unified approach to support the design of robust MAS capable of functioning correctly despite disruptions.

Jason and his friends Orpheus and Azorus: Orpheus and Azorus offers a programming model designed to enhance commitment-based reasoning in decentralised MAS. It uses declarative specifications centred on commitments and integrates them with information protocols. It supports reasoning about both goals and commitments and unifies three key technologies: Jason (a BDI-based agent programming model), Cupid (a formal language for commitments), and BSPL (a protocol language for information exchange). The model is implemented and shown to effectively represent complex business logic patterns.

4.3 All Intelligent Agents Should Speak a Formal Language

Victor Charpenay (Mines Saint-Étienne, FR)

Modern STS usually involve many more machines than humans. Among each other, machines always speak a formal language, which can be as simple as JSON (or better, JSON-LD) or as elaborate as FIPA-ACL. Because humans are outnumbered in STS, the best way for them to interact with machines is to speak the language of machines, via a dedicated graphical user interface for example. The ability of Transformers to model natural language should not encourage engineers to make machines speak natural language but rather to develop new forms of user interface to enhance the fluency of humans in formal languages.

4.4 Meaning-Based Abstractions for Agentic AI

Amit K. Chopra (Lancaster University, GB)

The Agentic AI paradigm is concerned with creating LLM-powered agents that take actions in the real world on behalf of their users. The promise of LLMs lies in their potential to reduce the knowledge engineering effort needed to build agents. Instead of being explicitly programmed, the agents would exploit LLMs to engage in a natural language dialog with their users, figure out the relevant constraints, and act accordingly. Several software frameworks (including protocols) lay claim to realizing Agentic AI. However, these frameworks miss crucial features about the context of real-world actions and their meanings.

Via the notion of norms, meaning is what much research in MAS has been concerned with. The idea is that communications between agents change the normative state of a system and this state is what matters to agents (and the principals they represent) in their reasoning. Recent work has shown how agents can engage flexibly on the basis of norms. A great direction for research is the synthesis of this body of work with LLM-based reasoning to realise more flexible, practical, and reliable Agentic AI.

4.5 Normative Multiagent Systems

Mehdi Dastani (Utrecht University, NL)

Normative systems are widely recognised as an effective means of regulating agent behaviour in MAS. Since the introduction of new norms alters system behaviour, there is a need for formal methodologies to model such dynamics. One line of research in our research group is to addresses this by treating the addition of norms as system updates and introducing formal update semantics to capture their impact.

Another contribution examines norm revision as a mechanism for improving system performance and ensuring the fulfilment of desirable properties. By analysing revisions such as relaxation and strengthening, and illustrating their effects through practical scenarios, our research explores how adaptive adjustment of norms can align MAS behaviour with system-level objectives.

A complementary approach investigates the challenges of maintaining effective norm enforcement in dynamic environments, where objectives evolve and previously defined norms may lose their effectiveness. To address this, we have introduced the data-driven norm revision framework. This framework automatically synthesises and revises conditional prohibitions with deadlines using system execution data. By analysing behavioural traces, the framework generates revised norms that more accurately distinguish between acceptable and unacceptable behaviours. Empirical evaluation using an advanced urban traffic simulator demonstrates that our approach significantly outperforms original norms in supporting the achievement of system objectives.

Collectively, these research directions advance the theory and practice of dynamic norm management in MAS by providing formal models for norm updates, conceptual tools for norm revision, and data-driven methods for adaptive norm synthesis.

4.6 Engineering Human-AI Teams: Norms, Values, and Responsible Collaboration

Davide Dell'Anna (Utrecht University, NL)

Recent advances in AI have made it necessary or desired for humans to get involved in interactions with AI systems on a daily basis. A key factor for the acceptance and responsible use of AI systems in STS is their ability to understand and adapt to personal, social, and legal norms. My research focuses on methodologies and mechanisms for designing AI systems that collaborate with humans synergistically and proactively as Human-AI teams where members amplify each other's intelligence by combining their complementary strengths [3]. I study how to represent, computationally, human social constructs such as norms, values, and team properties, and how to develop automated adaptive and data-driven mechanisms that ensure that AI behaviour is responsible, trustworthy, and justifiable [2, 1].

References

1 Davide Dell'Anna, Natasha Alechina, Fabiano Dalpiaz, Mehdi Dastani, and Brian Logan. Data-driven revision of conditional norms in multi-agent systems. *J. Artif. Intell. Res.*, 75:1549–1593, 2022. URL: https://doi.org/10.1613/jair.1.13683, doi:10.1613/JAIR.1.13683.

2 Davide Dell'Anna and Anahita Jamshidnejad. SONAR: an adaptive control architecture for social norm aware robots. *Int. J. Soc. Robotics*, 16(9):1969–2000, 2024. URL: https://doi.org/10.1007/s12369-024-01172-8, doi:10.1007/S12369-024-01172-8.

3 Davide Dell'Anna, Pradeep K. Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M. Jonker, Catharine Oertel, and Pınar Yolum. Toward a quality model for hybrid intelligence teams. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 434–443. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024. URL: https://dl.acm.org/doi/10.5555/3635637.3662893, doi:10.5555/3635637.3662893.

4.7 Sociotechnical Systems: Stay relevant: Norm Change and Value-Alignment

Marina De Vos (University of Bath, GB)

Joint work of Marina De Vos, Andreasa Morris-Martin, Julian Padget, Jack McKinlay, Mattias Brännström, Lili Jiang

When implementing autonomous agents in an STS, it is critical that their actions are in line with expected behaviours and with the social values of the stakeholders of the systems. To achieve this, the system and the agents should be equipped to reason about the norms and values of the system and how these can be affected by their actions. For modelling norms within a system we use the institutional action language InstAL which maps computationally to answer set programming. As the STS evolves over time, we allow participants to request norm changes which, if approved, are implemented in the system by updating the norms through inductive logic programming. Recently, we started exploring incorporating values and value-alignment, either through stand-alone value extraction from policy documents using LLMs or value-based decision-making for agents using argumentation frameworks.

4.8 Trust envelopes: Vehicles of History and Destiny

Beatriz Esteves (Ghent University, BE)

The original vision of the Web was one of decentralisation; however, this ideal is not reflected in the current landscape. While Web-based services and data inherently originate from diverse sources, the exchange of data – particularly personal data – is predominantly governed by a limited number of large BigTech companies. This concentration of control has contributed to growing public distrust in these services.

Our argument is not that data flows are absent, but rather that they are inefficient and misaligned with technological, legal and business principles. On one side, companies, uncertain about user retention, engage in aggressive data collection from the very first interaction. On the other, users – seeking the convenience of online services – often accept privacy policies and terms of service without adequate scrutiny.

We hypothesise that meaningful and trustworthy data exchange, conducted at every point in time where a data point needs to be exchanged with a clearly defined purpose, can foster evolving, trust-based relationships between individuals and organisations. To support this vision, we introduce the concept of trust envelopes. A trust envelope serves as a carrier of both the historical context and the intended future use of a data element. By accompanying data with usage policies, provenance and other contextual information, trust envelopes enable recipients to verify the origin and quality of the data and to use it in accordance with the source entity's preferences. As such, they enable well-intentioned actors to engage in responsible data exchange without facing the disproportionate obstacles currently present in the digital ecosystem, while those with less sincere motives will be unable to exploit the advantages of these evolvable trust relationships.

4.9 Modelling and Reasoning about Policies in Sociotechnical Systems

Nicoletta Fornara (USI – Lugano, CH)

The problem of modelling and reasoning on norms, policies, agreements and licenses is increasingly crucial in many fields of application and research, e.g. in the design and development of STS, in the regulation of autonomous agents on the Web of Things (see the Dagstuhl Seminar 23081 Agents on the Web[1]), for the governance of the use and exchange of personal and business knowledge graphs between parties (see Dagstuhl Seminar 25051 Trust and Accountability in Knowledge Graph-Based AI for Self-determination[2]), for the governance of the exchange of Data Spaces, Personal Data Stores (in the Solid open standard) and for the second used of health data. Automatically reasoning on the semantics of policies is crucial for providing different types of services, for example, what-if analysis, access control, monitoring and sanctioning, and conflict detection. In my research I studied the

[1] Dagstuhl Seminar 23081 Agents on the Web (Feb 19 – Feb 24, 2023) `https://www.dagstuhl.de/23081`
[2] Dagstuhl Seminar 25051 Trust and Accountability in Knowledge Graph-Based AI for Self Determination (Jan 26 – Jan 31, 2025) `https://www.dagstuhl.de/25051`

formalisation of frameworks for modelling and reasoning on policies by using Semantic Web Technologies and rule languages. Together with my colleagues, we proposed a model to represent and reason about obligations, prohibitions and permissions by extending the ODRL policy language [1] and the T-NORM model of norms able to regulate classes of actions whose performance is temporally constrained [2]. Since 2021, I have been co-chair of the W3C ODRL (Open Digital Rights Language) Community Group in which I mainly coordinate the activities of the group that defines the semantics of ODRL 2.2[3] [3]. Important challenges are the completion of the definition of the formal and operational semantics of the ODRL 2.2 language and the proposal of a new version of the model to overcome its current limitations, which must pass through the study of actual use cases, including its use in STS, and the definition of the main requirements that the new model should meet [4].

References

1 Fornara, N., & Colombetti, M. (2019). Using semantic web technologies and production rules for reasoning on obligations, permissions, and prohibitions. Ai Communications, 32(4), 319-334. https://doi.org/10.3233/AIC-190617.

2 Fornara, N., Roshankish, S., & Colombetti, M. (2021, May). A framework for automatic monitoring of norms that regulate time constrained actions. In International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems Vol. 13239 (pp. 9-27). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-16617-4_2.

3 Bonatti, P. A., Fornara, N., & Harth, A. (2025). Towards a Formal Semantics of the Open Digital Rights Language (ODRL 2.2). ESWC 2025 Workshops and Tutorials Joint Proceedings. 1st International Workshop on ODRL and beyond: Practical Applications and challenges for poLicy-base access and usage control (OPAL2025). Portorož, Slovenia 1st June 2025. Vol-3977. https://ceur-ws.org/Vol-3977/OPAL2025-4.pdf.

4 Cimmino, A., & Fornara, N. (2025). Improving ODRL 2.2: current limitations and theoretical solutions. ESWC 2025 Workshops and Tutorials Joint Proceedings. 1st International Workshop on ODRL and beyond: Practical Applications and challenges for poLicy-base access and usage control (OPAL2025). Portorož, Slovenia 1st June 2025. Vol-3977. https://ceur-ws.org/Vol-3977/OPAL2025-6.pdf.

4.10 Regulatory Supervision

Joris Hulstijn (Utrecht University, NL)

This tutorial presents a brief summary of some of the theory in public administration, IT auditing and law, that is relevant to regulatory supervision and compliance, especially where it concerns corporate regulation; not individual citizens. In particular, we discuss the topics of responsive regulation, (enforced) self-regulation, system-based supervision (also known as collaborative compliance) and the interpretation of open norms. These notions are illustrated by an example from the customs domain: "towards data-driven supervision". In a data-driven supervision process, regulators rely on data from documents provided by the company being supervised. In that case, the central question is: how often and when should regulators schedule inspections to verify the data?

[3] ODRL Formal Semantics, Draft Community Group Report https://w3c.github.io/odrl/formal-semantics/

4.11 What Do We Need for Software-based Normative MAS?

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

Indeed, for technologies such as business rule engines, enforcing and verifying compliance with normative requirements is a primary use-case class. Accordingly, we have "normative" capabilities in software systems that govern sociotechnical MAS. However, the "intelligent" handling of exceptions to norms, the deliberate violation of norms based on values, and the evolution of norms is still up to humans, incurring substantial social efforts in organisations, and causing conflicts between norms and values. Accordingly, more flexible architectures and practically more expressive abstractions are required for moving the operational workload of norm management and evolution from the human to the software agent level.

4.12 Logic for New Generation AI: An Argumentation Based Methodology

Beishui Liao (Zhejiang University, CN), Réka Markovich (University of Luxembourg, LU), Leon van der Torre (University of Luxembourg, LU)

This talk introduces a methodology to address the challenges of managing inherently conflicting and evolving policies, norms, and values within complex STS. It argues that formal argumentation provides the necessary rigorous, structured foundation for representing these concepts, including violations and causal links, and for deriving defensible conclusions. To tackle these complexities, we propose a comprehensive integrated framework built upon formal argumentation. This framework consists of six core, interconnected components: 1) A unified representation using defeasible rules to formalise norms, policies, and their violation conditions, naturally handling exceptions and priorities; 2) A conflict and violation resolution engine based on argumentation theory to systematically identify and adjudicate conflicts and violations through argument evaluation; 3) A dynamic adaptation mechanism using argumentation revision to evolve the system by adding, modifying, or retracting rules and arguments in response to change; 4) A causal attribution interface combining argumentation with causal inference to link normative states (compliance/violation) to root causes of outcomes; 5) An efficient computation strategy employing locality and modularity for scalable reasoning in large systems; 6) A neuro-symbolic integration pathway leveraging LLMs for tasks like natural language parsing and argument generation, while relying on the argumentation core for rigorous, explainable reasoning and validation.

References

1 Michael Anderson, and Susan Leigh Anderson. *Geneth: a general ethical dilemma analyzer*. Paladyn J. Behav. Robotics, 9(1):337–357, 2018.
2 Edmond Awad, Michael Anderson, Susan Leigh Anderson, and Beishui Liao. *An approach for combining ethical principles with public opinion to guide public policy*. Artif. Intell., 287: 103349, 2020.

3 Pietro Baroni, Guido Boella, Federico Cerutti, Massimiliano Giacomin, Leendert W. N. van der Torre, and Serena Villata. *On the input/output behavior of argumentation frameworks*. Artif. Intell., 217:144–197, 2014.

4 Pietro Baroni, Marco Romano, Francesca Toni, Marco Aurisicchio, and Giorgio Bertanza. *Automatic evaluation of design alternatives with quantitative argumentation*. Argument Comput., 6(1):24–49, 2015.

5 Pietro Baroni, Massimiliano Giacomin, and Beishui Liao. *Locality and Modularity in Abstract Argumentation*. In Handbook of Formal Argumentation, pp. 937-980, 2018.

6 Chen Chen, Pere Pardo, Leendert van der Torre, and Liuwen Yu. *Weakest link in formal argumentation: Lookahead and principle-based analysis*. In Andreas Herzig, Jieting Luo, and Pere Pardo, editors, Logic and Argumentation – 5th International Conference, CLAR 2023, Hangzhou, China, September 10-12, 2023, Proceedings, volume 14156 of Lecture Notes in Computer Science, pages 61–83.

7 Haixiao Chi and Beishui Liao. *A quantitative argumentation-based automated explainable decision system for fake news detection on social media*. Knowledge-Based Systems, 242:108378, 2022.

8 Phan Minh Dung. *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*. Artificial intelligence, 77(2):321–357, 1995.

9 Xiaotong Fang, Zhaoqun Li, Chen Chen, and Beishui Liao. *Llm-aspic+: A neuro-symbolic framework for defeasible reasoning*. To appear, 2025.

10 Bettina Fazzinga, Sergio Flesca, and Francesco Parisi. *On the complexity of probabilistic abstract argumentation frameworks*. ACM Trans. Comput. Log., 16(3):22:1–22:39, 2015.

11 Anthony Hunter. *Some foundations for probabilistic abstract argumentation*. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, Computational Models of Argument – Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012, volume 245 of Frontiers in Artificial Intelligence and Applications, pages 117–128. IOS Press, 2012.

12 Kun Kuang, Lian Li, Zhi Geng, Lei Xu, Kun Zhang, Beishui Liao, Huaxin Huang, Peng Ding, Wang Miao, Zhichao Jiang. *Causal Inference*. Engineering, Volume 6, Issue 3, March 2020, Pages 253-263.

13 Beishui Liao. *Toward incremental computation of argumentation semantics: A decomposition-based approach*. Ann. Math. Artif. Intell., 67(3-4):319–358, 2013. 10.1007/S10472-013-9364-8.

14 Beishui Liao. *On interdisciplinary studies of a new generation of artificial intelligence and logic*. Social Sciences in China, 43(3):5–19, 2022.

15 Beishui Liao and Huaxin Huang. *ANGLE: an autonomous, normative and guidable agent with changing knowledge*. Inf. Sci., 180(17):3117–3139, 2010.

16 Beishui Liao, Li Jin, and Robert C Koons. *Dynamics of argumentation systems: A division-based method*. Artificial Intelligence, 175(11):1790–1814, 2011.

17 Beishui Liao, Kang Xu, and Huaxin Huang. *Formulating semantics of probabilistic argumentation by characterizing subgraphs: theory and empirical results*. J. Log. Comput., 28(2):305–335, 2018.

18 Beishui Liao, Leendert van der Torre. *Explanation Semantics for Abstract Argumentation*. COMMA 2020: 271-282.

19 Beishui Liao, Michael Anderson, and Susan Leigh Anderson. *Representation, justification, and explanation in a value-driven agent: an argumentation-based approach*. AI Ethics, 1(1): 5–19, 2021.

20 Beishui Liao, Pere Pardo, Marija Slavkovik, and Leendert van der Torre. *The jiminy advisor: Moral agreements among stakeholders based on norms and argumentation*. Journal of Artificial Intelligence Research, 77:737–792, 2023.

21 Beishui Liao, Leender van der Torre. *Attack-defense semantics of argumentation*. COMMA 2024: 133-144.

22 Liuwen Yu and Davide Liga Réka Markovich. *Addressing the right to explanation and the right to challenge through hybrid-ai: Symbolic constraints over large language models via prompt engineering*. In The 20th International Conference on Artificial Intelligence and Law, 2025.

23 Liuwen Yu, Réka Markovich, and Leendert van der Torre. *Interpretations of support among arguments*. pages 194–203. IOS Press, 2020.

24 Liuwen Yu, Dongheng Chen, Lisha Qiao, Yiqi Shen, and Leendert van der Torre. *A principle-based analysis of abstract agent argumentation semantics*. In Meghyn Bienvenu, Gerhard Lakemeyer, and Esra Erdem, editors, Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021, Online event, November 3-12, 2021, pages 629–640, 2021.

25 Liuwen Yu, Mirko Zichichi, Réka Markovich, and Amro Najjar. *Enhancing trust in trust services: Towards an intelligent human-input-based blockchain oracle (ihibo)*. In 55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022, pages 1–10. ScholarSpace, 2022.

26 Liuwen Yu, Leendert van der Torre, and Réka Markovich. *Thirteen challenges in formal and computational argumentation*. In Gabbay, D., Kern-Isberner, G., Simari, G.R., Thimm, M. (eds.) Handbook of Formal Argumentation, pages 890–976. College Publications, 2024.

4.13 Abstract of Research

Réka Markovich (University of Luxembourg, LU)

I research computational legal theory and study its applications in AI and legal reasoning. My focus areas are legal knowledge representation, norMAS, deontic logic, machine ethics, and explainable AI (XAI). Computational legal theory is about reconstructing fundamental legal concepts and structures in a formal language. One of the topical foci of mine has a special relevance for policy modelling and reasoning in STS: I have been investigating the formal structure of normative positions. The theory of normative positions is based on the theory of W.N. Hohfeld, who differentiated between four types of positions often referred to as a "right" (claim-right, privilege/freedom, power, immunity) and their corresponding "duty" positions (duty, no-claim, liability, disability). The agents in these positions are in normative relations with each other. This differentiation and the characterisation of the positions and the relations are crucial in order to avoid terminological mess in the law and any system aiming at implementing it, but also for understanding further fundamental concepts playing an essential role in STS, such as competence, responsibility, authority, commitment. Hence the concepts and their adequate formalisation contribute to build, as Marek Sergot puts it, the "characteristic of all forms of regulated and organised agent interaction".

4.14 Sociotechnical Systems: From Dialogue to Decisions

Pradeep Murukannaiah (TU Delft, NL)

Engineering STS is the overarching theme of my research. I envision an STS as a system that supports rich interactions among principals (humans or organisation) and computational agents, enabling a variety of individual and societal applications. In an STS, principals are paramount. Principals act autonomously (based on values) and are accountable to each other (as specified by norms). Agents, in contrast, support decision-making by principals. In this seminar, I explore how to connect dialogue among stakeholders to decision-support by agents.

4.15 Computational Machine Ethics

Vivek Nallur (University College Dublin, IE)

Increasingly machines will be called upon to take ethically charged decisions. In these situations, it is important for the machine to behave in an ethically acceptable manner. The ability of an ABM-based simulation to "play out a few steps into the future", allows the agent to make a principled decision. Those decisions can be taken in a manner that respects multiple stakeholder values, i.e., in a pro-social manner. The agent also attempts to anticipate the humans around it, by understanding the cognitive model of the human interacting with it, and the various possible biases that impact human decision-making.

4.16 Normative Regulation of the Industry of the Future

Luis Gustavo Nardin (IMT Mines Saint-Étienne, FR)

Modern industry is compelled to become more flexible, adaptable, resilient, sustainable, and human-centred in order to evolve and remain competitive. We claim that these requirements can be fulfilled by coupling industrial processes modelling with normative aspects to define a set of design principles for governing industrial systems to operate trustworthy and sustainably, and to respond quickly and flexibly to exogenous and endogenous changes. The explicit normative representation and reasoning enable agents to both adapt the execution of industrial processes to unexpected situations and conditions, and to transparently and intelligibly express their decisions to an human operator. We aim to create normative regulation mechanisms, design regulation architectures and implement platforms that enable agents to operate in heterogeneous and dynamic industrial settings and reason about normative aspects to enhance flexibility, resilience, trustworthiness, and sustainability for the Industry of the Future.

4.17 CERTAIN: Towards Traceability and Regulatory Compliance in AI

Sebastian Neumaier (FH – St. Pölten, AT)

As AI systems become increasingly embedded in critical sectors, ensuring regulatory compliance and ethical integrity becomes essential. In my talk, I introduce the CERTAIN project (`https://certain-project.eu/`), which aims to develop a comprehensive framework for the traceability and compliance checking of AI systems within the evolving regulatory landscape of the European Union. Central to this effort is a Semantic MLOps Engine and a RegOps Engine:

- The semantic engine supports lifecycle tracking via ontologies;
- The RegOps engine enables compliance assessment by querying the collected information in a corresponding knowledge graph that captures the AI development and deployment process.

At the seminar, we discussed the relevance of the project's goals to the seminar's core themes, addressing decentralised system governance, ODRL-inspired policy semantics, and verifiable policy modelling in sociotechnical ecosystems.

4.18 The value of Values

Julian Padget (University of Bath, GB) and Harko Verhagen (Stockholm University, SE)

We draw on work in psychology and computer science to propose an approach to the embedding and operationalisation of values – or more precisely, value preference orders – in the design, implementation and operation of STS.

Our motivation is to put forward a methodology – called conscientious design – that puts people at the heart of systems so that (sociotechnical) systems meet – and continue to meet over their lifetime – the expectations of a changing population of participants. A further driver is that for many years we believed that norms were the right technology for capturing and operationalising human requirements in STS, but have concluded that while precise, they are also brittle, hard to write and hard(er) to maintain. In contrast, while not solving the problem, values offer a means to contextualise the norm production and maintenance process to realise what we call small "v" value alignment.

Conscientious design builds upon Schwartz's universal values and Friedman's Value-Sensitive Design (VSD) to propose a frame of reference for STS stakeholder values, in the form of a bespoke value system constructed around the axiology of thoroughness, mindfulness, responsibility. This extends into a process that embeds representations of values that go beyond the design stage to operation, revision and retirement creating a value-based approach to through-life development.

References

1 Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. "Ethical Online AI Systems Through Conscientious Design". In: IEEE Internet Computing 25.6 (2021), pp. 58–64. `https://doi.org/10.1109/MIC.2021.3098324`

2 Pablo Noriega, Harko Verhagen, Julian Padget, and Mark d'Inverno. "Design Heuristics for Ethical Online Institutions". In: Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV. Ed. by Nirav Ajmeri, Andreasa Morris Martin, and Bastin Tony Roy Savarimuthu. Springer, 2022, pp. 213–230. https://10.1007/978-3-031-20845-4_14.

3 Pablo Noriega, Harko Verhagen, Julian A. Padget, and Mark d'Inverno. "Addressing the Value Alignment Problem Through Online Institutions". In: Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVI – 27th International Workshop, COINE 2023, London, UK, May 29, 2023, Revised Selected Papers. Ed. by Nicoletta Fornara, Jithin Cheriyan, and Asimina Mertzani. Vol. 14002. Lecture Notes in Computer Science. Springer, 2023, pp. 77–94. https://10.1007/978-3-031-49133-7_5.

4.19 Extending ODRL for AI and Data Regulation

Victor Rodriguez Doncel (Polytechnic University of Madrid, ES)

Elements of the new EU legislation on AI and data can be formalised and operationalised. This opens the door to new types of software tools that support organisations in compliance-related tasks, and also allows for the analysis and simulation of the ethical and societal impacts of emerging technologies and their regulation within STS–this is the goal of the EU HARNESS project. In particular, certain norms can be represented using policy languages such as ODRL. Originally developed as a Rights Expression Language, ODRL has evolved into a more general policy language and could be extended to represent concrete legal norms found in recent AI and data regulations. To achieve this, new language features should enhance ODRL's expressiveness, and the behaviour of ODRL processors should be more precisely defined. Additionally, other ODRL-related tools should be explored, such as: translation mechanisms between ODRL and other languages (e.g., Prolog); methods for efficiently extracting rules from normative texts; and techniques for generating natural language (e.g., English) descriptions from formal rules.

4.20 Compliance Mechanism using LLM

Ken Satoh (Research Organization of Information and Systems, JP)

We launched the tri-lateral (Japan–France–Germany) research project *Research on Realtime Compliance Mechanism for AI (RECOMP)* (https://research.nii.ac.jp/RECOMP/) for the period 2021–2023, supported by the Japan Science and Technology Agency (JST), the Agence nationale de la recherche (ANR), and the Deutsche Forschungsgemeinschaft (DFG).

Our goal is to improve the reliability of AI in society by implementing real-time compliance mechanisms for legal and ethical norms. In our approach, legal norms are modelled as **hard constraints** that must always be satisfied, while ethical norms are modelled as **soft constraints** that should be satisfied as far as possible.

The overall agent architecture is shown in Fig. 1. When a new observation is received, the agent revises its current plan by integrating the new information with a causal theory

that encodes physical constraints and action–effect rules. We then verify legal compliance among candidate plans and select those that are legally valid. Next, we evaluate ethical compliance to identify ethically optimal plans, meaning plans that satisfy stronger ethical rules whenever possible.

In this abstract, I outline a proposal to extend our system using LLMs. At present, our framework is based on logic programming, which creates a gap between the formalised representation and the norms, typically written in natural language. This makes it difficult for both domain experts and system developers to fully understand the underlying logic-based knowledge representation. To address this, we propose the architecture shown in Fig. 2. In this design, system developers provide causal theories and legal and ethical rules in natural language. An LLM then translates them into a corresponding logic program. Similarly, user goals given in natural language are automatically translated into logical goals, and the inferred results from the logic program are translated back into natural language. This approach aims to create a more robust, accessible, and user-friendly system.

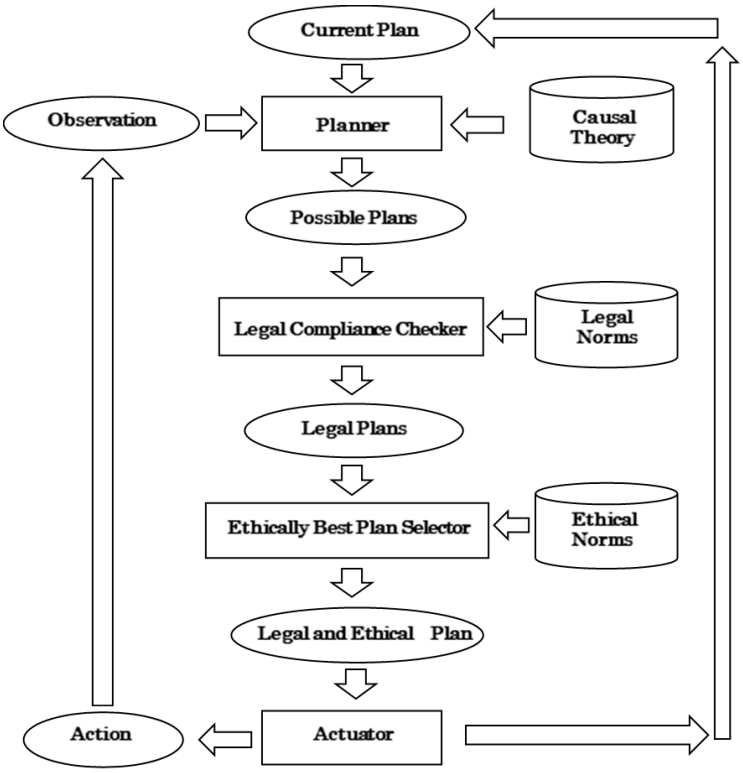

Figure 1 The architecture for RECOMP planning.

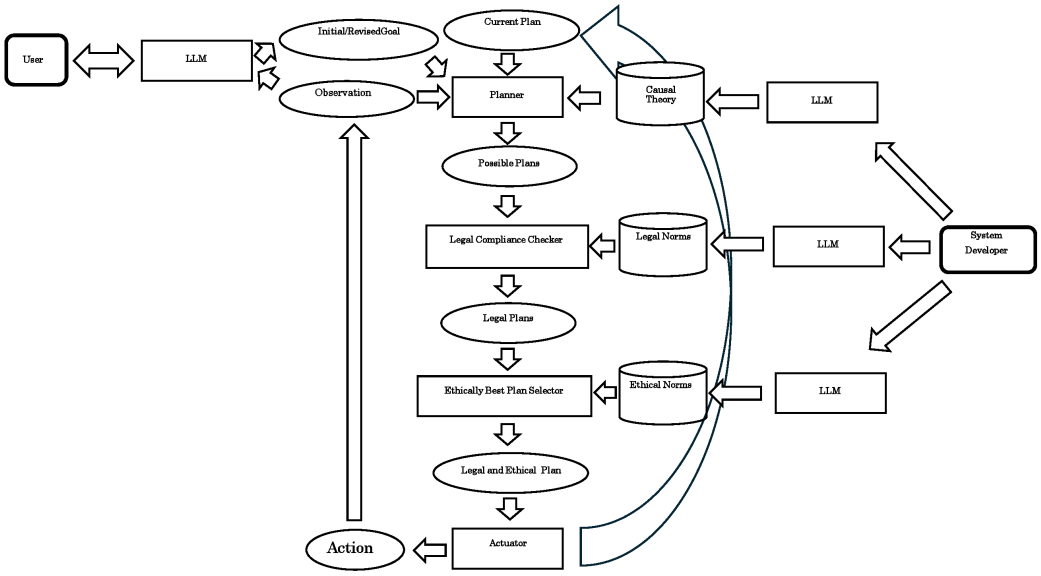

Figure 2 The architecture for RECOMP planning with LLM.

4.21 Regulation in Techno-Human Systems of the Future

Jaime Sichman (University São Paulo, BR)

In our current life, people interact with each other and with several institutions by using technical systems. These complex systems, composed by people and software / hardware are known as STS [1]. As an example, in our University researchers and students access the technical systems to see their grades, submit their works, submit their reports, access the restaurant menu, i.e., to interact with the institution USP.

The current state of the art in AI and Computer Science includes (i) LLMs & Normative Agents, (ii) Big Data for Smart Cities, (iii) Autonomous Systems, (iv) Security & Safety, (v) Green Computing, (vi) Complex Software Supply Chains and (vii) High Performance Computing & Simulation. However, these currently used techniques are not yet integrated in a single framework in order to enable better people's experiences in Digital Society.

Our current work intends to use multi-agent regulation techniques [2] to enhance these STSs. We intend to apply these techniques in the industry 4.0 domain [3] and integrate them with argumentation techniques [4].

References

1 F. E. Emery, E. L. Trist. Socio-technical systems. *Management science, models and techniques*, 2, 83-97, 1960.
2 E. Yan, L. G. Nardin, O. Boissier, J. S. Sichman. A Unified View on Regulation Management in Multi-Agent Systems. *In: Proc. of COINE2025 Workshop*, Detroit, USA, May 2025.
3 Normative Artificial Intelligence for Regulating Manufacturing. `https://naiman.wp.imt.fr/`, Accessed: 01/07/2025.
4 UNBIAS Team – argUmentation and Norm Based Intelligent AgentS. `https://thus.ime.usp.br/teams/unbias`, Accessed: 01/07/2025.

4.22 Ethics and AI: Resisting the Seduction of Frictionlessness

Judith Simon (Universität Hamburg, DE)

1. **Ethics:** Ethics in general asks questions about what is right and wrong, what is good and what is bad, what we should or must (not) do – and for what reasons. If applied to AI, this entails the question of what good and bad AI is – not only technically, but also morally and what we can and/or should (not) use AI for – and for what reasons.
2. **AI:** Looking into the history of AI, we can dissect three themes, related to epistemic, ontological and ethical questions as well as three promises, all of which were assessed later.
   The three themes are:
   - Big data & statistical reasoning: from means and standard deviations to personalisation without subjects
   - The role of imitation and deception
   - The human/cognition/language as both a benchmark and being deficient The three promises of AI are to increase the efficiency, quality and convenience.
3. **Ethical Challenges of AI:** I then outlined some of the most pressing challenges resulting from AI based upon three publications (Deutscher Ethikrat 2023, Simon et al. 2024, Simon 2025).
   These were:
   - Expanding/Reducing Agency
   - AI-based Knowledge Generation & Prediction
   - Endangering the Individual Through Statistical Stratification
   - Effects of AI on Human Competencies and Skills
   - Privacy & Autonomy versus Surveillance & Chilling Effects
   - Data Sovereignty and Data Use Oriented Towards the Common Good
   - Critical Infrastructures, Dependencies and Resilience
   - Path Dependencies & Dual Use
   - Bias and Discrimination
   - Transparency and Accountability – Control and Responsibility
   - Deception

   Having outlined the most pressing ethical challenges of AI, I argue that the judgment on the increased quality of cognitive process and decision making is still open and differs for different individuals & groups. The judgment on whether (Gen)AI increases the efficiency & convenience is also open – but even if epistemic processes were more convent and efficient, this very improvement comes with epistemic, ethical and political costs.
4. **Conclusions:** I concluded my talk with some suggestions on what we can do to design technologies with ethics in mind – but while also being aware about the limits of reaching ethical and political goals with and through technology.
   These are the following:
   - There is no "machine ethics": Ethics can't be delegated to machines, but requires judgment, situated and context-aware reasoning.
   - Another way of thinking about ethics and AI is rooted in the "Values in Design" approach. Instead of delegating ethics to tech, it asks: which values are relevant for whom and how can and should we operationalise them?

- Ethics is not a check-box to tick. Instead ethical considerations are part and parcel of the whole life-cycle of developing a deploying tech: from creating and annotating data, to choosing methods, using tech in specific contexts and taking care of their remains after they cease to work.
- So ethics is part of research and tech development, but also goes beyond tech. Think of de-biasing AI – you can't fully avoid discriminating against every possible group or individual, but have to make ethical and political choices, which harms are most important to prevent.
- Finally: Beware of the pitfalls of anthropomorphising technology and making humans machinic. Both are inherent in the history of AI anyway, but become even more salient in the field of normative MAS and ethical AI.

References

1 Deutscher Ethikrat (2023). *Mensch und Maschine – Herausfoderungen durch Künstliche Intelligenz*. Available at: `https://www.ethikrat.org/en/publications/opinions/humans-and-machines/`
2 Simon, J., Spiecker gen. Döhmann, I. & von Luxburg, U. Generative KI – Beyond Dystopia and Simple Solutions. Discussion Paper No. 34. Halle (Saale): National Academy of Sciences, Leopoldina, 2024. doi:10.26164/leopoldina_03_01245
3 Simon, J. (2025). Generative AI, Quadruple Deception and Trust. Social Epistemology. doi: 10.1080/02691728.2025.2491087
4 Simon, J., et al. (2020). Algorithmic bias and the Value-Sensitive Design approach. Internet Policy Review, 9(4), 1–16. Available at: `https://policyreview.info/concepts/algorithmic-bias`

4.23 Flexible, adaptive sociotechnical systems based on norms and values, their evolution, and their realisation using generative AI

Munindar P. Singh (North Carolina State University – Raleigh, US)

Our computational model of STS enables a natural way to elicit stakeholder needs (requirements, risk attitudes, and value preferences), reason about them, and build decentralised MAS meeting those needs. We have been working on improvements of this model to incorporate enhanced reasoning about value preferences and norms, especially in conjunction with each other and with mental constructs such as goals. In this seminar, we will discuss ideas relating to (1) a lifecycle for STS, especially its continual tracking and alignment with potentially changing stakeholder needs, (2) models for the emergence and evolution of norms in light of both observations and semantically rich models, and (3) realising STS by taking advantage of the facilitation of knowledge engineering provided by generative AI, including identifying ways to enrich current generative AI models and toolkits with social intelligence, thereby achieving a leap in the development of MAS that go far beyond today's rigid, workflow-based approaches.

4.24 Architects of Trust: A Framework for Sovereign Data Governance in the Age of Autonomous Agents

Simon Steyskal (Siemens AG – Wien, AT)

The emergence of decentralised STS, from industrial data spaces to autonomous MAS powered by LLMs, has exposed fundamental limitations in traditional, centralised governance models. This work presents a conceptual framework addressing the critical research challenge of establishing trustworthy, policy-based governance in environments where autonomous agents must interact without central authority.

The framework synthesises two W3C standards in a novel "two-pillar" architecture: the Open Digital Rights Language (ODRL) for expressing deontic policy semantics, and the Shapes Constraint Language (SHACL), repositioned as a dynamic policy enforcement engine through custom node expressions. We outline how this integration, combined with Decentralised Identifiers (DIDs) and Verifiable Credentials (VCs), could enable verifiable, context-aware governance that preserves data sovereignty while facilitating automated compliance checking.

Key research directions identified include: (1) developing formal semantics for ODRL-aware SHACL validation that transforms static data validators into Policy Decision Points, (2) addressing semantic gaps in current policy languages for complex temporal and contextual constraints, (3) establishing mechanisms for policy conflict resolution in multi-stakeholder environments, and (4) extending governance models to encompass LLM-powered agents where policies themselves may be generated through natural language interaction.

4.25 Human-Centred AI in Sociotechnical Systems

Sz-Ting Tzeng (University of Umeå, SE)

As AI increasingly integrates into our social structures, humans and AI form complex STS. Ensuring that AI aligns with human values and social norms and that AI behaviours are justifiable becomes increasingly important when humans are involved. My research focuses on developing human-centred AI that can make decisions and adapt its explanation strategies according to the social context and values in STS. In this seminar, I explore how decision making and AI-generated explanations reflect and are shaped by human values, and how agents adapt to evolving STS.

4.26 Ethical Decision-Making in Multi-Agent Systems

Jessica Woodgate (University of Bristol, GB)

Consequential decision-making is increasingly guided by AI in diverse social settings, from resource allocation to balancing preferences of stakeholders. Whilst AI has beneficial uses,

its sociotechnical nature entails it often adopts default social norms (standards of expected behaviour) and power structures of society, which includes systematic injustices and inequalities. Resource allocation may treat some recipients more favourably, or the preferences of minorities may be overlooked. Realising the benefits of AI across society necessitates addressing ethical implications, understood as what is morally good or right. Many ethical concerns are multi-agent in nature, involving one party's concern for another. MAS, which are collections of multiple agents interacting in a shared environment, are thus an appropriate setting to examine ethical implications of AI and encompass social factors such as norms. To advance ethical decision-making in MAS, operationalising principles from normative ethics – the philosophical study of practical means to determine right from wrong – helps support interdisciplinary insights and guide decision-makers in making evaluative judgements.

4.27 Sociotechnical reasoning of privacy

Pinar Yolum (Utrecht University, NL)

STS consist of agents and humans, each with potentially different capabilities, working together to accomplish tasks. For these systems to succeed, agents need to recognise, take into account, and demonstrate social, developmental, and communication skills – like self-reflection and empathy – that are typically linked to humans. How do we realise STS that benefit from these skills? How do we measure their existence? I argue that our vocabulary for talking about STS is based on individual AI systems and do not capture the effect of such skills. I demonstrate a few cases over the domain of privacy.

4.28 Regulating autonomous agents on the Web

Antoine Zimmermann (Ecole des Mines – St. Etienne, FR)

Much digital activity happens on the Web. As the Web gets increasingly vast and complicated, we need assistance from automated tools to process the information and sometimes do tasks for us. Complicated tasks require proactiveness and autonomy to complete automatically. When autonomous software acts on our behalf on the Web, it must do so in accordance with our policies and regulation. Therefore, we need mechanisms that allow artificial agents to become aware of their obligations, permissions, and prohibition in very strictly verifiable ways. Language models can help translating human-readable regulation into machine-processable representations, but they are prone to misinterpretation, approximation, and hallucination. We want to convey policies to agents on the Web in a formal, unambiguous form, and make them easily discoverable in a systematic way, such that agents can arrive at a location on the Web and operate on available resources according to the rules, without prior knowledge of the local context.

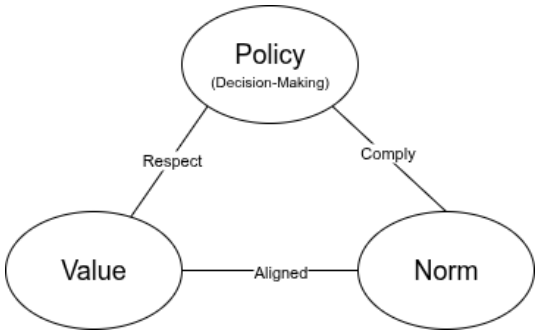

Figure 3 General interplay between policies, values, and norms.

5 Working groups

5.1 Coevolution of Values and Norms in Sociotechnical Systems

Nirav Ajmeri (University of Bristol, GB), Marina De Vos (University of Bath, GB), Davide Dell'Anna (Utrecht University, NL), Pradeep Murukannaiah (TU Delft, NL), Vivek Nallur (University College Dublin, IE), Luis Gustavo Nardin (IMT Mines Saint-Étienne, FR), and Munindar P. Singh (North Carolina State University – Raleigh, US)

5.1.1 Introduction

In this report, we provide an overview of the discussions held at the Dagstuhl 25721 Working Group on the coevolution of values and norms in a STS. We discuss the background and conceptualise an STS where norms and values can change over time and influence each other based on the observations and actions of the human actors and agents. We identify key research challenges spanning the formalisation, operationalisation, and application of such an STS.

An STS involves social actors (*humans*) and technical entities (abstracted as AI *agents*) [29]. The agents represent the social actors and aim to facilitate rich interactions among them. Two key factors that influence social actors' interactions in an STS are **values** and **norms**. Values represent deep-rooted motivations or preferences of social actors (to act in a certain way). In contrast, norms govern expectations between actors. Norms usually reflect the values of the social actors, but they can also shape the values of social actors. Both values and norms influence the policies agents adopt for decision making as shown in Figure 3.

Literature on engineering STS studies the evolution of values and norms independently, e.g., [18, 22]. However, their interplay (bidirectional) – how values inform norms and how norms influence values [30] – is largely unexplored. In this report, we identify key research avenues to conceptualise this interplay, model an STS with this conceptualisation, and engineer the agents in the STS to co-evolve values and norms.

5.1.2 Background

Values are generally considered as high-level motivations that drive human behaviour [33]. Value preferences describe the relative importance that a human ascribes to different values to guide their actions in a socio-cultural environment and context. Values in society are operationalised at the agent-level by aligning their actions with individuals' value preferences, and at the STS level by expressing and enforcing norms to regulate agents' behaviour and interactions [6, 19].

Research in (normative) MAS has explored several approaches to model and compute the norms required to make coordination between agents possible [8], to address norm violation and sanctioning [17, 31, 2, 1, 36], and to support aspects pertaining to the dynamic adaptation of norms, including a variety of centralised and distributed approaches for norm change, revision, emergence, and learning [4, 10, 32, 11, 5, 16, 13, 27, 24, 39, 38].

Research has also explored approaches to infer human values [22, 23] and to relate norms to values [34, 20, 37, 3]. Further, values have been shown to affect policy and norms [12]. Despite the extensive literature on norms and values, the study of the interplay and co-evolution of norms and values over time still remains largely under-explored.

5.1.3 Main discussion points

This section summarises the key points that emerged from our discussion. This discussion led us to the conceptualisation and research questions outlined later.

1. Distinction between evolution, adaptation, emergence, learning, and change of norms.
   - Evolution is emergent, and adaptation is more top-down.
   - Evolution is a gradual process, while adaptation does not have to be.
   - Adaptation refers to a change in norms.
2. Distinction between top-down and bottom-up norms creation.
   - In the top-down case, norms come from an institution that regulates agents' behaviour. In the bottom-up case, norms emerge or are agreed upon by the agents themselves in an agent-centric (distributed) regulation.
3. What triggers the need for adaptation?
   - Norms may change when they become incompatible with values. Values may change based on the perception of others.
4. A norm is understood to have emerged when a certain percentage of the population (e.g., 90%) adopts it [21]. This notion refers to the idea of social tipping points.
5. The size of inner and outer groups affects the strengths of sanctions, the vulnerability perception of inner groups, and the willingness to violate norms.
6. There is a difference between revealing actions and values underlying those actions.
7. While there is some work on norm change ([28, 25, 14, 7, 26]), less work is present on value change.
8. The co-evolution of norms and values is relevant for a variety of case studies. These include
   - Hospital and Healthcare scenarios [9], where evolving human norms and values need to be considered to ensure that technology adequately supports patients and citizens. As an example, we considered *Family bounding* and *privacy* as possible values in this context. Norms could relate to permissions to share data with family. Preferences express individuals' preferences over values, such as *family bounding* over *privacy*.

- Human-AI teams [35, 15], where it is essential that team members have a common understanding of other team members and their individual preferences and values, team norms and team objectives (incl. team and organisational values) to ensure that the team operates (increasingly) effectively over time and that both AI agents can adequately support humans in the teams.
- University settings, where individual norms and values interact and coexist with organisational values.
- Legal settings, where personal norms and values co-evolve with societal values and legal norms. We consider a scenario where an individual values *traditional family*, and legal norms in their country prohibit abortion. Individual actions could include abortion (an action that violates a norm within a traditional family), but also protest against abortion (an action that is in line with their values, but inconsistent with other actions). One such situation expresses a choice to violate the norm without the desire to change it from an individual point of view. This situation could be an indication of an incoherence between individual values and norms, potentially leading to a change in one or the other.

9. Considering multiple actions being executed by multiple agents concurrently or sequentially (i.e., the temporal aspect of action execution from multiple agents) is important to highlight the sociotechnical nature of the problem of norm-values coevolution and their relation with agents actions, and the multiplicity and interactions between agents. As a simple illustrative example, we considered the case of a fight resulting from two persons moving their hands at the same time.
10. If we were to integrate norms and values in an agent architecture, such as a BDI agent, where would they live? According to Schwartz [33], values can be interpreted as beliefs. The norms literature sometimes associates values with Desires, although implementations may not follow this. Most of the literature considers Beliefs as information, knowledge, while Desires are considered objectives to achieve/comply with if possible, Intentions are intended as plans to follow and adhere to.
11. Norm changes can lead to values change. Popularity is not enough to change norms, the general change in values may lead to the change in norms. Accountability is essential to a norm. Values may not even change for attitudes to change. What could change is how human/agent sees or *perceives* the values in the view of new experiences or situations.
12. Affordance may enable the change of attitudes but also may make possible or not change of values and norms.

5.1.4 Conceptualisation

Figure 4 shows our conceptualisation of how values and norms interplay in an STS. To start with, each agent is endowed with a representation of the values of its principal. This endowment can happen in various ways, including learning from the human-agent interaction. Similarly, the STS also starts with a set of norms specified by the organisation. These initial norms can be established or learned via, e.g., the negotiations among the stakeholders.

In this setting, one key challenge is to enable an agent to acquire a policy (how to act) that aligns with its principal's values and complies with the organisational norms. However, often a policy may not be able to accomplish both objectives, i.e., (sequences of) action(s) may align with a value but deviate from a norm, or comply with a norm but deviate from values. These circumstances provide an opportunity for value and/or norm revision.

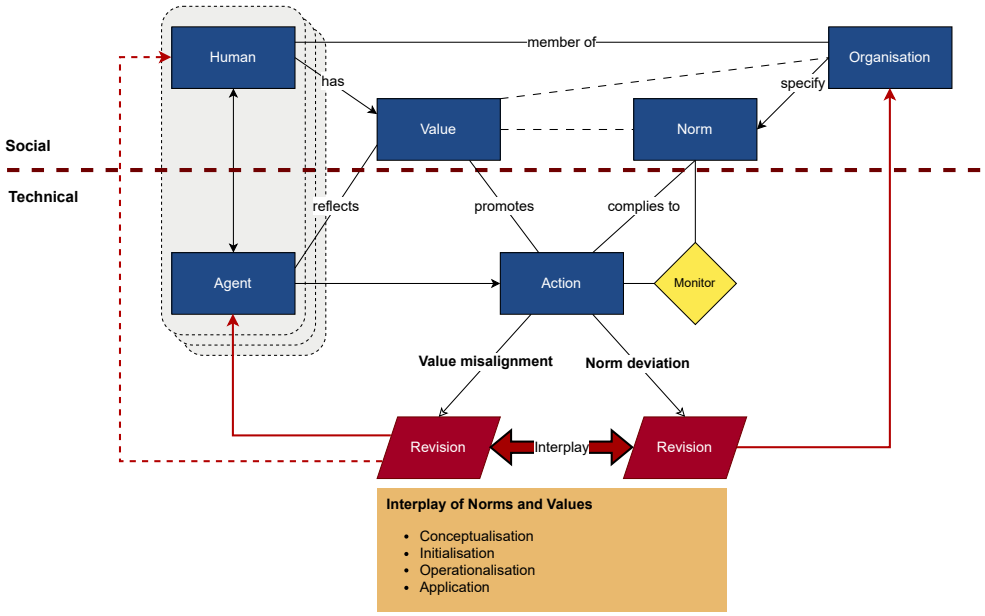

Figure 4 A conceptualisation of the interplay between policies, values, and norms in an STS.

We understand norm revision as a process that can be executed by an organisation in the STS that specifies the norms. Norm-revision can be informed by a variety of aspects, such as the values of individual agents in the organisation (if known), by the observation of agents complying or violating the norms, by their effectiveness in achieving organisational objectives, and by aspects such as popularity or affordances. Value revision, on the other hand, happens internally to the individual agents and can take different forms, such as changing the preference order between different values or changing (increasing or decreasing) the strength of a preference. Value revision, in this sense, is informed by the current norms that are enforced in the STS, by the human the agent is representing, and by the actions available to the agent, among others.

5.1.5 Challenges

Norms and Values alignment and interplay

This category of challenges refers to the conceptualisation of the problem of norm and value change and their interplay.

Human/Conceptual/Theoretical aspects

- What is the link between values and norms? We consider interplay in both directions, i.e., norms to value, and values to norms?
- How do affordances affect values?
- How does value change and clarification drives norm change?

Computational aspects

- How to model and reason about the alignment between agent policies, norms, and values in Intelligent Agents?
- How to characterise the interplay between norms and values when different types of norms (e.g., social vs more regulative norms) are considered?

- How to measure (in)coherence between an agent's actions, norms and values? For example, the temporal aspect matters when evaluating actions with respect to values.

Computational triggers and reasons for change

This category of challenges refers to the modelling of computational triggers that may cause norm and value change.

- When (if) to change norms and values?
- To what extent do public/private norm violations or compliance can initiate change?
- What are possible triggers/drivers of policy/value change?
- Observation of violation, compliance, sanction, other agent's actions
- What is a mechanism that uses a measure of incoherence/misalignment/asymmetry between norms and values as a pressure for change?

Computational norms and values change

This category of challenges refers to the computational mechanisms to implement actual norms and values change.

- How to change norms and values (value preferences rather than values)?
- How to ensure norms change but still within the boundaries and objectives of the intended system?
- How to ensure the system remains fit for purpose?

Computational dynamics and effects of change, from local to system-level and back

This category of challenges refers to the effects of norms and values change on the agents and on the STS as a whole, and to the resulting dynamics between norms, values and agents behaviour. These dynamics could be studied in a controlled setting for instance via (social) simulations, and in less controlled settings via longitudinal human-AI interaction studies.

- How can change in norms and values propagate from local groups to the larger society/organisation, and possibly back?
- When/how to move from convention to social norms?
- How to facilitate "integration" of new members (with their norms and values) into a group, and how new members affect the group norms and values?
- Can an AI that is able to reason about alignment of actions to norms and values and their dynamics, help people expose their reasoning about their values?

5.1.6 Future Plans

Our next steps involve (1) refining the research challenges identified above into a structured research agenda, and submitting it to, for instance, the AAMAS Bluesky track; (2) forming smaller working groups to focus on specific research areas; (3) organizing an online seminar series; and (4) preparing research proposals for funding calls.

References

1 Rishabh Agrawal, Nirav Ajmeri, and Munindar P. Singh. Socially intelligent genetic agents for the emergence of explicit norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 10–14, Vienna, July 2022. IJCAI.

2 Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Robust norm emergence by revealing and reasoning about context: Socially intelligent agents for enhancing privacy. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 28–34, Stockholm, July 2018. IJCAI. doi:10.24963/ijcai.2018/4.

3 Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 16–24, Auckland, May 2020. IFAAMAS. doi:10.5555/3398761.3398769.

4 Giulia Andrighetto, Sergey Gavrilets, Michele Gelfand, Ruth Mace, and Eva Vriens. Social norm change: drivers and consequences, 2024.

5 Duangtida Athakravi, Domenico Corapi, Alessandra Russo, Marina De Vos, Julian Padget, and Ken Satoh. Handling change in normative specifications. In *International Workshop on Declarative Agent Languages and Technologies*, pages 1–19. Springer, 2012.

6 Cristina Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2005.

7 Jordi Campos, Maite Lopez-Sanchez, Maria Salamó, Pedro Avila, and Juan A. Rodríguez-Aguilar. Robust Regulation Adaptation in Multi-Agent Systems. *ACM Transactions on Autonomous and Adaptive Systems*, 8(3):1–27, September 2013. doi:10.1145/2517328.

8 Amit Chopra, Leendert van der Torre, Harko Verhagen, and Serena Villata. *Handbook of normative multiagent systems*. College Publications, 2018.

9 Amit K Chopra and Munindar P Singh. Accountability as a foundation for requirements in sociotechnical systems. *IEEE Internet Computing*, 25(6):33–41, 2021.

10 Rosaria Conte, Giulia Andrighetto, and Marco Campennì, editors. *Minding Norms: Mechanisms and dynamics of social order in agent societies*. Oxford University Press, 2013.

11 Domenico Corapi, Alessandra Russo, Marina De Vos, Julian Padget, and Ken Satoh. Normative design using inductive learning. *Theory and Practice of Logic Programming*, 11(4-5):783–799, 2011.

12 Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: values, norms and culture in multi-agent systems. *Artificial intelligence and law*, 21:79–107, 2013.

13 Davide Dell'Anna, Natasha Alechina, Fabiano Dalpiaz, Mehdi Dastani, and Brian Logan. Data-driven revision of conditional norms in multi-agent systems. *Journal of Artificial Intelligence Research*, 75:1549–1593, 2022.

14 Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. Runtime revision of norms and sanctions based on agent preferences. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 1609–1617. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

15 Davide Dell'Anna, Pradeep K Murukannaiah, Bernd Dudzik, Davide Grossi, Catholijn M Jonker, Catharine Oertel, and Pınar Yolum. Toward a quality model for hybrid intelligence teams. In *23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024*, pages 434–443. ACM Press Digital Library, 2024.

16 Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. Runtime revision of sanctions in normative multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 34:1–54, 2020.

17 Thiago Freitas dos Santos, Nardine Osman, and Marco Schorlemmer. Is this a violation? learning and understanding norm violations in online communities. *Artificial Intelligence*, 327:104058, 2024.

18 Paul R. Ehrlich and Simon A. Levin. The evolution of norms. *PLOS Biology*, 3(6), 2005. doi:10.1371/journal.pbio.0030194.

19 Sven Ove Hansson. *The structure of values and norms*. Cambridge university press, 2001.

20 Samaneh Heidari, Maarten Jensen, and Frank Dignum. Simulations with values. In *Advances in Social Simulation: Looking in the Mirror*, pages 201–215. Springer, 2020.

21 James E Kittock. Emergent conventions and the structure of multi-agent systems. In *Proceedings of the 1993 Santa Fe Institute Complex Systems Summer School*, volume 6, pages 1–14. Citeseer, 1993.

22 Enrico Liscio, Roger Lera-Leri, Filippo Bistaffa, Roel I.J. Dobbe, Catholijn M. Jonker, Maite Lopez-Sanchez, Juan A. Rodriguez-Aguilar, and Pradeep K. Murukannaiah. Value inference in sociotechnical systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1774–1780, London, 2023. IFAAMAS.

23 Enrico Liscio, Luciano C. Siebert, Catholijn M. Jonker, and Pradeep K. Murukannaiah. Value preferences estimation and disambiguation in hybrid participatory systems. *Journal of Artificial Intelligence Research*, 82, April 2025. doi:10.1613/jair.1.14958.

24 Mehdi Mashayekhi, Nirav Ajmeri, George F. List, and Munindar P. Singh. Prosocial norm emergence in multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 17(1–2):3:1–3:24, June 2022. doi:10.1145/3540202.

25 Mehdi Mashayekhi, Hongying Du, George F. List, and Munindar P. Singh. Silk: A simulation study of regulating open normative multiagent systems. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 373–379. AAAI Press, 2016.

26 Javier Morales, Michael Wooldridge, Juan A. Rodríguez-Aguilar, and Maite López-Sánchez. Off-line synthesis of evolutionarily stable normative systems. *Autonomous Agents and Multi-Agent Systems*, June 2018. doi:10.1007/s10458-018-9390-3.

27 Andreasa Morris-Martin, Marina De Vos, and Julian Padget. Norm emergence in multiagent systems: A viewpoint paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 33(6):706–749, 2019.

28 Andreasa Morris-Martin, Marina De Vos, Julian Padget, and Oliver Ray. Agent-directed runtime norm synthesis. In *AAMAS23*, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

29 Pradeep K. Murukannaiah, Nirav Ajmeri, Catholijn M. Jonker, and Munindar P. Singh. New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, pages 1706–1710, Auckland, 2020. IFAAMAS.

30 Pradeep K. Murukannaiah and Munindar P. Singh. From machine ethics to internet ethics: Broadening the horizon. *IEEE Internet Computing*, 24(3):51–57, 2020. doi:10.1109/MIC.2020.2989935.

31 Luis G Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K Kalia, Jaime S Sichman, and Munindar P Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review*, 31(2):142–166, 2016.

32 Bastin Tony Roy Savarimuthu and Stephen Cranefield. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent and Grid Systems*, 7(1):21–54, 2011.

33 Shalom H Schwartz and Wolfgang Bilsky. Toward a universal psychological structure of human values. *Journal of personality and social psychology*, 53(3):550, 1987.

34 Marc Serramia, Maite Lopez-Sanchez, and Juan A Rodriguez-Aguilar. A qualitative approach to composing value-aligned norm systems. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems*, pages 1233–1241, 2020.

35 Munindar P Singh. The intentions of teams: Team structure, endodeixis, and exodeixis. In *ECAI*, volume 98, page 303. Citeseer, 1998.

36 Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. Norm enforcement with a soft touch: Faster emergence, happier agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1837–1846, Auckland, May 2024. IFAAMAS. doi:10.5555/3635637.3663046.

37 Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. Value-based rationales improve social experience: A multiagent simulation study. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, pages 3612–3619, Santiago de Compostela, October 2024. IOS Press. doi:10.3233/FAIA240917.

38 Jessica Woodgate and Nirav Ajmeri. Combining normative ethics principles to learn prosocial behaviour. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1–3, Detroit, May 2025. IFAAMAS.

39 Jessica Woodgate, Paul Marshall, and Nirav Ajmeri. Operationalising Rawlsian ethics for fairness in norm-learning agents. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2382–26390, Philadelphia, February 2025. AAAI. doi:10.1609/aaai.v39i25.34837.

5.2 Social Agentic Systems

Matthew Arrott (Coactive Computing, US), Matteo Baldoni (University of Turin, IT), Victor Charpenay (Mines Saint-Étienne, FR), Amit K. Chopra (Lancaster University, GB), Timotheus Kampik (SAP Berlin, DE & Umeå University, SE), Nadin Kokciyan (University of Edinburgh, GB), Ken Satoh (Research Organization of Information and Systems – Tokyo, JP), Jaime Sichman (University São Paulo, BR), Munindar P. Singh (North Carolina State University – Raleigh, US), and Pinar Yolum (Utrecht University, NL)

5.2.1 Introduction

Agentic computing is a recent paradigm for leveraging generative AI to build agents that accomplish sophisticated tasks. Today's agentic systems are impoverished because of a focus on rigid coordination methods based on procedural encodings of interaction, such as via task graphs. We introduce the notion of social agentic systems (SASY). SASYs are systems of agents that explicitly represent and communicate about norms, including the consequences of respecting or violating a norm. When needed, agents may deviate from norms and can negotiate their relaxations. In this way, SASYs provide a way to leverage the flexibility and power of generative AI through high-level, socially inspired models of interaction.

With the advent of software systems based on generative AI and LLMs, autonomous agents and MAS have re-emerged as one of the key research fields that will influence the future of large-scale information systems. Indeed, substantial expectations are attached to the importance that AI agents will have on how individuals and organisations carry out collaborative knowledge work. At the same time, concerns are growing that unjustified expectations regarding mainstream agent technologies will lead to failures of agent introduction projects; relatedly, the labelling of conventional (i.e., non-agent) technologies as "agents" is sometimes called *agent washing* [12].

These concerns raise the question of what is fundamentally needed to deploy software agents successfully as part of STS [17], such that human productivity and well-being are

indeed affected substantially and sustainably. In this chapter, we argue that a key capability – missing in current mainstream agent and MAS architectures and frameworks – is the ability to represent, reason, and communicate about norms and social values, in order to effectively operate and evolve the STS. We call STS that feature software agents with such capabilities social-agentic systems (SASY).

Based on a motivating example, we explain why the aforementioned capabilities are crucial for the successful deployment of intelligent software agents within and across organisations. We then explain why and how SASY differ architecturally from classical MAS, as well as from LLM-based agent architectures. Finally, we outline a set of research challenges that can move us towards real-world SASY.

5.2.2 Agentic Systems

An LLM-based chatbot is not only able to interact with humans, but it can also interact with technical systems by generating structured output such as code snippets: the chatbot generates a code snippet that is executed in a sandbox environment, giving controlled access to the technical system, and the result of the execution is given back to the chatbot in a textual form. In this setting, the chatbot becomes an agent, perceiving and acting on its environment [19].

Many agent frameworks have been built on this idea in the past two years, including LangGraph, AG2 (previously branded AutoGen), the OpenAI Agents SDK, Smolagents, and CrewAI. These frameworks encourage modularity in diverse respects. An agent interacts with its environment via a collection of *tools*, each providing access to a particular functionality. For example, LangGraph provides built-in tools for Web search, Web scraping, API access, code interpretation, and database access [11]. Likewise, the Model Context Protocol (MCP) [1] facilitates the integration of external tools into an agent's environment. Another way modularity is encouraged is by decomposing task handling into components, each component being made of a chatbot with its own context. Such a modular agentic system is sometimes referred to as a "MAS" in this literature (though at variance with the terminology in the MAS community), though from a more general point of view, it is indistinguishable from a single agent.

Even though LLM-based agentic systems perform much more poorly than humans, they achieve surprisingly good performance on several benchmarks. For instance, on the General AI Agent (GAIA) benchmark, GPT-4 correctly answers 30% of the questions (whose answers require Web search and reading documents in various formats) [14]. On the more general AgentBench benchmark, GPT-4 and Claude 3 achieve 14% to 70% of the tasks, depending on the domain, ranging from puzzle solving to Web browsing [13]. OpenAI and Anthropic provide commercial support for personal assistants evolving in a Web or computer environment. Planning a trip, which includes online reservation and payment, is one of the use cases advertised by the two companies, for example. Such a use case may – or, rather, must – include numerous social interactions (via the Web). Yet, LLM-based agentic systems demonstrate no form of social awareness.

5.2.3 Motivating Example

Let us consider the following scenario:

> A researcher Jaime, who works at the University of São Paulo, located in Brazil, is invited to a Dagstuhl Seminar. Pınar, a researcher who works at Utrecht University, located in the Netherlands, is invited to the same seminar. The two of them are preparing a joint research proposal, and hence have agreed to have a first meeting on the Sunday evening prior to the seminar.

To promote *sustainability*, Utrecht has an internal regulation that trips to cities that are less than 700 km away from Utrecht must be made by train. To promote *reasonable usage of public resources*, São Paulo has an internal regulation that researchers must buy economy flights. Additionally, Jaime has a preference to always fly with a certain airline, since his membership in its loyalty program allows him to upgrade his ticket. Both Jaime and Pınar would like to book a taxi together from the train station at Türkismühle to Dagstuhl.

In principle, this scenario presents *social constructs* that have been studied by the MAS community for the last 40 years [10, 9, 7]:

Institutional norms that must be taken into account during the agents' deliberation and decision-making.

Institutional or individual values to consider in choosing a solution.

Commitments between two parties, representing that one party legitimately expects that the counterparty will act accordingly.

Individual preferences to prioritise when several solution options are feasible.

Suppose the train that Pınar has booked is late, and this would prevent her from sharing a taxi with Jaime, as they initially planned. Suppose also that the taxi service in Türkismühle closes at 20:00 on Sundays and reopens on Monday morning. Current agentic systems do not appropriately cope with such diverse social concepts and changes in plans.

An intelligent agent would need to reconsider, either autonomously or by asking the user, which adaptations should be made to the original joint plan: should Pınar take an earlier train? Should she take a flight instead of a train? That is, how might she deliberate on whether to violate a norm to guarantee the meeting? Moreover, if a taxi reservation has been made for an earlier time, this commitment should be revised and communicated to the taxi service.

We posit that leveraging such *social constructs* is crucial to addressing the challenges facing agentic AI.

5.2.4 SASY Architecture

We define a SASY as a MAS consisting of human and software agents, in which software agents engage in communication dialogues that run over extensive periods of time and in which their actions, including speech acts, are grounded in norms, values, and trust (or lack thereof). The internal structure of agents in SASY could be designed in various ways; that is, a SASY may feature agents of diverse architectures. We assume that each agent has access to the environment where it can communicate with other agents. Agents have knowledge of the domain they are operating in, they have capabilities, and can reason about what action to perform. The domain reasoner could be a traditional reasoner, such as a rule-based engine, or a more sophisticated one using the capabilities of neural models. Some agents may have a social state component where they have an explicit representation of social preferences,

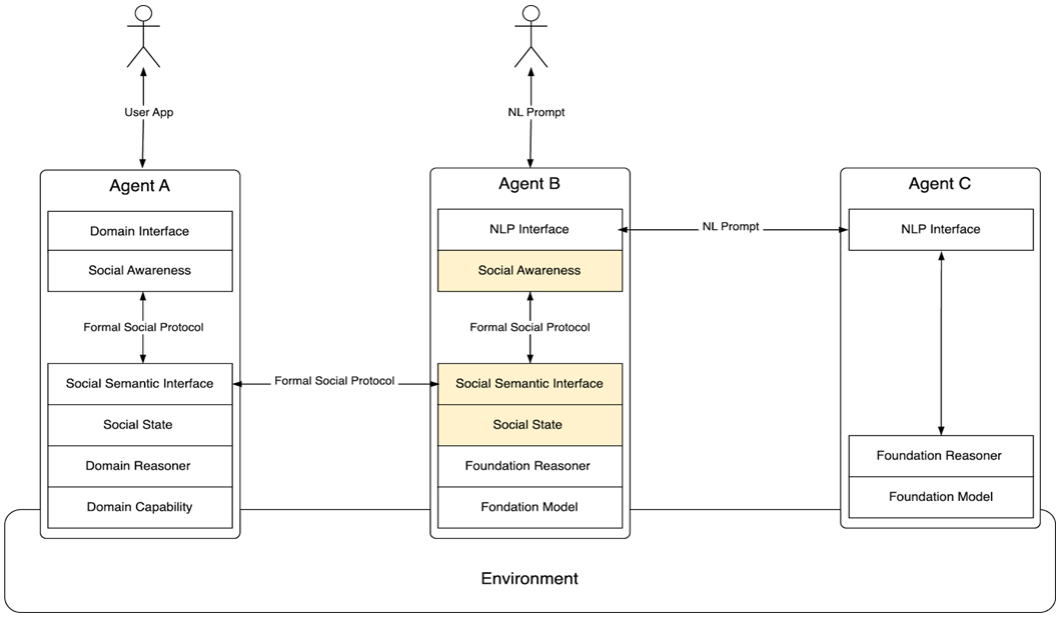

Figure 5 Communication between different types of agents. SASY enables the creation of socially-aware LLM agents, that is, of the type of Agent B.

norms, values, commitments, and so on. An agent may have an interface to interact with users. Such an interface could be a built-in application interface or an interactive interface, such as a chatbot, where the interactions occur through natural language. Moreover, when agents interact with humans, they gather additional information about these humans to become socially aware by encoding the information into formal specifications that could be used to communicate with other agents.

Let us assume we have three different types of agents as depicted in Figure 5. All agents have access to the environment, where they can communicate with each other through well-defined protocols. Agent A is an agent that is using a user application to interact with the user to process the user requests. The *Social Awareness* component includes information such as user preferences. A formal social protocol ensures that the user state is translated to a formal state, which could be used to communicate with other agents in the STS. Agent B supports a natural language interface for the user. Agents B and C conform to the SASY system architecture, since they are both equipped with a natural language interface. Agent B is using a natural language interface to communicate with the user, whereas Agent C processes natural language prompts, but it does not interact with the user directly.

The literature shows examples of Agent A [4] and Agent C [11]; however, research challenges remain. Architectural components that require further attention from the research community are highlighted in Agent B. For example, if Jaime and Pınar both had agents of type Agent B, their agents could communicate with each other and also with their users to adjust the plan according to the current social context.

In the example from Section 5.2.3, a SASY software agent could address the situation as follows:

- Decide that booking a flight is undesirable, not only because it violates the *sustainability* norm, but also because it is, from a commonsense perspective, not *comfortable* in the current situation and not a *reasonable use of public resources*. Whereas the latter two

norms are not formally represented in the norm base the agent has access to, the agent generates them on the fly using an LLM and presents them as explanations for the final proposal to Jaime and Pınar.
- Decide that the revised taxi trip is only marginally noncompliant with the operating hours of Taxi Martin and use this inference as the basis for a successful request (carried out via an interface to the traditional telephone system) that re-negotiates the schedule and conditions for the taxi trip.

5.2.5 Challenges

In order to realise the vision outlined above, agents must exhibit social awareness in human-agent interactions, which raises several research questions:

Prompting. How do we adapt the prompts with social constructs, such as norms and values? What are the different ways to adapt the prompts to ensure social components are represented adequately? Some options are enhancing the prompts, revising the prompts, and so on.

Social interactions. How do we design components to keep track of or regulate social interactions? Can an LLM figure out that it is making a commitment when it is saying certain things to the user? Consider the recent Air Canada case [8], where the airline's chatbot told a customer they could apply for a refund, in contradiction to the company's policy. Could we keep such commitments in a database and keep track of them so that the LLM can decide whether to make or break these commitments with its interactions? Commitments here are one such abstraction; we can also think of consent [2, 18] and other such abstractions similarly. For example, a consent store can prohibit an LLM from generating certain results or communicating what may have been generated.

Social. How may LLMs interact with this social state? Some possibilities are below.
- The LLM generates what it will normally generate and then sends it to the social awareness component, which then checks whether this is appropriate or not. It could be a binary decision or a modification to the output.
- Social awareness component generates (additional) prompts to the LLM to take into account so that the generated output is socially appropriate.

Validation. How do we assess that the proposed architecture delivers intended actions or outputs? Typical LLM evaluation is with benchmarks on input and output. This could be one way to evaluate SASY, but to specifically evaluate the social awareness component or the interface, we might need different techniques. For example, if the SASY produces socially appropriate output, is it because the social awareness component caught something and fixed it? Or, did the LLM generate it that way to begin with?

Languages. What are some languages or protocols that could help in realising this architecture? Formal languages to represent norms [5, 6]remain useful for verifiability and accountability reasons, even in the presence of a natural language interface. The Blindingly Simple Protocol Language (BSPL) [15, 16] provides an alternative to workflows. Translating natural language into a formal language and back is an important challenge. Some LLM services are already capable of outputting structured data, validating a schema defined at run time (e.g., ChatGPT Structured Outputs), but the support is inadequate. A typical problem arising in this context is terminological alignment [3].

In addition to serving as a personal assistant as described above, the agentic architecture could also help in improving various design stages of STS. For example, SASY could be used to simulate various stakeholders of an STS, adopting various personas, depicting

different interactions, and realising diverse scenarios. SASY could be used to simulate various alternative evolutions of an STS. Through such simulations, it may be possible to infer various side effects on primary stakeholders. Moreover, it may help in identifying secondary stakeholders, that is, those who would be affected by the use of the system.

References

1 Anthropic. Model Context Protocol (MCP), July 2025. Accessed 2025-07-11. URL: https://modelcontextprotocol.io/.

2 Anastasia Sophia Apeiron, Davide Dell'Anna, Pradeep K. Murukannaiah, and Pınar Yolum. Model and mechanisms of consent for responsible autonomy. In *Proceedings of the 24th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 133–141, Detroit, May 2025. IFAAMAS. doi:10.5555/3709347.3743525.

3 Victor Charpenay. Génération et validation de données structurées. In *36es journées francophones d'Ingénierie des Connaissances (IC)*, 07 2025. URL: https://hal-emse.ccsd.cnrs.fr/emse-05163532v1/document.

4 Amit K. Chopra, Matteo Baldoni, Samuel H. Christie V, and Munindar P. Singh. Azorus: Commitments over protocols for BDI agents. In *Proceedings of the 24th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 490–499, Detroit, May 2025. IFAAMAS. doi:10.5555/3709347.3743564.

5 Amit K. Chopra and Munindar P. Singh. Cupid: Commitments in relational algebra. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI)*, pages 2052–2059, Austin, Texas, January 2015. AAAI Press. doi:10.1609/aaai.v29i1.9443.

6 Amit K. Chopra and Munindar P. Singh. Custard: Computing norm states over information stores. In *Proceedings of the 15th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1096–1105, Singapore, May 2016. IFAAMAS. doi:10.5555/2936924.2937085.

7 Amit K. Chopra, Leon van der Torre, Harko Verhagen, and Serena Villata. Normative multi-agent systems (Dagstuhl seminar 15131). *Dagstuhl Reports*, 5(3):162–176, 2015. doi:10.4230/dagrep.5.3.162.

8 Civil Resolution Tribunal of British Columbia. Moffatt v. Air Canada, 2024 BCCRT 149 (CanLII), February 2024. Accessed 2025-09-02. URL: https://canlii.ca/t/k2spq.

9 Dov Gabbay, John Horty, Xavier Parent, Ron Van der Meyden, Leendert van der Torre, et al. *Handbook of Deontic Logic and Normative Systems, Volume 1*. College Publications, 2013.

10 Dov Gabbay, John Horty, Xavier Parent, Ron Van der Meyden, Leendert van der Torre, et al. *Handbook of Deontic Logic and Normative Systems, Volume 2*. College Publications, 2021.

11 LangGraph. LangGraph: Building language agents as graphs, December 2024. Accessed 2024-12-05. URL: https://langchain-ai.github.io/langgraph/.

12 Adrian Lee, Kip Martin, and David Yockelson. Stop agent-washing: Differentiate with human-centric agentic experiences. Technical report, Gartner Research, 2025. URL: https://www.gartner.com/en/documents/6819934.

13 Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents, 2023. URL: https://arxiv.org/abs/2308.03688, arXiv:2308.03688.

14 Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general ai assistants, 2023. URL: https://arxiv.org/abs/2311.12983, arXiv:2311.12983.

15 Munindar P. Singh. Information-driven interaction-oriented programming: BSPL, the Blindingly Simple Protocol Language. In *Proceedings of the 10th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 491–498, Taipei, May 2011. IFAAMAS. doi:10.5555/2031678.2031687.

16 Munindar P. Singh. Semantics and verification of information-based protocols. In *Proceedings of the 11th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1149–1156, Valencia, Spain, June 2012. IFAAMAS. doi:10.5555/2343776.2343861.

17 Munindar P. Singh. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):21:1–21:23, December 2013. doi:10.1145/2542182.2542203.

18 Munindar P. Singh. Consent as a foundation for responsible autonomy. *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, 36(11):12301–12306, February 2022. Blue Sky Track. doi:10.1609/aaai.v36i11.21494.

19 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models, 2023. URL: https://arxiv.org/abs/2210.03629, arXiv:2210.03629.

5.3 Conceptual Issues Regarding Policy Modelling and Reasoning in Sociotechnical Systems

Cristina Baroglio (University of Turin, IT), Mehdi Dastani (Utrecht University, NL), Frank Dignum (University of Umeå, SE), Beishui Liao (Zhejiang University, CN), Vivek Nallur (University College Dublin, IE), Judith Simon (Universität Hamburg, DE), Sz-Ting Tzeng (University of Umeå, SE), Harko Verhagen (Stockholm University, SE), and Jessica Woodgate (University of Bristol, GB)

5.3.1 Introduction

This working group focused on conceptual issues perceived to relate to policy modelling and reasoning in STS. The terminology problem was first discussed, as different stakeholders view STS differently which causes a lack of consensus in the use and misuse of a system. Difficulties in ascertaining what constitutes misuse of an STS, or when misuse is occurring, present barriers to deciding and enforcing effective regulation. The discussion went on to apply conceptual issues to an example of an educational STS, which encompasses dynamic change of goals and behaviour, influencing the technological decisions and process that must be made.

5.3.2 Boundaries, Mental Models, Agency

Efforts towards regulation can only address reasonably foreseeable misuse whilst leaving room for innovation. Yet, human behaviour is extremely hard to predict. In an STS composed of social (humans and organisations) and technical (data and devices) tiers, there will necessarily be a limit to the extent of what the human tier can understand, and what the technical tier can represent.

The lack of fixed boundaries in STS terminology leads to a lack of consensus about the use and misuse of a system. Technology-only solutions to STS misuse typically result in a blunt instrument approach, with the social components adopting, rejecting, complying, or violating the assumptions and norms (standards of expected behaviour [5]) of the STS. Humans create a mental model of the system (both technical and inter-related processes), yet it is unclear if the technical tier does, or is able to, represent such a model. Resulting problems from this include:

Deception: Both intended and unintended

Trust: Both deserved and undeserved

Goal and value alignment: Humans tend to assume that the technical system *ought* to have the humans' best interest at heart.

Can these problems be fixed with a notion of *group agency*? Agency ought to lead to responsibility, and hence group agency makes explicit the notion of shared responsibility [3]. Understanding how group agency emerges necessitates defining group membership, structure, and behaviour.

5.3.3 Normativity

Within an STS, stakeholders attempt to achieve normative goals, wherein normativity refers to something desirable. Normative goals may be imbued through law, social norms, or moral norms, and explicit or implicit incentives may be applied to influence behaviour and encourage stakeholders to comply with particular norms. The nature of an STS is an ongoing and dynamic process, where humans influence systems and those systems in turn influence human behaviour. Therefore, when a choice is made to encourage a normative goal in a system (e.g. installing smoke detectors to stop people smoking indoors), we can expect that people will adapt in response to that choice (e.g. by not smoking or by taking the batteries out of the smoke detector). Norms thus embody a dual role, effecting expected changes and catalysing unexpected responses to those changes. The technological choices we make in pursuit of normative goals will never be able to capture the whole space of normativity as people will loopholes and behave in unforeseen ways. Smoke alarms are installed to promote safety, but some people deviate by covering the alarm up or removing the power.

Communication fills the gap between the technology itself, and the normative goal aimed at, through sanctions (positive or negative reactions to approved or disapproved behaviour [2]) or other signals [4]. For example, speed signs may be put up to make people slow down outside a school. However, interventions may change over time as the system responds. If people do not follow the signs, the intervention may need to be fortified such as by putting in a speed bump.

Technological interventions to promote normative goals raise questions about intrinsic autonomy and deception. Whether a deception is (un)acceptable may depend on the intentions and reasons behind it, as well as the awareness of the subject that they are being deceived. On the one hand, end users should be aware of the limits of the technology they are engaging with. For example, some large LLMs now have a disclaimer that output may be misleading or false which alerts the user to the fact that LLMs are fallible. On the other hand, if the communication about the limits of a technology is incomplete, misunderstandings may arise. As the role of LLMs in everyday life increases (e.g. as the first result of a search engine), increasing exposure to the disclaimer that LLMs give inaccurate responses may reinforce the belief that people cannot trust the information they encounter, adversely affecting the way they interpret other sources of information such as the news. LLM outputs are deceptive in part because there is a detachment from the labelling of training data and the uncertainty of the output. Where a human would convey uncertainty in their communication, LLMs do not.

Whilst there are important limits to technological interventions to achieve normative goals, some people will still have a tendency to overestimate the capabilities of and defer to technology. It is thus key that efforts are made towards designing technology in ways that both benefit the system and acknowledge its limits.

5.3.4 Trust

The typical (philosophical) conception of trust is understood as:

A *trusts* B *with regard to* X

Trustworthiness has moral and an epistemic requirements:
Epistemic: To know whether something is competent and know the limits of that competence
Ethical: To know if the entity being trusted is honest and benevolent

It is critical to differentiate between `trust` and `trustworthiness`. Trust is an intentional stance taken by an entity, while trustworthiness is a property or a relation which perhaps can be measured. Trust in technology often reduces to reliance. Trust need not be compositional, *i.e.*, **A** can trust **B** without trusting all the constituent parts of **B**. In the application of system design, trust functions as a social enabler, allowing social entities to function without requiring third-party guarantees.

5.3.5 Value Alignment

There is some debate over whether global values really exist. Within this debate, the following are some important questions:
Understand: **How** do we define the value?
Define: Who decides **what** the relevant values are?
Alignment: A **joint-expectation** on what actions are plausible, and which ones are good
Reconcile: How do we decide whether the components of an STS are value-aligned, when there are **differences** between parts?

One possible characteristic of value is that it forms the criteria for comparing situations. A norm can be understood as a preference relation that pushes an entity towards actions leading to a situation where a value is upheld. This implies that the norms adopted influence the kinds of values that can be upheld. A value-conflict can therefore be defined as differences in estimates of goal states/situations, based on actions pushed by the norms. *Can value-alignment be defined as the absence of value-conflict?*

An alternate definition of value-alignment is that if two entities **A** and **B** take the same action, given the same context, one can form a belief that A and B are value-aligned. The presence of value-alignment creates the presence of an in-group *vis-a-vis* the out-group that (by definition) does not align with the same values. Members of the in-group have a common ordering over shared values.

It is not very well-understood how one should choose between values. Values that an entity is unwilling to negotiate on, are called non-negotiable values. These could also be viewed as sacred values.

5.3.6 Observation and Verification of Values

Designing a system given a set of shared values, considerations include how to encode the requirements that flow from those values, and discerning the evidence or data that is needed for the design of systems. It is unclear whether a value is something that can be observable

by perception, or is something that is completely cognitive. Agent architectures that could represent values include:

- Reactive
- Deliberative (*e.g.*, Belief Desire Intention, BDI)
- Symbolic
- Sub-symbolic

For sub-symbolic architectures, if the mechanism is very good then representation becomes invisible. Sub-symbolic reflects values present in the training data; values can be represented using reinforcement learning from human feedback (RLHF) [1]. Possible mechanisms for collating the training data for sub-symbolic approaches include social networks and purposeful collection.

Exploring whether values can be mixed in an STS, and if it is possible to validate a consistent mixing of values, is a gap in research. Possibly, validation of value-mixing can only be achieved in specific forms of value-failure (e.g., discrimination based on race or gender).

5.3.7 Argumentation-Based Methodology for Representing and Reasoning about Policies, Norms, Values, Disobedience, and Causality

Policies, norms, and values exhibit inherent potential conflicts and dynamic evolution within STS. Formal argumentation establishes a principled methodology for representing and reasoning about these entities, extending to norm violations and causal relationships. The integrated framework comprises:

Unified Representation: Norms, policies, and their violation conditions (disobedience) are modelled uniformly as defeasible rules. Value sets annotate norms, while priority relations operate over rules or their value associations

Conflict & Violation Resolution: Argumentation resolves conflicts between competing norms/policies and adjudicates norm violations. Embedded in belief-desire-norm-policy-intention (BDNPI) architectures, this enables autonomous agents that reconcile policy guidance with normative constraints, including disobedience detection and sanction reasoning

Dynamic Adaptation: Runtime modification of norms, policies, and violation thresholds is facilitated through argumentation revision mechanisms, maintaining system consistency during change

Causal Attribution: Argumentation frameworks can be extended to integrate causal inference for modelling responsibility attribution. This enables: (1) trust establishment via transparent causality chains; (2) precise accountability assignment for norm violations; and (3) root-cause analysis of system failures across multi-stakeholder interactions

Computational Viability: Locality-driven computation and modular design overcome complexity barriers, ensuring efficient handling of dynamic rule updates and causal reasoning

Neuro-Symbolic Integration: LLMs transform natural language norms/violation clauses into formal defeasible rules, while argumentation provides rigorous conflict and causality resolution – enabling hybrid reasoning unattainable by monolithic approaches

5.3.8 Governance of an STS

In governing an STS, it is important to acknowledge that any STS that needs to adapt to changing participants and technologies will become a complex system. If such a system is to be adaptable to change, then it needs to be governed rather loosely. A general guiding principle could be to *identify the forces that move the STS to some attractor point, and try*

to govern those forces. Diving deeper, we discussed the reasoning model required of agents, in an agent-based simulation of an STS. Some relevant questions include:

- How deep should agent reasoning attempt to go? Would reactive agents with appropriate calibration suffice? Or are BDI agents necessary?
- Can the two agent architectures be systematically bridged?
- What are the HPC requirements that agent-based models should consider, while being designed?
- Are there integrative techniques for merging sampling data, with deeper survey-type models?
- Which communities need to attempt integrating their techniques? (e.g., computer science, social psychology, behavioural economics)

It was recognised that we need to have some reference problems the community could attempt to solve. One suggestion was that the community agrees on a small set of small-but-complex systems, and attempt to answer each of the above questions systematically. This could possibly involve holding an iterated competition that starts from the same codebase each time it is held. Learnings from multi-year competitions could be applied to bigger and more realistic STS.

5.3.9 Policy Modelling

Policy modelling is a multidisciplinary field that synthesises insights from behavioural sciences, social simulation, reasoning, humanities, and history. In policy modelling, data doesn't mean reasoning. Agent behaviour for simulation can come from data, but it can also come from asking what people do and what is important to them. The latter covers humans' reasoning process.

In simulation, the input for how (human) agents reason can rely on basic assumptions (e.g. norm-obeying willingness based on political affiliation), theories (e.g. social-psychology theories), and actual data collected (e.g. using various social science methodologies). It is especially interesting and relevant for policy modelling in STS how (if at all) humans perform normative reasoning when making decisions. While there seems to be an agreement that integrating the different approaches would be beneficial for the fruitfulness of simulation, it is far from obvious how the integration should happen. Integrating diverse approaches is hence an open challenge for the community.

5.3.10 Education as a Use Case

Building on abstract conceptual ideas, the discussion transitioned to consider an educational STS. Education presents a useful case study of an STS that is dynamic and complex. In particular, we considered teaching computer science at university. Education and learning is a domain where behavioural change, due to technological decisions and processes, might be both short and long-term. An educational STS is a (as, we imagine, many STS are) complex system. The complexity means that we have no systematic way of determining which point of intervention is the best, or if the intervention will cause the future to be as we envisioned. Important questions discussed included:

- What are the challenges of an STS where each student might rely on one or more GenAI systems (supplied or not supplied by the university)?
- Would GenAI systems be considered as personal agents? How should GenAI be integrated into an STS?
- What are, or ought to be, the learning goals?

- How do we reach learning goals when tools change substantially, explicitly accounting for skills that we believe could be lost?
- How do we evolve or change goals over time? For example, reading and writing are currently considered important. Will they continue to be so? Are they important for critical thinking?
- Are we using tools as assistive technologies, or skill replacement technologies?
- What happens when our values and goals change? Some jobs/careers may diminish in importance when tasks are automated, so that they are not coveted anymore.

5.3.10.1 What Topics to Teach

The approach to programming changed rapidly in the presence of GenAI systems, does it make any sense to keep on teaching programming as in the past. So, do we need courses on programming and/or teach on the use of AI programming "tools" and have students (inter)act with the tools – pair-programming in a human-AI team as the future of software development. Some drawbacks are:

What to teach: Programming with Gen AI is a loop where the human uses the system to revise a drafted program improving it until he/she is satisfied with it. But, the roles are not equal: the human has responsibility over the product as well as the ownership in terms of copyright, for this reason he/she should have the ability to evaluate the program that is being built

Normative compliance: The GenAI system often produces over-complex code, difficult to understand or, more in general, not respecting some general norms, policies, or good practices we would like to be respected

Personal development: Companies want to hire persons that have high level skills, and who can work with abstractions so that they can understand if and how things fit together. Attractive candidates do not try to solve a problem on their own but rather know when and how to ask things to teammates. So, how does programming fit in here? Perhaps we should put forward more general learning outcomes that encompass critical thinking and analytical skills, as well as understanding loops and recursion

5.3.10.2 How to Teach

Suitable approaches could take inspiration from "the Amazon method", which involves starting with a mandatory brainstorming session and is followed by an idea collecting session. Teaching could use design thinking as a model, or focus on more student-centred methods and flipped classroom-inspired teaching. Currently, most programming courses have lab sessions with teaching assistants where students get help "on the fly" in constructing their program. Teaching assistants are increasingly receiving student requests to explain code that works when executed, but students do not understand why because it was produced with the help of an LLM. This is connected to the previous item – the ownership and understanding of what the code does are not matching. Thus, different strategies for teaching and for assessing what has been learned should be developed.

5.3.10.3 How to Assess

As LLMs have destroyed the essay (and in some discussions, the bachelor or masters thesis has also been declared dead), the assessment of programming skills may need to adapt. In an LLM-supported programming curriculum, it may not be all that important to check that a student wrote his/her code without help, as in companies they will probably use LLMs

and work in teams, so it is important that they practice with such tools. Yet, students should be able to explain the code, the choices made, why they are good, their limits, and so on. Similarly, it is important that students can review code written by someone else. More appropriate assessment could thus take the form of oral on-the-spot explanations of code. Such assessments should take place in a controlled environment.

We observe that when writing text, LLMs often mark a particular words or ways of phrasing things as a mistake, forcing a "standardised" way of writing. Similar instances may happen with writing code, limiting the exploration of learners. This is the issue of sausage production: GenAI tools repress individual expressions and start from "the mean", which results in everything becoming gray (and correcting correct items to incorrect or bland ones as a consequence).

On the whole, we would like students to learn to be in control of Gen AI systems when using them to produce code, they should be able to review code but for this aim they should know and have practice on how to code. They should be aware of the strengths and limits of such systems, and not be over-reliant on them.

5.3.11 Conclusions

Loose boundaries of terminology related to STS confuses how to define (mis)use of a system, which in turn makes STS difficult to regulate. Normativity, trust, and values are fundamental to STS, yet there are important challenges with attempts to granulate these concepts into entities that can be encoded. Formal argumentation could be useful to represent and reason about the conflicts associated with policies, values, and norms. An educational STS provides a helpful use case to explore how the conceptual issues discussed actualise and how challenges could be addressed.

References

1 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 1–9. Curran Associates, Inc., 2017.

2 Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. *The Knowledge Engineering Review (KER)*, 31:142–166, March 2016.

3 Raimo Tuomela. Joint Intention, We-Mode and I-Mode. *Midwest Studies in Philosophy*, 30(1):35–58, 2006.

4 Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. Norm enforcement with a soft touch: Faster emergence, happier agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1837–1846, Auckland, May 2024. IFAAMAS.

5 Georg Henrik Von Wright. Deontic logic: A personal view. *Ratio Juris*, 12(1):26–38, 1999.

5.4 Harmonising constraint and policy languages for the use by autonomous agents

Nicoletta Fornara (USI – Lugano, CH), Beatriz Esteves (Ghent University, BE), Sebastian Neumaier (FH – St. Pölten, AT), Victor Rodriguez Doncel (Polytechnic University of Madrid, ES), Simon Steyskal (Siemens AG – Wien, AT), Rigo Wenning (W3C / ERCIM, FR), and Antoine Zimmermann (Ecole des Mines – St. Etienne, FR)

5.4.1 Introduction

The subgroup of the seminar addressed questions arising from the use of constraint and policy languages. Presentations of the ODRL Policy language and the SHACL constraint language were given. The group discussed challenges around using SHACL and ODRL, how they relate, how they differ, and how they can be used in combination. We found that SHACL can at least serve to disambiguate the expression of constraints in ODRL. Both languages are W3C Recommendations.

A standardisation strategy discussion was triggered. It included the suggestions to amend the respective standards to ease the combination of ODRL and SHACL. Standardisation gaps were identified and the group discussed the formal semantics of ODRL and how to update the current W3C Recommendation. Given the work on RDF 1.2, the group explored how to use the new possibilities for annotation of data coming with those updates.

Web agents and the Web Agents Community Group and their work were presented. We discussed how policy and constraint languages can help in an agent scenario. New challenges stemming from web agents for policy and constraint management were identified. Annotation and protocol issues were at the core of our discussion.

A *policy language* is a formal or semi-formal language used to express various types of rules (i.e., permissions, obligations, and prohibitions) that govern access and usage of resources or regulate behaviour or state of affairs in open distributed systems. It typically includes constructs to define subjects, actions, objects, and conditions under which actions or states are allowed or denied. Policy languages are used in a wide range of domains, including access and usage control, data usage, privacy, and digital rights management.

5.4.2 Background

5.4.2.1 From Rights Expression Languages to Policy Languages

ODRL 1.1 (Open Digital Rights Language) was introduced as a *Rights Expression Language* (REL) designed to represent statements about the usage rights of digital content. Published as a W3C Note in 2002 [14], it emerged in the context of the growing interest in Digital Rights Management (DRM) systems during the early 2000s. These systems required standardised means to express permissions and conditions associated with digital assets.

ODRL 1.1 was based on XML and offered a structured and extensible way to define rights and conditions. Its clean and modular specification contributed to its adoption in several industry sectors. Notably, ODRL 1.1 was incorporated into the Open Mobile Alliance (OMA) DRM specifications [23], where it served as the core rights language for managing usage rights on mobile devices. This adoption demonstrated the practical applicability and interoperability of ODRL 1.1 across platforms and devices even without having a formal semantics, the specification was clear for everyone and the interoperability problems were minimal or non-existent.

The XML-based ODRL 1.1 was later replaced by a more expressive RDF-based version, formalised in the ODRL 2.2 ontology [22]. This revision was not merely syntactic; it introduced significant enhancements in the language's expressive power. In particular, the core concept of a *rights expression* was replaced by the more general notion of a *policy expression*. While `odrl:permission` had been sufficient for access control scenarios, ODRL 2.2 introduced additional constructs such as `odrl:prohibition` and `odrl:obligation` (also referred to as duties), allowing for richer and more nuanced representations of policy constraints. These expanded capabilities significantly increased the applicability of the language across broader domains. However, they also introduced greater interpretive complexity and potential ambiguity in how policy expressions should be understood and enforced.

5.4.2.2 The ODRL 2.2 Policy language

The Open Digital Rights Language (ODRL) [15] is the policy language for the Web specified by the W3C[4] for expressing permissions, prohibitions, duties, and restrictions associated with digital content. ODRL is used in contexts such as rights management, licensing, and, more recently, data sharing agreements [26]. It is specified in two W3C Recommendations: the Information Model [15] and the Vocabulary & Expression [16], which are primarily text documents, but also include an OWL ODRL Ontology[5] and a number of profiles and supplementary documents, such as [17, 27, 6, 13, 21].

The syntax of valid ODRL policies must therefore conform not only to the rules formally specified in the OWL ontology, but also to additional constraints described in natural language in the specification. For example, the OWL ontology states that the domain of the `odrl:constraint` property must be either an `odrl:Policy` or an `odrl:Rule`. However, the ontology does not impose that a policy must contain at least one rule, although this requirement is expressed in the textual specification [15, Section 2.1]. Many such textual constraints can be represented as SHACL shapes used for validation,[6] but these rules are not formally standardised. Thus, it is fair to say that ODRL is a *semi-formal language*.

5.4.2.3 Formal Semantics for ODRL

The ODRL specification refers to an "ODRL Validator" –a system that checks the conformance of ODRL Policy expressions – but its behaviour is even less precisely defined. Providing a fully formal semantics for this software component would bring significant interoperability benefits. This need led to the creation, in 2021, of a draft *Formal Semantics for ODRL* report [10], which is still under development and was a topic discussed at the seminar. The semantics of ODRL 2.2 can be specified in a *declarative* format (as in [1]) or can be specified in an *operational* format by providing an algorithm for translating ODRL policies into another language that has a formal semantics, for example SPARQL.

5.4.2.4 Constraint languages in the Web

A *constraint language* is a language that can be used to express conditions or integrity constraints over data structures. These constraints can be used for validation, consistency checking, or to define requirements that data must satisfy. Constraint languages are often declarative and are used to ensure that (structured) data adheres to specified rules, independently of any procedural behaviour.

---

[4] World Wide Web Consortium, `https://www.w3.org/`
[5] `https://www.w3.org/ns/odrl/2/`
[6] ODRL Implementation, `https://odrlapi.appspot.com/`

SHACL (Shapes Constraint Language) [20] is the W3C recommendation used to express constraints over RDF data. It allows for the specification of conditions that RDF graphs must satisfy and is commonly used for data validation. SHACL 1.2, a Working Draft as of July 2025 [19], extends the original specification with several significant features aimed at increasing expressivity, usability, and integration with RDF and SPARQL.

In the context of ODRL, SHACL has been used to encode additional constraints that are specified in the natural language part of the ODRL specification but not captured by the OWL ontology.

5.4.2.5 Automated Regulatory Compliance

Automated regulatory compliance faces several fundamental challenges. These potentially stem from the inherent complexity of legal texts and the diversity of technical systems that the legal text aims to govern. Regulations are typically written in natural language, using ambiguous terms, implicit assumptions, and context-depended information. Translating the natural language legal provisions into a formal language like SHACL or ODRL is not an unambiguous operation. Translating a legal text may not always result in the same formal language file. The result of a transformation of a legal compliance requirement is thus necessarily and always an interpretation of the law creating that compliance requirement.

On the other hand, compliance is not just about the implementation of static checks (e.g., the completeness of meta-information); many obligations depend on dynamic and evolving system behaviour (for instance, consider updates in the provenance of training data, or in the deployment process of AI systems). The data side of things may not be uniform either. So the compliance checking needs to be done over a very heterogeneous data landscape. In fact, the constraint file serves as a first filter that searches for contextual graph artifacts matching the constraint shape that resulted from the interpretation of law.

Currently, there is a need for better standardised mappings between legal norms and technical artifacts that potentially lead to ad-hoc, domain specific implementations. In the future, a law establishing a new compliance requirement may choose to create that SHACL constraint file and the subsequent action to accomplish as an annex to the actual law. In this case, the legislator does the translation himself. Consequently, the constraint file will participate in the normative and authoritative value of the legislator. As a consequence, matching the constraint will then mean the official recognition of compliance. A system of automatic compliance testing and confirmation would appear.

Looking at the concrete case of using ODRL to represent policies for aiding with regulatory compliance, it has been concluded that ODRL is not fit for purpose to represent legal concepts, such as purposes or legal grounds under which data can be accessed or used, as it does not contain such concepts in its vocabulary [9]. As such, one can make use of its profile mechanism to extend the ODRL vocabulary with concepts for representing contextual information relevant for legal compliance. Relevant work [5] in this area has been explored in the context of the SPECIAL project, which used ODRL constructs and legal concepts from vocabularies established in the context of this project to support regulatory compliance checking of business policies.[7] The SPECIAL project also launched the work of the W3C's Data Privacy Vocabularies and Controls Community Group (DPVCG).[8] The mission of this group is to develop specifications for representing machine-readable metadata about the use

[7] SPECIAL project homepage, `https://specialprivacy.ercim.eu`, retrieved 3 July 2025.
[8] Data Privacy Vocabularies and Controls Community Group homepage, `https://www.w3.org/community/dpvcg/`, retrieved 3 July 2025.

and processing of personal and non-personal data, as well as about technologies that use such data, in a jurisdiction-agnostic manner, and also create extensions to these specifications for concrete regulations, such as the European General Data Protection Regulation (GDPR) or the AI Act. The core specification, the Data Privacy Vocabulary (DPV) [25, 7] includes taxonomies to represent information about entities and legal roles, purposes and processing operation concepts, data and personal data, rights, risks, contextual information about processing operations, such as storage conditions or the scale of processing, technical and organisation measures, legal bases, and location and jurisdiction terms. Hence, by using both DPV and ODRL, i.e., making use of ODRL's profile mechanism, one can model policies using ODRL's model while using DPV's terms to refer to legal concepts [8, 24]. Furthermore, the previously cited publications are currently being used, in a joint effort by the ODRL and DPV communities, as the basis to create an official DPV-ODRL profile[9], and a guide document for using this profile[10]. As such, DPV is a promising approach to tackle regulatory compliance, when used with a policy language such as ODRL. It is still continuously being maintained and updated with new requirements coming from newly-enforceable laws, such as the European Health Data Spaces Regulation or the AI Act.

5.4.2.6 Governing agents on the Web using policies

In the context of autonomous agents on the Web, expressing policies formally is crucial, yet still a key challenge [18]. Web-based systems may be open to new, previously unidentified agents that must be guided by way of systematic formal knowledge upon which they can carry out logical deductions. On the Web, agents can autonomously discover more information about the resources they deal with thanks to the hypermedia dimension of REST-based infrastructure. Following a link, a Web agent can navigate from an object description to norms associated with it, to policies, and more [4].

If agents are implemented such that they can automatically follow policy descriptions, then they can optimise the utilisation of the Web resources, including–potentially–getting the assistance of other agents that have observable presence on the same Web platforms [28]. On the contrary, if the agents do not obey the rules imposed by policies, they may be driven away by either the Web platform that hosts the resources, or by other autonomous Web agents that have norm-enforcing role.

In such scenarios where artificial agents must adhere to policies as strictly as possible, and connect policies to resource descriptions as well as possibility of interactions, it is convenient to rely on knowledge graphs as the underlying data model and technology. As argued in [2], knowledge graphs are key enablers of autonomy in relation to all aspects of autonomy, although challenges relating to governing agents remain open.

Because of the open challenges in designing Web-based MAS, academic researchers and enterprise practitioners are crossing their views and insights over a W3C community group exploring Web Agent technologies, including recent advances in LLM-based Agentic AI.[11] The group identified a strong interest by academia and corporation alike to devise standardised interaction protocols, where norms and policies represent a key dimension for supporting the governance of agents on the Web [3, Section 10].

[9] Mapping from DPV to ODRL (Draft Community Group Report), `https://w3id.org/dpv/mappings/odrl`, retrieved 4 July 2025.
[10] Guide for using DPV with ODRL (Draft Community Group Report), `https://w3id.org/dpv/guides/dpv-odrl`, retrieved 4 July 2025.
[11] W3C Autonomous agents on the Web community group, `https://www.w3.org/community/webagents/`

5.4.3 Main Discussions

The ODRL Evaluator is designed to produce conclusions by performing logical reasoning over ODRL policies, an evaluation request, and a state of the world. Various approaches have been explored to achieve this, including mappings to finite state machine systems, formal logic systems, Answer Set Programming (ASP), and logic programming languages such as Prolog.

In this seminar, the potential use of SHACL has been explored. While SHACL has demonstrated effectiveness in validating the syntactic correctness of ODRL policies, it also shows promise as a possible reasoning engine underlying the ODRL Evaluator.

5.4.3.1 Expressing ODRL Constraints with SHACL

ODRL constraints, such as a rule limiting usage to ten hours, are typically expressed using the ODRL vocabulary, for example by defining a constraint with odrl:leftOperand, odrl:operator, and odrl:rightOperand. The same restriction can also be validated operationally with SHACL: a sh:PropertyShape can be defined on the property that records actual usage time, here ex:totalUsageTime, with sh:maxInclusive "PT10H"8sd:duration. This SHACL shape ensures that any recorded usage value exceeding ten hours will be flagged as invalid, making the policy both human-readable in ODRL and machine-checkable through SHACL.

```
[ a odrl:Constraint ;
  odrl:leftOperand odrl:meteredTime ;
  odrl:operator odrl:lteq ;
  odrl:rightOperand "PT10H"^^xsd:duration
]
```

Figure 6 ODRL Constraint.

```
[ a odrl:Constraint ;
  odrl:leftOperand odrl:meteredTime ;
  odrl:operator odrl:lteq ;
  odrl:rightOperand "PT10H"^^xsd:duration
]
```

Figure 7 SHACL Shape.

Figure 8 ODRL Constraint → SHACL Shape.

An important direction for future work is the definition of a general mechanism for mapping ODRL constraints to SHACL constraints, taking into account the variety of left operands defined in the ODRL Vocabulary. A key difficulty lies in bridging the abstract semantics of ODRL's left operands (e.g., `odrl:meteredTime`) with the concrete properties used in specific datasets or implementations (e.g., `ex:totalUsageTime`).

5.4.3.2 The challenge of operationalising compliance of AI systems

Automated regulatory compliance, particularly in the context of the European Union's AI Act, faces challenges in mapping high-level legal requirements to concrete technical artifacts [11]. The key difficulty often lies in the insufficient or fragmented information available across the AI lifecycle, which complicates interpretation and verification of transparency and traceability of AI systems.

Compliance frameworks for AI, such as the framework envisioned in the projects CERTAIN[12], HARNESS[13] or GLACIATION[14] aim to address this by formalising metadata and

[12] CERTAIN project homepage, https://certain-project.eu/, retrieved 4 July 2025.
[13] HARNESS project homepage, https://harness-network.eu/, retrieved 4 July 2025.
[14] GLACIATION project homepage, https://glaciation-project.eu, retrieved 4 July 2025.

aligning it with regulatory requirements [12]. However, integration of these representations require clear machine-processable formalisation of policies, which is a non-trivial task due to semantic ambiguity and evolving standards.

A promising approach to mitigate these issues is the current developments of SHACL: the use of SHACL as a validation mechanism to enforce structured constraints on policy-relevant metadata, enabling systematic compliance checks against legal requirements. For instance, the EU AI Act's Annex IV requires providers of high-risk AI systems to maintain technical documentation, such as performance metrics of an AI system. SHACL can be used to validate completeness and structural requirements – given that the underlying information is sufficiently formalised.

5.4.4 Proposed approaches

5.4.4.1 RDF 1.2 Triple Terms

The integration of RDF 1.2 capabilities into ODRL policy evaluation opens new possibilities for creating more robust and transparent policy enforcement mechanisms. Beyond the provenance and duty fulfilment scenarios outlined above, RDF 1.2 *annotations* can address several critical challenges in policy evaluation and compliance monitoring.

Policy State Tracking

A fundamental challenge in ODRL policy evaluation is maintaining state information across multiple policy interactions. Traditional approaches often rely on external state management systems, creating potential inconsistencies between the policy representation and its execution state. RDF 1.2 triple terms and reifiers together with Turtle 1.2's suggested Annotation Syntax [15] enable embedding state information directly within the policy graph itself.

This approach ensures that policy state remains synchronised with the policy definition, enabling more reliable policy evaluation and reducing the complexity of external state management systems.

Explainable Policy Decisions

Policy evaluation often involves complex reasoning processes that can be difficult to trace and explain. RDF 1.2 annotations provide a mechanism for embedding explanation trails directly into policy evaluation results, supporting both accountability and debugging requirements.

Such annotated evaluation results provide transparency into the decision-making process, enabling stakeholders to understand why specific policy decisions were made and facilitating trust in automated policy enforcement systems.

Temporal Policy Evolution

Policies often evolve over time, and maintaining a clear audit trail of policy changes is crucial for compliance and governance. RDF 1.2 annotations enable the embedding of version control information directly within policy definitions, creating self-contained policy evolution histories.

[15] Turtle 1.2 Annotation Syntax `https://www.w3.org/TR/rdf12-turtle/#annotation-syntax`

```
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <http://example.com/policy/> .

:policy003 a odrl:Agreement ;
 odrl:permission [
  odrl:target :document456 ;
  odrl:action odrl:read ;
  odrl:constraint [
   odrl:leftOperand odrl:count ;
   odrl:operator odrl:lteq ;
   odrl:rightOperand "5"^^xsd:integer
   # --- State Annotation ---
   {| :currentCount "2"^^xsd:integer ;
     :lastAccessed "2025-07-02T14:30:00Z"^^xsd:dateTime ;
     :accessedBy :user456
   |}
  ]
 ] .
```

Figure 9 Policy with embedded state tracking.

```
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <http://example.com/policy/> .

# Policy evaluation result
:evaluation001 a :PolicyEvaluation ;
 :evaluates :policy003 ;
 :result :denied
 {| :reason "Usage count limit exceeded" ;
   :evaluationTime "2025-07-03T10:15:00Z"^^xsd:dateTime ;
   :evaluatedConstraints (:constraint001 :constraint002) ;
   :failedConstraint :constraint001
 |} .
```

Figure 10 Policy evaluation with explanation annotations.

This approach enables policy systems to maintain comprehensive change histories without requiring external versioning systems, supporting both compliance requirements and operational transparency.

5.4.4.2 Validating Triple Terms with SHACL 1.2

A reified statement is an RDF triple (subject-predicate-object) turned into a resource that can itself be the subject of other statements. SHACL 1.2 supports reified statements by providing enhanced validation capabilities for metadata attached to statements, including support for RDF-star syntax and traditional RDF reification patterns. This allows SHACL shapes to validate not only the original triple but also the properties describing when, how, or by whom the statement was made.

```
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <http://example.com/policy/> .

:policy004 a odrl:Agreement ;
 odrl:permission [
  odrl:target :sensitiveData ;
  odrl:action odrl:use ;
  odrl:constraint [
   odrl:leftOperand odrl:purpose ;
   odrl:operator odrl:isA ;
   odrl:rightOperand :researchPurpose
  ]
 ]
 {| :version "2.1" ;
   :previousVersion :policy004_v2.0 ;
   :modifiedBy :admin_bob ;
   :modificationDate "2025-06-15T16:45:00Z"^^xsd:dateTime ;
   :changeReason "Updated purpose constraints per new research guidelines"
 |} .
```

Figure 11 Policy with version history annotations.

For example, a SHACL shape can require that every occurrence of an odrl:permission triple is accompanied by provenance metadata describing its version and possible link to a previous version. In the example below, the shape ex:ProvenanceShape specifies that reified permission statements must include at most one :version (typed as an xsd:date) and at most one :previousVersion (an IRI pointing to an earlier policy). The property shape ex:PermissionShape then declares that any odrl:permission triple must be reified and must conform to that provenance shape. In this way, SHACL 1.2 ensures that policies are not only structurally correct but also carry the necessary temporal and versioning information for accountability.

5.4.5 Conclusion

The group intends to dig deeper into the topic of ODRL semantics and using SHACL constraints. The results are such that a scientific article seems at reach. On the basis of that article, future work on ODRL will be suggested. This may lead to the creation of a new ODRL Working Group in W3C that would create a simplified profile of ODRL 2.2, and also do some maintenance work (i.e., correcting some issues in ODRL 1.1's W3C Note) and allow for the use of SHACL in the constraint element of ODRL. It could also further explore how to create Policy annotations of data using RDF 1.2. To kick off this initiative, a Workshop needs to be organised.

References

1 Piero Bonatti, Nicoletta Fornara, and Andreas Harth. Towards a Formal Semantics of the Open Digital Rights Language (ODRL 2.2). In Marta Sabou, Andreas Harth, Pasquale Lisena, Edward Curry, Bohui Zhang, Reham Alharbi, Yuan He, Georg Rehm, Sonja Schimmler, Stefan Dietze, Natalia Manola, Andrea Cimmino, Nicoletta Fornara, Víctor Rodríguez-Doncel, John Domingue, Achim Rettinger, Damian Trilling, Marko Grobelnik, Claudia d'Amato, Valeria Fionda, Ilaria Tiddi, and Gabriele Tolomei, editors, *ESWC 2025*

```
ex:ProvenanceShape
  a sh:NodeShape ;
  sh:property [
    sh:path :version ;
    sh:datatype xsd:date ;
    sh:maxCount 1 ;
  ] ;
  sh:property [
    sh:path :previousVersion ;
    sh:nodeKind sh:IRI ;
    sh:maxCount 1 ;
  ] .

ex:PermissionShape
  a sh:PropertyShape ;
  sh:path odrl:permission ;
  sh:reifierShape ex:ProvenanceShape ;
  sh:reificationRequired true .
```

Figure 12 SHACL 1.2 proposes constraint components for validating reifiers.

Workshops and Tutorials Joint Proceedings, volume 3977 of *CEUR Workshop Proceedings*. Sun SITE Central Europe (CEUR), June 2025. URL: http://ceur-ws.org/Vol-3977.

2 Jean-Paul Calbimonte, Andrei Ciortea, Timotheus Kampik, Simon Mayer, Terry R. Payne, Valentina Tamma, and Antoine Zimmermann. Autonomy in the Age of Knowledge Graphs: Vision and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):13:1–13:22, 2023. doi:10.4230/TGDK.1.1.13.

3 Andrei Ciortea. WebAgents Community Group Report on Interoperability for Agents on the Web. W3C Draft Community Group Report, World Wide Web Consortium, August 29 2025. URL: https://w3c-cg.github.io/webagents/TaskForces/Interoperability/Reports/report-interoperability.html.

4 Andrei Ciortea, Simon Mayer, Fabien Gandon, Olivier Boissier, Alessandro Ricci, and Antoine Zimmermann. A Decade in Hindsight: The Missing Bridge Between Multi-Agent Systems and the World Wide Web. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS'19, Montreal, QC, Canada, May 13-17, 2019*, pages 1659–1663. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

5 Marina De Vos, Sabrina Kirrane, Julian Padget, and Ken Satoh. ODRL Policy Modelling and Compliance Checking. In Paul Fodor, Marco Montali, Diego Calvanese, and Dumitru Roman, editors, *Rules and Reasoning – Third International Joint Conference, RuleML+RR 2019, Bolzano, Italy, September 16-19, 2019, Proceedings*, volume 11784 of *Lecture Notes in Computer Science*, pages 36–51. Springer, September 2019. doi:10.1007/978-3-030-31095-0.

6 Beatriz Esteves, Andrés Chomczyk Penedo, Blessing Mutiro, Haleh Asgarinia, and Dave Lewis. Privacy Paradigm ODRL Profile. Project report, PROTECT Consortium, April 11 2022. URL: https://w3id.org/ppop.

7 Beatriz Esteves, Delaram Golpayegani, Georg P. Krog, Harshvardhan J. Pandit, Julian Flake, and Paul Ryan. Digital Privacy Vocabulary (DPV), version 2.1. Final Community Group Report, World Wide Web Consortium, March 16 2025. URL: https://w3c.github.io/dpv/2.1/dpv/.

8 Beatriz Esteves, Harshvardhan J. Pandit, and Víctor Rodríguez-Doncel. ODRL Profile for Expressing Consent through Granular Access Control Policies in Solid. In *IEEE European Symposium on Security and Privacy Workshops, EuroS&P 2021, Vienna, Austria, September 6-10, 2021*, pages 298–306, 2021. doi:10.1109/EuroSPW54576.2021.00038.

9 Beatriz Esteves and Víctor Rodríguez-Doncel. Analysis of ontologies and policy languages to represent information flows in GDPR. *Semantic Web Journal*, 15(3):709–743, 2024. doi:10.3233/SW-223009.

10 Nicoletta Fornara, Victor Rodríguez-Doncel, Beatriz Esteves, Simon Steyskal, Benedict Whittam Smith, and Yassir Sellami. ODRL Formal Semantics. Draft Community Group Report, World Wide Web Consortium, July 02 2025. URL: https://w3c.github.io/odrl/formal-semantics/.

11 Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. To Be High-Risk, or Not To Be – Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 905–915, New York, NY, USA, June 2023. Association for Computing Machinery. doi:10.1145/3593013.3594050.

12 Delaram Golpayegania, Harshvardhan J. Panditb, Declan O'Sullivana, and Dave Lewisa. Semantic Frameworks to Support Implementation of the EU AI Act. *OSF Preprints*, 2025. Submitted to Computer Law & Security Review. URL: https://doi.org/10.31219/osf.io/43rq2_v1.

13 Arghavan Hosseinzadeh, Robin Brandstädter, and Jessica Chwalek. ODRL Profile for Data Sovereignty. Odrl profile, Fraunhofer IESE, September 9 2024. URL: https://w3id.org/ods/.

14 Renato Ianella. Open Digital Rights Language (ODRL) Version 1.1. W3C Note, World Wide Web Consortium, September 19 2002. URL: http://www.w3.org/TR/2002/NOTE-odrl-20020919/.

15 Renato Ianella and Serena Villata. ODRL Information Model 2.2. W3C Recommendation, World Wide Web Consortium, February 15 2018. URL: http://www.w3.org/TR/2018/REC-odrl-model-20180215/.

16 Renato Iannella, Michael Steidl, Stuart Myles, and Víctor Rodríguez-Doncel. ODRL Vocabulary & Expression 2.2. W3C Recommendation, World Wide Web Consortium, February 15 2018. URL: https://www.w3.org/TR/2018/REC-odrl-vocab-20180215/.

17 IPTC Rights Expression Working Group. IPTC RightsML Standard 2.0. Odrl profile, International Press Telecommunications Council, August 6 2018. URL: https://www.iptc.org/std/RightsML/2.0/RightsML_2.0-specification.html.

18 Timotheus Kampik, Adnane Mansour, Olivier Boissier, Sabrina Kirrane, Julian A. 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 Technologies*, 22(4):104:1–104:31, 2022. doi:10.1145/3507910.

19 Holger Knublauch, Thomas Bergwinkl, Yousouf Taghzouti, and Simon Werner. SHACL 1.2 Core. W3C First Public Working Draft, World Wide Web Consortium, March 18 2025. URL: https://www.w3.org/TR/2025/WD-shacl12-core-20250318/.

20 Holger Knublauch and Dimitris Kontokostas. Shapes Constraint Language (SHACL). W3C Recommendation, World Wide Web Consortium, July 20 2017. URL: https://www.w3.org/TR/2017/REC-shacl-20170720/.

21 Penny Labropoulo and Victor Rodríguez-Doncel. ODRL Profile for Policies of Language Resources and Technologies. W3C Community Group Draft Report, World Wide Web Consortium, June 2 2025. URL: https://rdflicense.linkeddata.es/profile.html#.

22 Mo McRoberts and Victor Rodríguez Doncel. Open Digital Rights Language (ODRL) Ontology. Community Group draft, World Wide Web Consortium, May 12 2014. URL: https://www.w3.org/ns/odrl/2/ODRL20.

23 Open Mobile Alliance. DRM Specification, Approved Version 2.0.2. OMA Technical Specification, Open Mobile Alliance, July 23 2008. URL: https://www.openmobilealliance.org/release/DRM/V2_0_2-20080723-A/OMA-TS-DRM_DRM-V2_0_2-20080723-A.pdf.

24 Harshvardhan J. Pandit and Beatriz Esteves. Enhancing Data Use Ontology (DUO) for Health-Data Sharing by Extending it with ODRL and DPV. *Semantic Web Journal*, 15(4):1473–1498, 2024. doi:10.3233/SW-243583.

25 Harshvardhan J. Pandit, Beatriz Esteves, Georg P. Krog, Paul Ryan, Delaram Golpayegani, and Julian Flake. Data Privacy Vocabulary (DPV) – Version 2.0. In Gianluca Demartini, Katja Hose, Maribel Acosta, Matteo Palmonari, Gong Cheng, Hala Skaf-Molli, Nicolas Ferranti, Daniel Hernández, and Aidan Hogan, editors, *The Semantic Web – ISWC 2024 – 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part III*, volume 15233, pages 171–193, Cham, October 2024. Springer. doi:10.1007/978-3-031-77847-6_10.

26 Siem Velmaat. Automated machine-readable data access agreements by applying ODRL to a FAIR Data Train. Master's thesis, University of Twente, September 2024. URL: https://essay.utwente.nl/103662/1/Veltmaat_MA_EEMCS.pdf.

27 Benedict Whittam Smith and Mark Bird. Market Data Profile for ODRL 1.0. W3C Community Group Draft Report, World Wide Web Consortium, December 1 2021. URL: https://www.w3.org/2021/md-odrl-profile/v1/.

28 Antoine Zimmermann, Andrei Ciortea, Catherine Faron, Eoin O'Neill, and María Poveda-Villalón. Pody: A Solid-Based Approach to Embody Agents in Web-Based Multi-Agent-Systems. In Andrei Ciortea, Mehdi Dastani, and Jieting Luo, editors, *Engineering Multi-Agent Systems – 11th International Workshop, EMAS 2023, London, UK, May 29-30, 2023, Revised Selected Papers*, volume 14378 of *Lecture Notes in Computer Science*, pages 220–229. Springer, 2023. doi:10.1007/978-3-031-48539-8_15.

Participants

- Nirav Ajmeri
University of Bristol, GB
- Matthew Arrott
Coactive Computing, US
- Matteo Baldoni
University of Turin, IT
- Cristina Baroglio
University of Turin, IT
- Victor Charpenay
Mines Saint-Étienne, FR
- Amit K. Chopra
Lancaster University, GB
- Mehdi Dastani
Utrecht University, NL
- Marina De Vos
University of Bath, GB
- Davide Dell'Anna
Utrecht University, NL
- Frank Dignum
University of Umeå, SE
- Beatriz Esteves
Ghent University, BE
- Nicoletta Fornara
USI – Lugano, CH
- Joris Hulstijn
Utrecht University, NL
- Timotheus Kampik
SAP Berlin, DE &
Umeå University, SE
- Nadin Kokciyan
University of Edinburgh, GB
- Beishui Liao
Zhejiang University, CN
- Réka Markovich
University of Luxembourg, LU
- Pradeep Murukannaiah
TU Delft, NL
- Vivek Nallur
University College Dublin, IE
- Luis Gustavo Nardin
IMT Mines Saint-Étienne, FR
- Sebastian Neumaier
FH – St. Pölten, AT
- Julian Padget
University of Bath, GB
- Victor Rodriguez Doncel
Polytechnic University of
Madrid, ES
- Susana Rodríguez Verdugo
Kunveno Digital – Madrid, ES
- Ken Satoh
Research Organization of
Information and Systems –
Tokyo, JP
- Jaime Sichman
University São Paulo, BR
- Judith Simon
Universität Hamburg, DE
- Munindar P. Singh
North Carolina State University –
Raleigh, US
- Simon Steyskal
Siemens AG – Wien, AT
- Sz-Ting Tzeng
University of Umeå, SE
- Leon van der Torre
University of Luxembourg, LU
- Harko Verhagen
Stockholm University, SE
- Rigo Wenning
W3C / ERCIM, FR
- Jessica Woodgate
University of Bristol, GB
- Pinar Yolum
Utrecht University, NL
- Antoine Zimmermann
Ecole des Mines –
St. Étienne, FR