Towards a Shared Understanding of Democratic Values for AI Governance in Multi-Agent Systems

Sinead Sadler[1,2,*,†], Anelia Kurteva[1,†]

[1]School of Computer Science, University of Birmingham, United Kingdom

Abstract

The pursuit for trustworthy Artificial Intelligence is a global challenge and founded upon core democratic values of transparency, accountability, safety, human oversight, personal privacy, diversity, and wellbeing. With the prevalence of multi-agentic systems, the need for a clear demonstration of these principles within autonomous decision making is paramount for enhancing trustworthiness. This paper seeks to raise awareness of semantic technologies as a technological solution to this challenge of providing a shared understanding of democratic values in multi-agent environments. In doing so, it proposes a human-centred, multi-agent framework that utilises semantic technologies to support value alignment and the understanding of democratic values for AI governance.

Keywords

AI Governance, Democratic Values, Multiagent Systems, Ontology, Knowledge Graph

1. Introduction

Artificial Intelligence (AI), in particular Generative and Agentic AI, is being adopted at a faster rate than any previous technological advancement [1], which raises scrutiny over its development and adoption safeguards. Frontier AI labs, that develop state-of-the-art models and conduct AI research, are arguably not developed with sufficient democratic legitimacy and regard for public interests [2] to govern AI in a democratic manner. The reasoning for this being that they are privately owned companies, often lacking in transparency and holding a monopoly within the technology market [2]. Despite this, they do allow for the participation and access of AI to the public through their open source AI models [3]. While this supports access and participation, it doesn't necessarily incorporate democratic value alignment, which is an essential component of trustworthy AI. AI governance seeks to make AI trustworthy and it does this through the protection and upholding of human values, fairness, trust, transparency, safety, security, inclusive growth and wellbeing [4]. Alignment and governance of these values within AI is a global challenge, as demonstrated by the establishment of over 1000 OECD policy initiatives in 70 jurisdictions to combat these issues [5]. The fragmented yet broad application of AI governance illustrates the need for a shared understanding of democratic values particularly with the prevalence of multi-agent systems.

Multi-agent systems by definition are a collection of LLMs that interact together to solve tasks on behalf of the user in a multi step way [6]. The agents within the system tend to be autonomous by design [7], and possess 'heterogeneity, autonomy, capacity to process' and act on behalf of the user [8]. Increasingly, they are assuming roles with human decision making, so they need to exhibit the same levels of trust, accountability and permissions [9] that underpin human processes and judgement. This compounded with known risks of goal misalignment, behavioural drift, tool misuse and emergent coordination failures [9], give a need for a governance that enshrines democratic values that enhance trustworthy AI.

NeXt-generation Data Governance workshop, SEMANTiCS 2026

*Corresponding author.

†These authors contributed equally.

sas693@student.bham.ac.uk (S. Sadler); a.kurteva@bham.ac.uk (A. Kurteva)

https://aneliakurteva.com (A. Kurteva)

0009-0002-3528-9492 (S. Sadler); 0000-0003-4512-5969 (A. Kurteva)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Democratic values themselves uphold fundamental rights and freedoms whilst providing legal remedies for infringement upon them [10]. AI governance embeds these through the principles of transparency, accountability, safety, human oversight, personal privacy, diversity, and wellbeing [11][3]. Failure to adhere to these can cause jeopardise the rights of an individual and their protections from harm.

One of the most prolific examples of an encroachment of AI on democratic values was the Cambridge Analytica case [12]. Their misuse of algorithms and violations on personal data was used to influence voting behaviour [12]. This was not just an invasion of privacy, but a clear interference in democratic processes and lack of transparency surrounding data and algorithms [13]. Risk of an individuals safety, wellbeing, and privacy by AI can result in severe harms, such as death, as illustrated by the case of Raine vs Open AI [14]. Summary judgements noted that model designs could be coercive and impede autonomy by purposeful psychological dependency [14]. Consequentially, this can undermine consent [15] as per Article 7 of the GDPR [16]. Based on this need for a shared understanding and technical implementation of democratic values within multi-agent systems, this paper will propose a collaborative, human centred framework that upholds democratic values. In doing so, it will provide an overview of semantic technologies and their potential to achieve this [17][18].

2. Technologies supporting Democratic Value Implementation

Semantic technologies utilised in multi-agent systems can enable robust mechanisms to uphold privacy, compliance, autonomy, data sovereignty, transparency, and accountability.

2.1. Ontologies: Building Shared Understanding

Vocabularies and ontologies function as a vital bridge between abstract ethical principles and actionable AI governance. By formalising semantic concepts and complex object relationships, these technologies cultivate a shared understanding of democratic values leading to transparency, privacy, and accountability. Crucially, this approach enables the translation of normative goals into functional safeguards. For example, cross-domain ontological solutions such as those employing descriptive-based logical approaches [19] enable encoding of legal instruments. As a result, this upholds the rule of law and allows for its protections to be embedded into system architectures. Furthermore, this integration supports legal certainty by ensuring that agentic AI behaviours remain aligned with democratic standards and legislation. From a governance viewpoint, this enhances transparency, trust and protection of the individuals rights and freedoms enshrined within legal instruments.

Widely adopted standards, such as the Data Privacy Vocabulary (DPV) [20], play a critical role in elevating privacy from a high-level democratic ideal to a technical requirement. By providing a standardised, machine-readable lexicon for data handling, the DPV creates a shared, unambiguous language for privacy governance. This codification ensures that privacy protections are not merely abstract, policy goals but foundational and operational components within the infrastructure. As such, it supports systems to interpret and uphold privacy requirements consistently and reliably across multi-agentic environments.

Specialised vocabularies such as the AI risk ontology, AI Risk Ontology(AIRO) [21], and Vocabulary of AI Risks (VAIR) [22], specialise in the operability of risk management concepts as defined by the EU AI Act [23]. The integration of these semantic technologies helps ensure that AI systems remain compliant and increases understanding between different users, thus making it suitable for multi-agentic environments when using autonomous decision making. Moreover, these tools foster cross-contextual privacy interoperability, with a strong focus on GDPR compliance [20] and use of FAIR principles. The collaborative, interdisciplinary methodology involved in their development significantly enhances semantic depth and precision [20], thereby increasing design integrity and trust for the user. Although not exhaustive in their entirety, vocabularies such as DPV, provide a sufficiently broad, conceptual framework for a range of data handling scenarios while maintaining rigorous protections [20].

2.2. Knowledge Graphs: Making Data FAIR and Operational for AI

Context is pivotal not only for establishing values but also understanding the nuances of a culture or society in which they seek to operate [24]. For AI, a lack of context, means that people cannot be engaged in a 'meaningful way' [25]. Knowledge graphs provide a solution for this added complexity of context, particularly in multiagent environments, whereby agents are interacting with other agents in various contexts and tasks. They enable a more coherent and structured understanding of relationships between entities, can provide a form of verification and ground AI outputs. Consequentially, they reinforce concepts of transparency and trust on decision making though 'reliability, truth and ability' [19].

Moreover, knowledge graphs enable the implementation of FAIR data principles through metadata labelling [26], thus enhancing democratic values to be embedded within AI due to improved transparency, accountability and legal certainty. Legal knowledge graphs legitimise the rule of law and AI governance into the architecture it operates in. Examples of successful work in utilising legal knowledge graphs for compliance include the SmashHit Legal Knowledge graph powered by the smashHitCore ontology for GDPR consent verification [27]. Knowledge graphs have shown considerable promise in mitigating against hallucinations [28], helping ensure outputs do not deviate against values which could cause harm to a user's wellbeing or jeopardise trust. To further increase trust and transparency, knowledge graphs can assist in the interpretation of data and AI for human users via visualisations [29]. Most notably, this has been used to enhance informed consent, therefore promoting self sovereignty and autonomy for users, such as the use of visualisations for cookie comprehension [30]. It can ensure that consent goes beyond that of a superficial requirement and be truly participatory [30]. By default therefore, it supports democratic values and based on this reasoning could enhance trustworthy outputs in multi-agentic environments.

2.3. Open Digital Rights Language (ODRL): Defining Clear Access and Usage Controls for Greater Accountability and Safety Guardrails

ODRL is a machine readable policy language that expands on vocabularies and ontologies [23]. ODRL can express permissions, prohibitions, access controls and duties in different contexts [31]. It can implement them through algorithms [32], making it a suitable guardrail for agentic decision making and multi-agentic systems. The ability to explicitly state and enforce policies in a way that is compliant and customised to the policy owner's requirements, helps reinforce autonomy and self sovereignty. Additionally, an extra security mechanism is the ODRL compliance report model [33], which can ensure that autonomous decision making is properly evaluated and implemented.

2.4. Shapes Constraints Language (SHACL): Quality Control and Verification

SHACL shapes provide a data verification mechanism [34] by using through structural constraints and RDF nodes without needing dedicated infrastructure making it more FAIR through its accessibility and interoperability [34]. It's potential to check compliance through validating machine readable components of legislation against SHACL shapes, means that it can implicitly support the rule of law and regulation that upholds values such as privacy through the GDPR. Recent research has suggested there is potential for automated compliance checking in this way [35]. However, there exists limitations with 'inferring sub-class or super-class relations' to interpret the high-risk rule in the EU AI Act [23], so more research needs to be conducted around in this area.

2.5. Privacy Enhancing Technologies: Preserving Privacy by Design

Privacy-enhancing technologies (PETs) such as differential privacy [36], federated learning [37], and zero knowledge proofs [38] have established themselves as a foundational standard for robust data governance, enabling the extraction of high-utility information from sensitive datasets while mathematically guaranteeing the preservation of data privacy [18]. As discussed, in a recent OECD policy paper

on PETS and AI [39], PETS minimise the risk of privacy violations by reducing the need for sharing sensitive, personally identifiable data. Additionally, it notes the co-creation aspect of AI models by implementation of PETS, which allows for greater participation in technology [39].

By upholding the right to privacy, they support autonomy, sovereignty and freedom of the individual though the notion of consent. Trust is built upon informed consent [40], so technologies which support this within AI architectures, naturally help to promote AI trustworthiness as per the OECD objectives [5]. In multi-agent systems, the risk to privacy violations is higher, so PETs play a pivotal role in data flows to protect privacy [40]. When it comes to multi-agent systems, it is important to remember that decision making delegation does not necessarily mean relinquishing autonomy, if mechanisms for informed consent and transparency are incorporated [15].

PETs, such as federated learning, allow organisations to access and train a model in a shared way, but keep data locally within their respective locations [18]. This allows for users democratic right of privacy to be respected [18]. Knowledge sharing and participation can still happen without the compromising of sensitive data or the autonomy of the individual. Offering solutions such as this fosters trust and transparency with data subjects. In health domains, it can enhance the wellbeing of individuals and progression of technology by still being able to access insights from models [18].

Differential privacy is used in the context of deep learning [18]. It is useful in multi-agent environments that use LLM architectures because they can also protect privacy [36]. By adding noise to the data to avoid revealing information, they successfully protect user privacy [18], improve interoperability and support FAIR principles capabilities in data-sensitive domains [36].

Lastly, zero knowledge proofs are a type of a PET that allow statements to be validated, as a form of access control, from one party to another without revealing the secrets that make it true [18]. In the context of multiagent environments, where data needs to be communicated between agents within or outside the ecosystem, zero knowledge proofs provide an enhanced solution for privacy and wellbeing, because statement truths are formulated without personally identifiable information that could subject users to risk or harm if revealed.

2.6. Existing frameworks for democratic and trustworthy AI

Governance frameworks exist at a conceptual level to help understand the connection of democratic values and AI systems. For instance, the AI Democratisation Framework [3] holistically provides an understanding of how social, economic, and technological factors create opportunities and challenges for AI governance [3]. Others are value specific, such as the autonomy focused framework [15], which strives to understand organisational layers and invoke self reflection through questions [15]. The autonomy governance framework is a tool that provides practical measures for autonomy-centred AI [15]. Specific frameworks such as this, demonstrate how democratic value alignment is happening within the domain.

More technical frameworks such as, Singapore's Model AI Governance Framework for Agentic AI [41], focuses on risk, human accountability, technical controls and user responsibility as opposed to heavy reliance on the rule of law through normative regulation [41]. Internationally recognised, it presents a different world-view on trust within a 'communitarian' rather than 'rights-based, individualistic orientation of Western ethical approaches', promoting freedom of expression and inclusivity of cultural perspectives [42]. While democratic objectives of trust, accountability and risk mitigation are foundational, they may present differently depending on the context in which they operate. This provokes a deeper question of *how do multi-agents operate cross-jurisdictionally across competing values and goals?* Technically, this specific framework, focuses on the multi-agent architecture design as a form of control and guardrails, which encompass the development lifecycle [41][42].

The GAIA-X [1] architecture model assists in providing a shared understanding for AI governance between different stakeholders. In doing so, it allows for participation and autonomy for stakeholders to influence governance in a way that is democratic, transparent and trustworthy. It incorporates

[1] https://gaia-x.eu/, accessed 1 July 2026.

FAIR principles through verifiable and composable, self descriptions [43]. As per the framework's objectives, it explicitly aims for participation, trust, transparency with identified accountability and liability [43]. Therefore, not only does it enshrine democratic values but it also provides remedies and accountability as democratic mechanisms. Furthermore, it utilises semantic technologies for a 'self-description taxonomy', vocabulary and ontology for enhanced shared semantic understanding, value alignment and interoperability [43] providing a clear example of semantic technologies supporting AI trustworthiness.

3. Proposed framework for shared democratic value alignment

The motivation for this framework, as shown in Figure 1, is to provide a human centric, collaborative multi-agentic system grounded by an ontology, representing democratic principles and values, that leverages and aligns with the FAIR data principles. Consequentially, this would support democratic value alignment in multi-agentic workflows across various contexts. Semantic technologies are leveraged in a way that provide grounding and constraints to ensure compliance and value alignment. Despite these guardrails, they provide sufficient flexibility to incorporate the dynamic nature of human agents and their ever-changing personal preferences, needs and contexts as well as their agentic counterparts operating across multiagent systems. FAIR principles enable interoperability and consistent data sharing practices for multi-agent communication without compromising individual rights such as privacy. All of which enhance the notion of trustworthiness within the multi-agent workflow. Encapsulated within the design are various subagents that help provide collaboration, value alignment and facilitate checks and balances. Similar to the Singapore's Model AI Governance Framework for Agentic AI [41] and Gaia-X[43], the multi-agentic workflow becomes a participatory and transparent process.

Human Agent Component. The human agent acts as a central point of contact to share its values and preferences, thus enhancing self sovereignty and autonomy. This bilateral communication flows to and from the orchestrator agent, who delegates tasks among its dedicated subagent models, allowing for dynamic interaction. The orchestrator can delegate tasks between the three agentic components embedded in the layer below; the knowledge agent, security and compliance agent and the gateway agent.

Knowledge Agent. This agent acts as a reserve of knowledge through the utilisation and implementation of semantic technologies in collaboration with other agents and the human agent. Building on standard vocabularies like DPV, creates a foundation for the ontology, as do frameworks of risk in the EU AI act [23]. However, a broader ontological work that encompasses political science, democratic theory and law, to fully understand the granular nuances of concepts, entities and relationships within the wider democratic context is proposed. The democratic ontology would be shared and accessible to all agents and humans to provided grounded and aligned outputs that comply with the jurisdiction in which it operates. Using this knowledge and the human input, it constructs a personal knowledge graph to inference and reason over, providing a technical solution that upholds liberty, autonomy and privacy of the individual. Implementations of personal knowledge graphs to enhance trust and autonomy have already been considered within the healthcare domain [19] and could be extended for value alignment. The knowledge graph would be used to represent a visualisation of data to the human user [40], facilitating human agentic collaboration, enhancing consent, verification, transparency and ultimately creating trust rooted in procedural integrity.

Compliance Agent. The compliance agent is embedded in the layer overseen by the orchestrator and checks actions carried out by other agents are compliant with ODRL policies. The ODRL itself is formulated by inference of the personal knowledge graph and the policy owner, human agent or organisation preferences. The contribution of this agent is not only policy grounded autonomy for the user but maintenance of privacy and a filtering before being sent through the gateway agent. Differential privacy can be used here as a governance mechanism when sending aggregated data to and from the gateway to mitigate against risks and malicious behaviour. Zero knowledge proofs would be used to prove compliance without revealing any sensitive data within the multi-agent workflows.

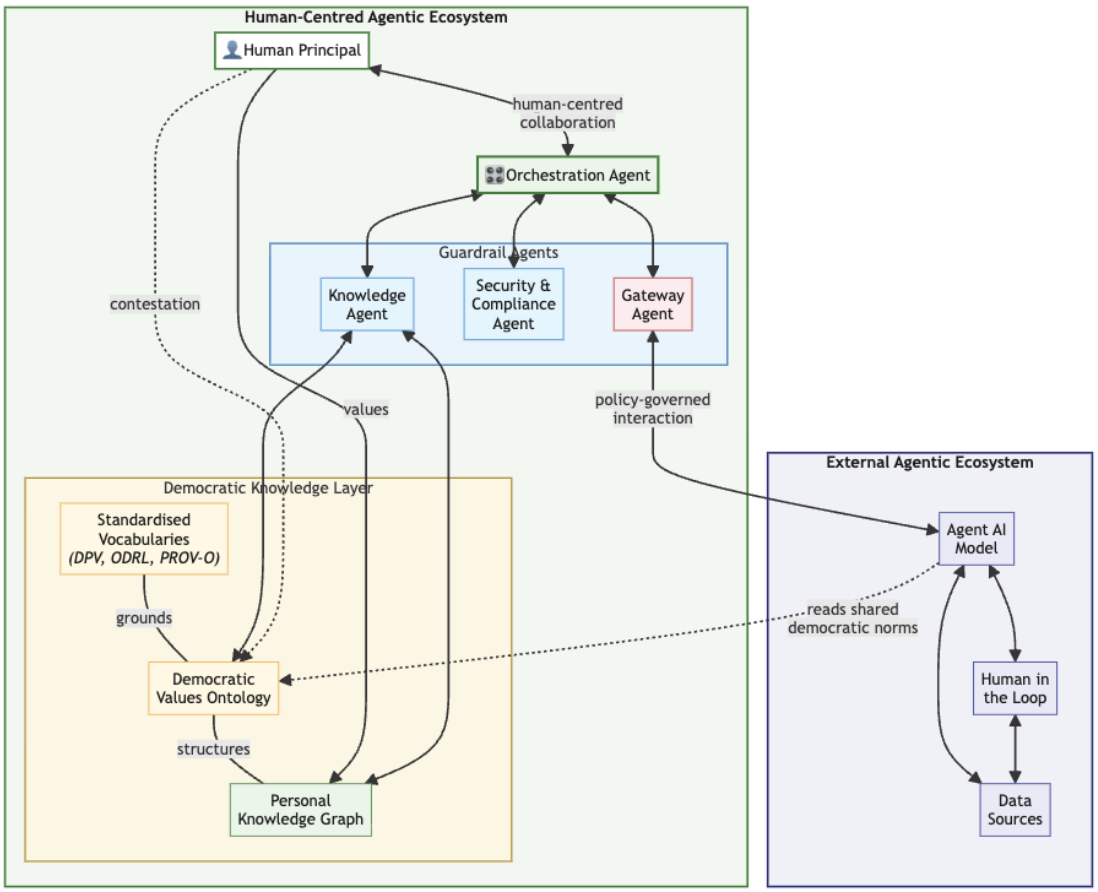

Figure 1: Framework for a multi-agentic, human-centred ecosystem that enables shared democratic alignment

Gateway Agent. Lastly, the gateway agent, acts a boundary between the internal and external agents that they are interacting with, providing a check and balance in any interaction before data enters or leaves the internal ecosystem. It can filter the incoming and outgoing flows of data, logs audits of what is shared and provide compliance checks on requests using the knowledge from the compliance and knowledge agents. It can also identify the jurisdictions it is operating in if the multi-agentic communication involves global interaction and cross reference against its internal agents, helping provide initial steps towards understanding the different contexts of values. However, the deeper semantic interpretation of values sits within the other agentic layers.

4. Conclusion

In conclusion, this paper was intended to invoke consideration and discussion of democratic values within multi-agentic environments and highlight the potential of semantic technologies to support further research in this area. The framework proposes a dynamic and collaborative model of governance, that focuses on the human as the centre point with semantic technologies acting as constraints to adhere to democratic value alignment. It extends previous semantic research to beyond the interpretation of AI governance values, into one that understands legal and political theory, whilst balancing the personal values and preferences of humans through their personal knowledge graph.

However, there are many points to consider such as, is this framework truly feasible at scale and how, when and why do we need to prioritise one value over another? As mentioned previously, AI

governance is a global issue and therefore needs to consider these challenges of implementation. For instance, democratic values often take priority over that of the individual preferences, but how does that dynamic establish itself against the individuals right to freedom of expression within the European Convention on Human Rights [44] or constitutions with a strong freedom of speech protection such as the First Ammendment [45] in the USA? Some research in this area has been done with norm conflict detection and resolution [46] but needs to be understood more closely within the current technological innovations of Generative AI and in a cross-jurisdictional setting. Moreover, are the decisions of priority traceable and explainable within the multi-agentic system? Another point to consider, is the creation of artefacts and LLM reasoning within the multi-agentic architecture, because there exists the dilemma of the black-box which limits transparency. This ultimately, leads to the question of how strong are the mechanisms of the framework and implementation of values to counteract this and provide sufficient provenance, accountability and trust?

Future research will explore these open ended questions and also propose the formal development of a democratic ontology that extends across the various disciplines beyond governance and computer science, and potentially invoke multi jurisdictional concepts for interoperability and fairness. Additionally, the development of the personal knowledge graph pipeline will be an important research direction to provide a way for humans to express themselves and their values autonomously. Inevitably, as an initial framework, this would need significant testing and modelling to assess and evaluate it's effectiveness in meeting the democratic values. Evaluation methods themselves would need to consider if they truly measure and evaluate the procedural impact of such a framework, such as value alignment. Research in this area with quantitative evaluation measures and human values exists [47], but not specifically with democratic values. Additionally, the evaluation methods proposed are in themselves not without limitations [47].

While the future strives to make the world ever more agentic, this initial framework instils the importance of the human as its own autonomous agent, for collaboration and preservation of self-sovereignty. Collaborative stakeholder engagement should no longer be a 'nice-to-have' but an important component in ensuring that AI not only embeds democratic values but encourages the democratisation of AI in its entirety.

References

[1] A. Bick, A. Blandin, D. J. Deming, The rapid adoption of generative AI, Working Paper 32966, National Bureau of Economic Research, 2024. URL: https://www.nber.org/papers/w32966. doi:10.3386/w32966.

[2] E. Seger, A. Ovadya, D. Siddarth, B. Garfinkel, A. Dafoe, Democratising ai: multiple meanings, goals, and methods, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 715–722. URL: https://doi.org/10.1145/3600211.3604693. doi:10.1145/3600211.3604693.

[3] C. J. Costa, M. Aparicio, S. Aparicio, J. T. Aparicio, The democratization of artificial intelligence: theoretical framework, Applied Sciences (Switzerland) 14 (2024). doi:10.3390/app14188236.

[4] High-Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy AI, Technical Report, European Commission, 2019.

[5] Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence, Technical Report, OECD, 2019. OECD-LEGAL-0449, amended 2024.

[6] N. M. Intelligence, Multi-agent AI systems need transparency, Nature Machine Intelligence 8 (2026) 1. URL: https://doi.org/10.1038/s42256-026-01183-2. doi:10.1038/s42256-026-01183-2, editorial.

[7] D. Filip, D. Lewis, An ontology for standardising trustworthy ai, in: Artificial Intelligence - Recent Trends and Developments, IntechOpen, 2021. URL: https://www.intechopen.com/chapters/76436. doi:10.5772/intechopen.97230.

[8] J. Rocha, I. Boavida-Portugal, E. Gomes, Introductory chapter: multi-agent systems, in: J. Rocha (Ed.), Multi-agent Systems, IntechOpen, 2017. doi:10.5772/intechopen.70241.
[9] W. E. Forum, AI agents in action: foundations for evaluation and governance, Technical Report, World Economic Forum, 2025.
[10] D. Leslie, C. Burr, M. Aitken, J. Cowls, M. Katell, M. Briggs, Artificial intelligence, human rights, democracy, and the rule of law: a primer, Technical Report, The Alan Turing Institute, 2021. Prepared for the Council of Europe's Ad Hoc Committee on Artificial Intelligence (CAHAI). arXiv:2104.04147.
[11] Organisation for Economic Co-operation and Development, Recommendation of the council on artificial intelligence, Technical Report, OECD, 2019. doi:10.1787/eedfee77-en.
[12] J. Isaak, M. J. Hanna, User data privacy: Facebook, Cambridge Analytica, and privacy protection, Computer 51 (2018) 56–64. doi:10.1109/MC.2018.3191268.
[13] M. Hu, Cambridge Analytica's black box, Big Data & Society 7 (2020). doi:10.1177/2053951720938091.
[14] M. Raine, M. Raine, Raine v. OpenAI, Inc., Superior Court of the State of California, County of San Francisco, Case No. CGC-25-628528, 2025. Filed Aug. 26, 2025.
[15] P. Subías-Beltrán, I. Unceta, I. de Lecuona, O. Pujol, Asking the right questions: a governance approach to uphold human autonomy in artificial intelligence, AI and Society (2025). doi:10.1007/s00146-025-02611-4.
[16] European Parliament and Council of the European Union, Regulation (EU) 2016/679 (General Data Protection Regulation), 2016.
[17] A. Meroño-Peñuela, E. Simperl, A. Kurteva, I. Reklos, KG.GOV: Knowledge graphs as the backbone of data governance in AI, Journal of Web Semantics 85 (2025). doi:10.1016/j.websem.2024.100847.
[18] J. M. Auñón, D. Hurtado-Ramírez, L. Porras-Díaz, B. Irigoyen-Peña, S. Rahmian, Y. Al-Khazraji, J. Soler-Garrido, A. Kotsev, Evaluation and utilisation of privacy enhancing technologies—A data spaces perspective, Data in Brief 55 (2024). doi:10.1016/j.dib.2024.110560.
[19] L.-D. Ibáñez, J. Domingue, S. Kirrane, O. Seneviratne, A. Third, M.-E. Vidal, Trust, accountability, and autonomy in knowledge graph-based ai for self-determination, Transactions on Graph Data and Knowledge 1 (2023).
[20] H. J. Pandit, D. O'Sullivan, D. Lewis, Creating a vocabulary for data privacy, in: H. Panetto, C. Debruyne, M. Hepp, D. Lewis, C. Ardagna, R. Meersman (Eds.), On the Move to Meaningful Internet Systems: OTM 2019 Conferences, volume 11877 of *Lecture Notes in Computer Science*, Springer, Cham, 2019, pp. 714–730. doi:10.1007/978-3-030-33246-4_44.
[21] D. Golpayegani, H. J. Pandit, D. Lewis, AIRO: an ontology for representing AI risks based on the proposed EU AI Act and ISO risk management standards, in: Towards a Knowledge-AwareAI, volume 55 of *Studies on the Semantic Web*, 2022, pp. 51–65. doi:10.3233/SSW220008.
[22] D. Golpayegani, H. J. Pandit, D. Lewis, Vocabulary of AI risks (VAIR), 2023. URL: https://w3id.org/vair. doi:10.5281/zenodo.10894914.
[23] D. Golpayegani, H. J. Pandit, D. O'Sullivan, D. Lewis, Semantic frameworks to support implementation of the EU AI Act, Computer Law & Security Review 61 (2026) 106315. URL: https://doi.org/10.1016/j.clsr.2026.106315. doi:10.1016/j.clsr.2026.106315.
[24] J. McKinlay, M. De Vos, J. Hoffmann, A. Theodorou, Understanding the process of human-ai value alignment, Journal of Artificial Intelligence Research 85 (2026). URL: http://dx.doi.org/10.1613/jair.1.18846. doi:10.1613/jair.1.18846.
[25] B. Zhang, A. Meroño-Peñuela, E. Simperl, Towards explainable automated knowledge engineering with human-in-the-loop, Semantic Web 16 (2025). doi:10.1177/22104968251382171.
[26] L. Vogt, P. Strömert, N. Matentzoglu, et al., Suggestions for extending the fair principles based on a linguistic perspective on semantic interoperability, Scientific Data 12 (2025) 688. URL: https://doi.org/10.1038/s41597-025-05011-x. doi:10.1038/s41597-025-05011-x.
[27] A. Kurteva, T. R. Chhetri, A. Tauqeer, R. Hilscher, A. Fensel, K. Nagorny, A. Correia, A. Zilverberg, S. Schestakov, T. Funke, E. Demidova, The smashHitCore ontology for GDPR-compliant sensor

data sharing in smart cities, Sensors 23 (2023). doi:10.3390/s23136188.
[28] Y. Chen, M. Chen, Y. Zhu, J. Pei, S. Chen, Y. Zhou, Y. Wang, Y. Zhou, H. Li, S. Zhang, Leverage knowledge graph and large language model for law article recommendation: A case study of chinese criminal law, J. Web Semant. 89 (2026) 100879. URL: https://doi.org/10.1016/j.websem.2026.100879. doi:10.1016/J.WEBSEM.2026.100879.
[29] T. R. Chhetri, A. Kurteva, R. J. Delong, R. Hilscher, K. Korte, A. Fensel, Data protection by design tool for automated GDPR compliance verification based on semantically modeled informed consent, Sensors 22 (2022). doi:10.3390/s22072763.
[30] G. Bushati, S. Rasmusen, A. Kurteva, A. Vats, P. Nako, A. Fensel, What is in your cookie box? Explaining ingredients of web cookies with knowledge graphs, Semantic Web 15 (2023) 1593–1609. doi:10.3233/sw-233435.
[31] A. Cimmino, N. Fornara, Improving odrl 2.2: current limitations and theoretical solutions, in: Joint Proceedings of the ESWC 2025 Workshops and Tutorials, volume 3977 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-3977/OPAL2025-6.pdf.
[32] P. A. Bonatti, N. Fornara, A. Harth, Towards a formal semantics of the Open Digital Rights Language (ODRL 2.2), in: Joint Proceedings of the ESWC 2025 Workshops and Tutorials co-located with 22nd Extended Semantic Web Conference (ESWC 2025), Portorož, Slovenia, June 1–2, 2025, volume 3977 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-3977/OPAL2025-4.pdf.
[33] W. Slabbinck, J. Rojas Meléndez, B. Esteves, P. Colpaert, R. Verborgh, Interoperable interpretation and evaluation of ODRL policies, in: The Semantic Web: 22nd European Semantic Web Conference (ESWC 2025), pages = 192–209,, 2025. doi:10.1007/978-3-031-94578-6_11.
[34] M. Šimkus, I. Varzinczak (Eds.), Reasoning Web. Declarative Artificial Intelligence, volume 13100 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2022. URL: https://link.springer.com/10.1007/978-3-030-95481-9. doi:10.1007/978-3-030-95481-9.
[35] J. Anim, L. Robaldo, A. Z. Wyner, A SHACL-Based approach for enhancing automated compliance checking with RDF Data, Information 15 (2024). doi:10.3390/info15120759.
[36] Y. I. Alzoubi, A. Mishra, Differential privacy and artificial intelligence: potentials, challenges, and future avenues, EURASIP Journal on Information Security 2025 (2025). doi:10.1186/s13635-025-00203-9.
[37] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y. Arcas, Communication-efficient learning of deep networks from decentralized data, in: A. Singh, J. Zhu (Eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.
[38] S. Goldwasser, S. Micali, C. Rackoff, The knowledge complexity of interactive proof-systems, in: Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing, 1985, pp. 291—-304. doi:10.1145/22145.22178.
[39] OECD, Sharing trustworthy AI models with privacy-enhancing technologies, Technical Report 38, OECD Publishing, 2025. URL: https://doi.org/10.1787/a266160b-en. doi:10.1787/a266160b-en.
[40] A. Kurteva, J. Domingue, Towards cultivating cecentralised data privacy, interoperability and trust with semantic PETs and visualisations, in: Proceedings of the 1st NeXt-generation Data Governance Workshop (NXDG 2024) co-located with 20th International Conference on Semantic Systems (SEMANTiCS 2024), volume 3891 of *CEUR Workshop Proceedings*, 2024. URL: https://ceur-ws.org/Vol-3891/paper1.pdf.
[41] Infocomm Media Development Authority, Ministry of Digital Development and Information, Model AI governance framework for agentic AI, Technical Report, Infocomm Media Development Authority (IMDA), Singapore, 2026. URL: https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf, accessed: July 8, 2026.
[42] J. G. Allen, J. Loo, J. Luna, Governing intelligence: Singapore's evolving ai governance framework, Cambridge Forum on AI: Law and Governance 1 (2025). doi:10.1017/cfl.2024.12.
[43] Gaia-X AISBL, Gaia-x architecture document - 22.04 release, Official Gaia-X document, 2022. URL:

https://gaia-x.eu/, accessed: 2026-07-08.
[44] Council of Europe, European Convention on Human Rights, 1950. URL: https://www.echr.coe.int/documents/d/echr/convention_ENG.
[45] United States Congress, The constitution of the united states: Amendment i (freedom of speech, press, religion, assembly, and petition), 1789. URL: https://www.archives.gov/founding-docs/bill-of-rights-transcript.
[46] W. W. Vasconcelos, M. J. Kollingbaum, T. J. Norman, Normative conflict resolution in multi-agent systems, Autonomous Agents and Multi-Agent Systems 19 (2009) 124–152. doi:10.1007/s10458-008-9070-9.
[47] M. Peterson, P. Gärdenfors, How to measure value alignment in AI, AI and Ethics 4 (2024) 1493–1506. doi:10.1007/s43681-023-00357-7.

Declaration on Generative AI

During the preparation of this work, the author(s) used generative AI to improve the initial diagrams of the framework as shown in Figure 1. Generative AI was also used as a literature discovery tool to identify preliminary sources within specific domains. All literature references obtained through the use of Generative AI were validated and thoroughly checked.