PART 5: FORESIGHTED SOLUTIONS FOR PRESENT PROBLEMS

22. Rewriting the Rules of the Game: Epistemological and Ontological Challenges at the Intersection of Legal Science and Data Science

Matheus Alles

Abstract. This study examines the epistemological and ontological challenges emerging from the intersection of data science and law. Through a multidisciplinary lens, it analyzes how predictive and language models in the legal domain challenge traditional legal theory and redefine key legal categories. The research highlights the tension between data-driven approaches and conventional legal reasoning, emphasizing the need for a reflexive legal rationality. This framework aims to critically integrate data science insights while maintaining the nuanced interpretative nature of legal thought. The article explores the implications of algorithmic decision-making in law, addressing issues of transparency, accountability, and the changing nature of legal knowledge. The methodology employs a multidisciplinary approach, combining conceptual analysis and literature review. By proposing a balanced approach that harnesses the power of data science without compromising legal principles, this study contributes to the ongoing dialogue on responsible integration of technology in legal practice and theory.
Keywords. Data science, legal epistemology, legal ontology, reflexive rationality, legal theory.

Introduction

The intersection between data science and law has become increasingly prominent, bringing not only opportunities for innovation and efficiency in legal practice but also profound challenges to the theoretical and conceptual foundations of law. This article proposes a critical reflection on how data science not only transforms legal practice but also disrupts the epistemological and ontological bases of law, demanding a radical reconstruction of legal theory.

This study adopts a global perspective, recognizing the diversity of legal systems and cultural contexts in which the intersection of legal science and data science manifests. In doing so, it seeks not only a comprehensive analysis but also an understanding of the nuances and specific challenges faced by different jurisdictions in the digital era

The central problem addressed in this study is the epistemological and ontological tension that arises when traditional legal reasoning, rooted in a hermeneutic and argumentative tradition, confronts the new forms of rationality and knowledge introduced by data science. This tension manifests itself in various dimensions, from the application of predictive models in the legal context to the reconfiguration of fundamental legal categories mediated by algorithms.

Faced with this scenario, the objective of this article is to investigate the epistemological and ontological implications of applying data science in law, identifying the challenges and opportunities that this intersection presents for legal theory and practice. It seeks to contribute to the development of a reflexive legal rationality, capable of critically integrating insights from data science without relinquishing the interpretive richness and contextual sensitivity of legal reasoning.

To achieve this objective, the study adopts an interdisciplinary methodological approach, combining conceptual analysis and literature review. Starting from an exploration of the epistemological and

ontological foundations of law, the article critically examines concrete examples of the application of data science in the legal context, such as the use of predictive models to anticipate judicial decisions.

Through this analysis, the study aims to identify the points of tension and the possibilities of reconciliation between traditional legal rationality and the new forms of knowledge introduced by data science. In doing so, the article seeks to contribute to the development of theoretical and methodological frameworks that enable a responsible and transparent integration of data science into law, promoting justice, equity, and social well-being in the digital age.

The methodology adopted in this study is rooted in a multidisciplinary approach, reflecting the complex nature of the intersection between legal science and data science. The research methodology comprises two main components.

First, a conceptual analysis, in which the study conducts a thorough examination of the epistemological and ontological foundations of law, drawing from established legal theory and philosophy. This provides the theoretical framework for understanding the traditional bases of legal reasoning and knowledge.

Second, research analyses recent scholarly works, case studies, and practical applications of data science in the legal domain. This includes examining predictive models for judicial decisions, the use of language models in legal drafting, and the application of machine learning algorithms in legal analysis. Through this dual approach, the study aims to identify the points of tension and potential reconciliation between traditional legal rationality and the new forms of knowledge and analysis introduced by data science.

The article is also structured in two main parts. The first part explores the epistemological disruption caused by data science in law, examining how new forms of rationality and knowledge challenge the traditional foundations of legal theory. The second part, in turn, addresses the ontological disturbance generated by the application of data science in law, investigating how fundamental legal categories are reconfigured and re-signified in this process.

Throughout these two parts, the article develops an argument in favour of the need for a reflexive legal rationality, capable of critically integrating insights from data science without relinquishing the interpretive richness and contextual sensitivity of legal reasoning. It is through this critical and constructive engagement, it is suggested, that it will be possible to face the epistemological and ontological challenges imposed by the intersection between law and data science, harnessing its transformative potential to promote justice and social well-being in the digital age.

22.1. Discussion

Law, as a discipline of study, is rooted in an epistemological tradition that values hermeneutic interpretation, logical argumentation, and the construction of coherent narratives (Dworkin, 2014). This tradition, which dates back to Roman jurisprudence and medieval exegesis, views law as a system of norms and principles that must be interpreted and applied to concrete situations through a process of legal reasoning (Berman, 1983).

Legal reasoning, in turn, is an intellectual process, not only methodological, but also from the intersection of this methodological tradition that involves an interpretive process in which there is a

dialogue between the norm and social facts and the subsumption of one to the other, in a system of network connections, with legal rationality being the guiding thread and also the foundation that fills a certain gap in this system.

However, this legal reasoning faces a new scientific challenge that promotes a dialogue through a distinct epistemological paradigm - data science.

The latter, unlike legal science, is extracted through a quantitative and qualitative analysis of large volumes of information and the identification of patterns and correlations (Hey, Tansley & Tolle, 2009) that promotes, in the combined application with legal science, a new challenge to the hermeneutic interpretation based on the understanding of phenomena with legal repercussions.

At first glance, it is assumed that this approach between legal science and data science is similar to a system of evidence in common law as a source of law, beyond what is exclusively posited, but in the face of what is practiced at the time of the application of the law, however with a maximization of the capacity of processed, stored and provided information.

However, data science goes beyond the empirical decision-making capacity, reflecting on the axis of the information used until entering the result of the decision and its comparison with a specific concrete situation.

Elements originated from automation condition correlations of segregation, influence on decision-making, and reliability, which ends up generating an epistemological tension between these two sciences.

An example is the predictability of judicial decisions that extend beyond a debate that is merely empirical from the perspective of precedents, but from a study of techniques for learning the predictability of decisions. In 2017, three authors wrote the article "A general approach for predicting the behavior of the Supreme Court of the United States" where a machine learning model was presented to predict the behavior of the Supreme Court of the United States of America. The model sought to encompass both the individual votes of Supreme Court justices and the overall outcomes of cases between 1816 and 2015 (Katz, Bommarito & Blackman, 2017).

The method used was called *Random Forest*, which evolves over time, taking advantage of feature engineering techniques that comprised more than 240,000 judge votes and 28,000 case outcomes. This interpretation was based on three principles: generality, consistency, and applicability (Katz, Bommarito & Blackman, 2017). The three principles aimed at general application, stable performance, and the possibility of repercussion outside the analyzed samples.

For this, the measurement of common and diverse elements between identification criteria, disagreement of the courts of origin, procedural aspects, and variables of historical behavior of decisions that encompass political directions, rate of disagreement, and reversal were used. During the analyzed period, the system obtained predictability criteria of 70.2% at the case levels and 71.9% at the individual vote levels (Katz, Bommarito & Blackman, 2017).

The system, which is beneficial in the face of an adaptation of precedents in the Supreme Court, runs the risk of generating a boomerang effect - which, when thrown, manifests the articulation of the petitioners in the face of knowledge about the vote of the decision-maker and how to adapt their

petition for analysis of agreement, disagreement, or overcoming (distinguishing and overruling). On the other hand, the return of the throw is to the detriment of critical thinking, considering that the agenda of discussion moves away from the legal repercussion of the case discussed to the line of the decision-maker's vote.

The link between claim and decision, in this context, transcends the mere resolution of social controversies under legal protection, shifting to an adaptation of the petitioners' reasons to the decision-maker's decision-making history. A phenomenon that raises a critical reflection on the core of legal discussion, which momentarily moves away from the reason of the law itself to orbit around the reason of the decision-maker, potentially compromising the integrity of the hermeneutic process.

From the perspective of Luhmann's Theory of Social Systems (2011), there is a reconfiguration of the dynamics between subsystems, where the guiding thread between fact and law is replaced by a connection between fact and precedent, the latter influenced by the decision-maker's history of motivations. In parallel, Dworkin's Theory of Law as Integrity (2014), originally conceived to strengthen legal certainty through a coherent system of precedents, is challenged by this new reality.

This intersection highlights the complexity of legal reasoning, especially in the face of technological advances and the proliferation of data, demanding a continuous re-evaluation of the epistemological foundations of legal science. In this scenario of transformation, the need to rethink legal epistemology emerges, seeking new approaches that can reconcile the hermeneutic tradition with the potentialities offered by data science.

The disruption caused by data science requires a critical re-evaluation of traditional legal epistemology, with the development of new theoretical frameworks that recognize the contribution of quantitative analysis without relinquishing the interpretive richness of legal reasoning.

However, the mere epistemological approach proves insufficient to face the emerging challenges. The intersection between data science and law raises a deeper disturbance that reaches the core of legal ontology. This disturbance transcends questions about methods and the nature of legal knowledge, reaching fundamental inquiries about the very nature and structure of legal reality.

This is because the introduction of data science in the legal field not only challenges the understanding of how law is known but questions what constitutes legal reality itself - which generates an ontological disturbance that emerges when traditional legal categories, constructed over centuries of legal thought, confront new forms of representation and analysis provided by data science (Hildebrandt, 2018).

Therefore, in rethinking legal epistemology, there is an inevitable conduction to reconsider the ontological bases of law, initiating a profound reflection on traditional legal categories and their adequacy to the contemporary technological context. As Mireille Hildebrandt (2018) observes, the integration of data technologies into law not only alters legal practices but challenges fundamental conceptions about what constitutes law and its entities.

Traditional legal ontology is built around abstract categories, such as "contract," "civil liability," and "crime," which emerge from interpretive and argumentative processes (Schauer, 2009). These categories are treated as real and objective entities that exist independently of the social and discursive practices that constitute them (Zheng, Jiang, Ding & Zaheer, 2022).

Data science, on the other hand, operates at a sub-symbolic level, identifying patterns and correlations that may not correspond to existing legal categories (Bengio, Lecun, Hinton, 2015).

Machine learning algorithms, for example, can identify clusters of cases or behaviors that do not fit into traditional legal taxonomies, revealing an alternative ontology based on statistical regularities (Mackenzie, 2015).

This ontological tension is aggravated by the problem of algorithmic opacity. Many machine learning algorithms operate as black boxes, producing results without providing a clear explanation of how those results were achieved (Pasquale, 2015). This lack of transparency raises questions about the accountability and legitimacy of algorithm-based decisions, especially when those decisions have significant legal consequences (Selbst & Barrocas, 2018).

The ontological tension and algorithmic opacity mentioned above find a pertinent illustration in the analogous and contemporary scenario of investments in Artificial Intelligence (AI).

As reported by Futurism (2023), Silicon Valley investors and Wall Street analysts are expressing growing concerns about the ability of technology companies to effectively monetize their AI initiatives. This case exemplifies how the introduction of new technologies, such as AI, can challenge not only established practices but also fundamental conceptual categories in a specific field.

The article highlights that, despite the enormous investments in AI - with Google, for example, projecting capital expenditures in excess of $49 billion in 2023 - there is growing uncertainty about the financial return on these technologies. This situation reflects the tension between traditional expectations of return on investment and the emerging reality of technologies whose value and impact are difficult to quantify in conventional ways.

The opacity mentioned in the original text finds a parallel in the difficulty investors have in understanding how these AI technologies will generate significant revenues. As noted by Jim Covello, senior analyst at Goldman Sachs, "Despite its expensive price tag, the technology is nowhere near where it needs to be in order to be useful" (Futurism, 2023).

The situation described in the article also resonates with the idea of an alternative ontology based on statistical regularities. AI models, by processing vast volumes of data and identifying patterns that may not correspond to traditional categories of business analysis, are effectively creating a new ontology of value and utility that challenges established conceptions in the world of investments.

The example of Silicon Valley and the world of AI investments demonstrates how the introduction of new technologies can disrupt not only operational practices but the fundamental ontological structures in fields such as finance, technology, and, by extension, law.

In parallel, there is an overreliance on AI as a science of analysis for other sciences, where the rupture of this reliance is already apparent when there is unverified use.

This situation of overreliance on AI, followed by growing concern about its real effectiveness, reflects what Luciano Floridi (2014) calls the infosphere - an increasingly complex and interconnected informational environment.

In the legal context, the infosphere challenges not only traditional epistemology but ontology, as there is an expectation that AI can autonomously and profitably solve complex legal problems, echoing a still limited understanding of the nature of legal knowledge and legal reality itself.

In law, this translates into a need to rethink fundamental legal categories and the very processes of legal reasoning and decision-making, especially considering the ethical and social implications of information technology.

The ontological disturbance raised by data science in the legal sphere is not an isolated event but mirrors a broader propensity for technological disruption in various spheres. The case of AI investments in Silicon Valley exemplifies how the implementation of new technologies can challenge established practices and basic conceptual categories.

This situation echoes the notion of an increasingly intricate and interconnected infosphere (Floridi, 2014), in which excessive trust in AI is followed by growing apprehension about its effectiveness.

In the legal field, this reality implies the need to rethink elementary legal categories and the very processes of reasoning and decision-making, seeking a balance between the efficiency promised by AI and principles such as justice, equity, and transparency.

Thus, the disturbance of legal ontology lies in its need to develop to encompass not only new technological entities but also new values and ethical precepts that emerge from the interaction between law and AI.

The ontological disruption caused by data science in the legal field demands a profound rethinking of the very nature of legal entities and categories. It is not merely a matter of adapting existing legal concepts to new technological realities but of recognizing that these technologies may fundamentally alter the ontological landscape of law.

This recognition requires a shift from a static, essentialist view of legal categories to a more dynamic, relational understanding of legal ontology. Legal entities and concepts must be seen not as fixed, pre-given realities but as emergent, context-dependent constructs that are shaped by the socio-technical practices in which they are embedded (Hildebrandt, 2018).

Such a relational ontology would acknowledge the constitutive role of data science in shaping legal reality while also preserving the normative and interpretive dimensions of law. It would require a dialogical approach that brings together legal expertise, technical knowledge, and ethical reflection to navigate the complex terrain of law in the age of data (Danaher, 2016).

Moreover, this ontological shift must be accompanied by a commitment to transparency, accountability, and public engagement. The opaque and proprietary nature of many data science tools and methods poses significant challenges to the rule of law and democratic governance (Pasquale, 2015). Ensuring that these technologies are developed and deployed in a transparent, accountable, and inclusive manner is crucial for maintaining the legitimacy and integrity of the legal system.

In conclusion, the ontological disturbance generated by the intersection of data science and law represents a profound challenge to the foundations of legal thought and practice. Addressing this challenge requires not only new theoretical frameworks and methodological approaches but also a fundamental rethinking of the nature of legal reality and the role of law in the digital age.

By engaging critically and constructively with the ontological implications of data science, the legal community can develop a more adaptive, responsive, and ethically grounded approach to the challenges and opportunities of the 21st century. This engagement is essential not only for the future of law but for the future of society as a whole, as we navigate the complex terrain of the infosphere and the emerging realities of the data-driven world.

The epistemology and ontology previously analyzed necessitate the exercise of legal rationality, traditionally based on a series of assumptions about the nature of legal reasoning, including the belief in the logical coherence of the legal system, the possibility of arriving at correct answers through rational argumentation, and the autonomy of law in relation to other forms of knowledge.

However, data science introduces a new form of rationality in law, an algorithmic rationality that operates differently from traditional legal rationality, privileging correlation over causality, probability over certainty, and efficiency over coherence.

It is this algorithmic rationality that challenges the autonomy of law, suggesting that legal decisions can be influenced by patterns and trends identified in data external to the legal system.

But how can this conflict between rationalities be overcome?

First, it is essential to recognize that both forms of rationality have valuable contributions to offer to the legal field.

Traditional rationality, with its emphasis on hermeneutic interpretation, logical argumentation, and the construction of coherent narratives, is essential to maintain the integrity and legitimacy of the legal system. On the other hand, algorithmic rationality, with its ability to identify patterns and correlations in large volumes of data, can provide valuable insights and support more efficient and evidence-based decision-making.

To reconcile these two forms of rationality, it is necessary to develop theoretical and methodological frameworks that allow for the responsible and transparent integration of data science into law. This implies not only establishing clear guidelines for the collection, processing, and use of data in the legal context, ensuring privacy protection, fairness, and non-discrimination.

The core lies in the cognizable demonstration of the algorithms used in a transparent manner, allowing data-based decisions to be understandable and appropriate to the ontological and epistemological context in which legal hermeneutics operates, echoing the connections and ruptures that relate to social dynamism.

This translates into the demonstration of legal reasoning associated with a demonstration of data reasoning, which tends to mitigate the risks associated with algorithmic opacity and strengthen confidence in the use of data science in law.

This context is reflected in the exercise of human activity associated with the exercise of activity developed by AI.

For example, the use of advanced language models to assist in drafting more equitable and inclusive contracts and policies can be seen as a positive application of data science in the legal domain. These models have the potential to identify linguistic biases, suggest more neutral alternatives, and promote

more accessible and understandable language. In this sense, they can contribute to the promotion of justice and equality, aligning with fundamental legal principles.

However, it is crucial to consider the epistemological and ontological implications of this approach. From an epistemological point of view, it is necessary to question the extent to which language models can adequately capture and represent the complexity and subtlety of legal reasoning from a human perspective. The drafting of contracts and policies involves not only the choice of words but also the interpretation of legal concepts, the weighing of principles, and the consideration of specific contexts. Can language models, however advanced they may be, encompass this epistemological depth?

Furthermore, there is an ontological concern about how these models can influence the very nature and meaning of legal concepts. If the drafting of contracts and policies becomes mediated by algorithms, this may lead to a reconfiguration of traditional legal categories. Notions such as equity, inclusion, and justice may acquire new meanings and interpretations, shaped by the logic and limitations of language models. This ontological disruption can have profound implications for legal theory and practice.

The core of the problem lies precisely in the exercise of human rationality so that it is not replaced by the predictive activity of AI or the application of concepts obtained through its rationality. It is not a matter of a resolutive method, but of constant and interdisciplinary exercise, as a mechanism for preserving the integrity and autonomy of legal reasoning.

This means the emerging convergence between data science and law through the development of a reflexive legal rationality, capable of critically questioning its own assumptions and adapting to new forms of knowledge and rationality introduced by data analysis. This reflexive rationality implies a commitment to transparency, comprehensibility, and accountability of algorithmic systems used in law, as well as an openness to interdisciplinary and collaborative forms of legal knowledge production.

22.2. Conclusion

Data science presents a new epistemological paradigm that challenges traditional legal reasoning. To reconcile these two forms of rationality, it is necessary to develop theoretical and methodological frameworks that allow for the responsible and transparent integration of data science into law.

This implies establishing clear guidelines for the collection, processing, and use of data in the legal context, ensuring privacy protection, fairness, and non-discrimination. The core lies in the cognizable and transparent demonstration of the algorithms used, allowing data-based decisions to be understandable and appropriate to the ontological and epistemological context of legal hermeneutics.

Data science goes beyond the empirical decision-making capacity, reflecting from the axis of the information used to the result of the decision and its comparison with the concrete situation. Elements such as automation, segregation, decision influence, and reliability generate an epistemological tension with legal science.

Recent studies, such as the machine learning model to predict the behavior of the U.S. Supreme Court, exemplify the disruptive potential of data science in law. However, to fully harness this potential responsibly, an interdisciplinary effort involving jurists, data scientists, and public policy makers is needed.

Furthermore, it is essential to consider the ontological implications of applying data science in law. The use of advanced language models to assist in drafting more equitable and inclusive contracts and policies, for example, raises questions about the ability of these models to capture the complexity and subtlety of legal reasoning. It is necessary to critically assess the extent to which these models can encompass the epistemological depth involved in interpreting legal concepts, weighing principles, and considering specific contexts.

Another concern is the influence these models can have on the very nature and meaning of legal concepts. Algorithmic mediation in the drafting of contracts and policies may lead to a reconfiguration of traditional legal categories, with notions such as equity, inclusion, and justice acquiring new meanings shaped by the logic and limitations of language models. This ontological disruption requires a deep reflection on the implications for legal theory and practice.

The implications of this study for future research are significant. There is a pressing need for empirical investigations into how different legal systems are adapting to the integration of data science. Furthermore, the development of ethical and regulatory frameworks for the application of AI in law emerges as a critical area for future research. In practice, the results of this study can inform the development of legal curricula, public policies, and organizational strategies for a more effective and ethical integration of data science in the legal field.

Faced with this complex scenario, the need emerges for the legal community to develop a reflexive rationality, capable of critically questioning its own assumptions and adapting to new forms of knowledge and rationality introduced by data analysis. This reflexive rationality implies a commitment to transparency, comprehensibility, and accountability of algorithmic systems used in law, as well as an openness to interdisciplinary and collaborative forms of legal knowledge production.

References

Alexy, R. (1989). A theory of legal argumentation: The theory of rational discourse as theory of legal justification. Oxford: Clarendon Press.

Bengio, Y., LeCun, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Berman, H. J. (1983). Law and revolution: The formation of the Western legal tradition. Cambridge, MA: Harvard University Press.

Dworkin, R. (2014). O império do direito (3rd ed.) (J. L. Camargo, Trans.). São Paulo: Martins Fontes.

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. Oxford: Oxford University Press.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). The fourth paradigm: Data-intensive scientific discovery. Redmond, WA: Microsoft Research.

Hildebrandt, M. (2018). Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics. University of Toronto Law Journal, 68(supplement 1), 12-35.

Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE, 12(4), e0174698.

Luhmann, N. (2011). Introdução à teoria dos sistemas (3rd ed.) (A. C. A. Nasser, Trans.). Petrópolis: Vozes.

Mackenzie, A. (2015). The production of prediction: What does machine learning want? European Journal of Cultural Studies, 18(4-5), 429-445.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Cambridge, MA: Harvard University Press.

Rouvroy, A., & Berns, T. (2013). Algorithmic governmentality and prospects of emancipation: Disparateness as a precondition for individuation through relationships? Réseaux, 1(177), 163-196.

Schauer, F. (2009). Thinking like a lawyer: A new introduction to legal reasoning. Cambridge, MA: Harvard University Press.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham Law Review, 87(3), 1085-1139.

Zheng, Y., Jiang, S., Ding, W., & Zaheer, A. (2022). Ontology-based knowledge representation and semantic topic modeling for intelligent trademark legal precedent research. World Patent Information, 68, 102098.

23. People-Centered Justice AI: Data Dimensions for Embracing a Responsible Digital Transformation

Julio Gabriel Mercado

Abstract. The digital transformation of justice, driven by AI, must be guided by a people-centered approach to ensure its responsible and effective implementation. Simply digitizing the system will not close existing access gaps. Instead, adopting Open Justice principles, emphasizing transparency, accountability, and public participation, is crucial to driving the necessary cultural and organizational shifts. Particularly, Open Justice promotes the publication of judicial data in open, reusable formats, which is key to fostering innovation and inclusivity in AI-driven systems. The quality of the available data will largely determine whether AI's benefits are distributed equitably. To achieve this, five critical dimensions for the publication of data, i.e., standardization, accessibility, completeness, cybersecurity, and privacy, must be addressed. Tackling these issues requires coordinated efforts at national and global levels to ensure that AI advancements serve the public interest and uphold human rights.
Keywords. People-centered Justice - Artificial Intelligence - Open Justice - Open Judicial Data - Access to Justice

Introduction

The integration of artificial intelligence (AI) into the justice sector represents a transformative opportunity, but one that must be grounded in a people-centered approach to ensure a responsible implementation. As highlighted by the OECD (2023), achieving this approach requires collaboration across the entire justice system, including courts, prosecutors, police, and correctional institutions, in order to establish robust regulatory frameworks and the necessary institutional strategies.

While digital transformation has progressed in some areas, fundamental change in justice systems has been slow, with core functions remaining largely unchanged for centuries (Muller et al., 2013). Meanwhile, a significant justice gap persists, with 1.5 billion people worldwide unable to resolve their justice problems and two-thirds of the global population lacking meaningful access to justice (Task Force on Justice, 2019). This gap disproportionately affects vulnerable groups such as women, low-income individuals, and ethnic minorities, exacerbating structural inequalities.

In this context, AI offers the potential to expedite judicial processes, reduce case backlogs, and assist in decision-making. However, a simple reliance on technology without addressing deeper cultural and organizational shifts may lead to failure, and even to increased digital exclusion (Addo et al., 2024). The principles of Open Justice, which emphasize transparency, accountability, and public engagement, provide a framework for addressing these challenges (Elena et al., 2019).

In particular, by encouraging the publication of judicial data in open, reusable formats, Open Justice ultimately supports the development of responsible AI systems that respect human rights and help address bias. The governance of data not only ultimately shapes the governance of AI but also largely determines the extent to which its benefits might be distributed equitably and the risks associated with its implementation can be mitigated (Datasphere Initiative, 2024). In this sense, addressing some critical dimensions of judicial data publication is key to ensure AI's eventual success in assisting the delivery of justice in a responsible manner.

This paper aims to define the key dimensions of judicial data (i.e., standardization, accessibility, completeness, cybersecurity, and privacy) that must be addressed to embrace a responsible deployment of AI in the justice sector. It explores the challenges of balancing these dimensions and the importance of coordinated national and global efforts to align AI's use in justice with public interest, fundamental rights, and fairness. Ultimately, it advocates for a people-centered approach that emphasizes inclusivity, transparency, and accountability, ensuring that the benefits of digital transformation are distributed equitably, contributing to closing the global justice gap.

23.1. Digital Transformation, Open Justice and AI

The digital transformation of justice, particularly in the context of the use of AI, should aim at the adoption of a people-centered approach (OECD, 2023). A people-centered approach to justice can be defined as one that prioritizes a unified vision and purpose aimed at making the justice system more responsive to peoples' needs. This can be done through designing and delivering services based on the justice journey of different groups, with a focus on those populations that face the greatest barriers to accessing justice.

According to recent figures, a total of 1.5 billion people worldwide experience justice problems that they cannot resolve. They are victims of unreported violence or crime, or they have civil or administrative justice problems that they cannot resolve. Meanwhile, a total of 5.1 billion people, representing two-thirds of the world's population, are currently considered to lack meaningful access to justice. This justice gap is both a reflection of and a contributor to structural inequalities, which most often affect individuals and collectives in disadvantaged situations, such as women, low-income persons, gender-diverse persons, or persons belonging to ethnic minorities (Task Force on Justice, 2019).

The persistence of this justice gap, which can be defined as a person's inability to obtain an effective, legally sound and actionable resolution to a dispute, makes the use of AI and its promise to expedite judicial processes, reduce persistent case backlogs, and assist judicial decision-making, a critical area of focus. However, transforming justice requires more than just deploying digital tools. The adoption of Open Justice principles by judicial institutions can support the cultural shift that they need to advance digital transformation in a people-centered and inclusive manner.

Open Justice is a vision that calls for transparency and accountability, making the workings of the justice system clear and accessible (Elena et al., 2019). It also promotes the publication of justice data in open formats, fostering informed public engagement and enabling evidence-based innovation. Open Justice also calls for collaboration between justice institutions and stakeholders to drive digital transformation in a way that focuses on creating social value through more responsive and tailored processes that meet people's justice needs.

Open Justice provides tools for the development of responsible AI in justice, understood as the creation and deployment of AI systems that seek to ensure positive social impact, respect for people's rights, and minimization of bias and error, based on compliance with current legal standards and shared ethical principles (Adams, 2024).

23.2. The role of judicial data for delivering people-centered AI

Open Justice provides judicial institutions with a starting point for adopting mechanisms for transparency, accountability, and public oversight of the quality of the data they publish. This can include establishing publication priorities, participatory and collaborative mechanisms to protect the rights and interests of the people they serve throughout the data publication cycle, particularly by ensuring that the published data reflects the experience with justice of individuals and groups in vulnerable situations.

The availability and quality of judicial data are crucial for developing AI systems. Open Justice fosters a robust data ecosystem that promotes data publication and reuse, supporting advocacy, innovation, and the redesign of processes to better meet people's legal needs (World Justice Project, 2023). However, merely making data available is insufficient; publication policies must include safeguards to ensure data quality and integrity, as well as protect the privacy of individuals involved in judicial proceedings, particularly when AI systems rely on these data.

The importance of data in the development of AI systems is not new, but it becomes increasingly critical as their use becomes more ubiquitous (UNESCO, 2024a), while institutions strive to understand and regulate a technology whose evolution and scope have yet to be fully grasped. In 2018, the European Commission for the Efficiency of Justice (CEPEJ) presented ethical principles for the use of AI in justice through its European Ethical Charter. This charter emphasized the importance of using high-quality, certified data to train AI systems, while maintaining the traceability of this data to prevent changes that could influence judicial decisions. It also underlined the need to address privacy concerns related to data used in the development of AI systems (CEPEJ, 2018). In line with their original position, in a more recent informative note the CEPEJ emphasizes the fact that any existing gaps or biases in judicial data can significantly impact the overall validity of AI-generated results, thus reducing the effectiveness and fairness of AI systems in the justice sector (CEPEJ, 2024).

The recently approved European AI Regulation is a significant step towards the establishment of a common global framework for the development of responsible AI, which has clear implications for its use in the field of justice (Regulation (EU) 2024/1689). This regulation emphasizes the clear impact that AI can, and most likely will, have on democracy, the rule of law, and individual rights. In particular, as a result of this regulation, justice AI enters a category considered as high-risk, which is therefore subject to stringent transparency, documentation, and oversight requirements.

The European AI regulation also highlights the importance of high-quality data and access to it as a means of structuring and ensuring the safe operation of AI systems, while avoiding becoming an additional source of social discrimination. To this end, it urges the establishment of appropriate management and governance practices for the data used in the context of AI systems, in order to achieve high quality datasets for training, validation and testing. In this regard, attention is drawn to biases that are considered to be inherent in the datasets used and, as such, may affect the outcomes of AI systems, thereby perpetuating and amplifying existing discrimination against certain vulnerable groups. To this end, it is established as a requirement that datasets be as complete and error-free as possible, which should not affect the use of techniques to protect the privacy of individuals.

While this regulation provides a starting point for global regulation on this topic, there have been significant advancements, particularly from Global South countries like Brazil, in addressing the availability and quality of data for the development of AI. The National Justice Council (Conselho Nacional de Justiça, CNJ), the institution responsible for overseeing the administration of justice in the

country, has been working for several years on establishing common standards for the publication of cases and documents, particularly aimed at facilitating their reuse in AI systems. As this paper is being written, the CNJ is proposing an enhanced regulation that reflects the data needs and risks arising from the increasing use of generative AI tools.

Meanwhile, from a universal standpoint, UNESCO is currently developing Guidelines for the use of AI systems in courts and tribunals. These guidelines will emphasize key principles for AI training data, such as transparency, quality, integrity, and data governance. Key aspects addressed by these Guidelines will include the need for robust data governance frameworks and infrastructures to protect personal data and promote responsible data-sharing practices, enhanced privacy protections, enhancing transparency regarding training data, and empowering deployers and users to effectively evaluate the quality and integrity of data (UNESCO, 2024b).

23.3. DATA NEEDS FOR A PEOPLE-CENTERED JUSTICE AI

Bias in legal data significantly influences the development of AI applications, potentially leading to unfairness or errors in prediction tasks, or biased information generation in question-answer tasks (Sargeant et al., 2024). For AI to work well, it needs a large volume of diverse and accurate data. Biases in AI systems can result from incomplete or unrepresentative data, leading to unfair outcomes and perpetuating existing disparities.

To address these challenges, it is critical to identify key needs for the availability and publication of judicial data, recognizing the transformative impact that AI systems can have on the delivery of justice, while also addressing various existing frameworks that come into play in the development of these systems, particularly when it comes to aligning them with the protection of fundamental rights, compliance with legal requirements, promotion of sustainability, and the maximization of the public interest (Belli et al., 2024). In this regard, it is essential to understand that the generation, classification, and use of these data must be conducted in a responsible and inclusive manner, through transparent, accountable, and participatory mechanisms that emphasize achieving more people-centered justice by ensuring that people's rights and needs are respected and represented throughout the data lifecycle.

There are five main dimensions that are critical to the effective publication and use of justice data in the context of AI. These require action by justice institutions that seek to promote the genuine inclusion of all communities in shaping the development of a people-centered justice AI. The first aspect is **standardization**. It is vital to establish unified standards for data publication to ensure consistency and quality. Currently, judicial data is often published in an *ad-hoc* manner, which leads to inconsistencies and difficulties in using that data to inform system-wide innovations. Standardizing data formats and protocols can enhance the interoperability and reliability of judicial data, facilitating its use in AI systems, as well as in other applications.

The second aspect is **accessibility**. To be effectively used, judicial data must be easily accessible. This requires making data available through open, reusable formats, while ensuring that it can be integrated from multiple sources. Open judicial data portals allow for direct access, verification, and reuse of data. This aspect supports transparency and, therefore, improves the quality of AI systems by providing a reliable and traceable resource. However, balancing open access with privacy concerns remains a challenge that publication policies need to address.

Thirdly, the **completeness** of data is crucial for ensuring that AI systems can offer fair and equitable responses. In many justice systems, data often lacks representation of diverse groups, such as women, ethnic minorities, low-income persons, gender-diverse persons, or persons with disabilities. This underrepresentation can limit the effectiveness of AI systems and perpetuate existing inequalities. Efforts must be made to improve data collection and representation, guided by principles such as data equity, which emphasizes the need for inclusive data practices that respect human rights and promote fairness.

The fourth aspect is **privacy**, whose balance with data accessibility and completeness is often complex but necessary. As AI systems increasingly rely on large datasets, it is essential to protect personal information while allowing for a meaningful and effective use of data. To achieve this balance, publishing institutions can resort to various measures, which should encompass the whole data publication cycle. These include conducting risk assessments to identify and mitigate privacy risks, applying data minimization principles to ensure that only the necessary data is collected and used, and maintaining ongoing human oversight to ensure that privacy concerns are continuously addressed and managed.

Finally, the aspect of **cybersecurity** clearly impacts the necessity to protect both the data and the systems used, as well as to address ethical and privacy issues related to data handling. While this dimension encompasses a broader context within the field of AI (i.e., focusing on the security of the systems themselves) the approach to data usage requires a holistic perspective. This perspective should encompass not only the protection of data in its initial dimensions (capture, storage, and management) but also risk controls surrounding the information security involved.

These five dimensions generate tensions and balances that must be addressed and discussed by judiciaries at two levels. One is the intra-systemic or primary level of justice institutions or systems. This first level can be dealt with through national or sectoral AI strategies or public policies, whereby the institution or system takes a position and combines its interests with those of the stakeholders in its direct sphere of influence. These initiatives should aim to guide the ethical and responsible development of AI systems in the judiciary through measures and approaches that could range from soft measures, such as the establishment of ethical guidelines or standards, to new regulations or legislation (OECD, 2024).

On the other hand, the traditional approach to judicial data governance, which focuses solely on the legal requirements within each specific jurisdiction (i.e., on the scope within which each institution carries out its jurisdictional work), is not compatible with the nature of AI system development. This development is inherently polycentric (Xue, 2024), and therefore transcends the boundaries of national jurisdictions. Consequently, it is essential for the above-mentioned four dimensions to also be discussed at a secondary, inter-systemic level involving various actors, including legislative bodies, data-publishing institutions, companies that develop and use AI systems, and justice system users, at a global scale. In this regard, multi-stakeholder forums, such as the Internet Governance Forum's Dynamic Coalition for Artificial Intelligence, are working to connect and empower all populations, ensuring that AI systems are developed from a people-centered perspective and can therefore help them reap the benefits from digital transformation, in terms of closing the justice gap that prevent them from meaningfully accessing their rights (Belli et al., 2024).

CONCLUSION

To be people-centered, digital transformation processes in justice should not be limited to the adoption of new technologies. They must be supported by Open Justice policies that guide the necessary cultural and organizational shifts to mitigate digital exclusion and ensure that the benefits of digitalization reach all individuals equitably.

The role of data in developing responsible AI systems for justice must be addressed, as there is a key interrelation between these data, how their governance is conducted, and the need to mitigate biases and errors that can affect equity and justice in the application of AI tools within the judicial process.

Therefore, implementing AI systems in the justice sector requires a rigorous and well-structured approach to data publication and management. Based on these considerations, five fundamental dimensions should be addressed, each presenting challenges and opportunities that must be tackled jointly, both at the primary level by the institutions publishing the data and at the secondary level, given the polycentric nature of the AI value chain, by the various actors involved in the process.

Standardization, accessibility, completeness, privacy and cybersecurity are the main five dimensions identified that should be taken into consideration to govern judicial data, as a prerequisite to ensure a successful integration of AI in the provision of justice that focuses on peoples' needs and on closing the justice gap effectively. Addressing these five dimensions in a collaborative and open manner will be key to facilitate a digital transformation of justice that serves the justice needs of society as a whole.

References

Adams, R. (2024, May 27th). Responsible AI practices for business leaders (Episode 13). [Video podcast episode] *Unpacked*. Lab 45. https://lab45thinktank.com/podcast/responsible-al-practices-for-business-leaders-with-dr-rachel-adams-ceo-gcg/

Addo, P. M., Baumann, D., McMurren, J., Verhulst, S.G., Young, A. & Zahuranec, A.J. (2021). *Usages émergents des technologies au service du développement : un nouveau paradigme des intelligences.* Policy Paper. AFD. https://www.afd.fr/fr/ressources/technologies-developpement

Belli, L. & Gaspar, W. (2024). AI Transparency, AI Accountability, and AI Sovereignty: An Overview. En L. Belli & W. Gaspar (eds.). *The Quest for AI Sovereignty, Transparency and Accountability.* Official Outcome of the UN IGF Data and Artificial Intelligence Governance Coalition. FGV, 21-28. https://direitorio.fgv.br/en/publication/quest-ai-sovereignty-transparency-and-accountability

CEPEJ (2024). *Use of Generative Artificial Intelligence (AI) by judicial professionals in a work-related context.* CEPEJ Working group on Cyberjustice and Artificial Intelligence. https://www.coe.int/en/web/cepej/-/information-note-on-the-use-of-generative-artificial-intelligence-ai-by-judicial-professionals-in-a-work-related-context

CEPEJ (2018). *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their Environment.* https://www.coe.int/en/web/cepej/cepej-european-ethical-charter-on-the-use-of-artificial-intelligence-ai-in-judicial-systems-and-their-environment

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Datasphere Initiative (2024, June 25th). 6 reasons why Data matters for AI. *The Datasphere*. https://www.thedatasphere.org/news/6-reasons-why-data-matters-for-ai/

Elena, S. & Mercado, J.G. (2019). A Theoretical Approach to Open Justice. In Elena, S. (coord). *Open Justice: An Innovation-Driven Agenda for Inclusive Societies*. SAIJ, 17-40. http://www.bibliotecadigital.gob.ar/items/show/2569

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence.* https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Muller, S. & Barendrecht, M. (2013). *The Justice Innovation Approach: How Justice Sector Leaders in Development Contexts Can Promote Innovation.* The World Bank Legal Review: Legal Innovation and Empowerment for Development, Vol. 4, 17-30. https://doi.org/10.1596/9780821395066_CH02

OECD (2024). *Governing with Artificial Intelligence: Are Governments Ready?* OECD Artificial Intelligence Papers, No. 20. OECD Publishing. https://doi.org/10.1787/26324bc2-en

OECD (2023). *Recommendation of the Council on Access to Justice and People-Centered Justice Systems* (OECD/LEGAL/0498). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0498

World Justice Project (2023). *Disparities, Vulnerability, and Harnessing Data for People-Centered Justice*. WJP Justice Data Graphical Report II. https://worldjusticeproject.org/our-work/research-and-data/wjp-justice-data-graphical-report-ii

Task Force on Justice (2019). *Justice for All. Final Report.* Center on International Cooperation. https://cic.nyu.edu/wp-content/uploads/2023/02/english_task_force_report_27jun19-min_compressed.pdf

Sargeant, H. & Magnusson, M. (2024). *Bias in Legal Data for Generative AI*. Workshop on Generative AI and Law (GenLaw) at International Conference on Machine Learning 2024. https://blog.genlaw.org/pdfs/genlaw_icml2024/9.pdf

UNESCO (2024a). *Survey on the Use of AI Systems by Judicial Operators*. UNESCO Global

Judges' Initiative. https://unesdoc.unesco.org/ark:/48223/pf0000389786

UNESCO (2024b). *Draft UNESCO Guidelines for the Use of AI Systems in Courts and Tribunals*. UNESCO Global Judges' Initiative. https://unesdoc.unesco.org/ark:/48223/pf0000390781

Xue, J.H. (2024). Polycentric Theory Diffusion and AI Governance. In C. Aguerre, M. Campbell-Verduyn & J.A. Scholte (eds.), *Global Digital Data Governance. Polycentric Perspectives*. London: Routledge, 223-237. https://www.researchgate.net/publication/377135755_Polycentric_Theory_Diffusion_and_AI_Governance

24. Fostering AI Research And Development: Towards A Trustworthy LLM. Mitigating Compliance Risks Illustrated via Scenarios

Liisa Janssens, Saskia Lensink and Laura Middeldorp

Abstract

The rapid growth of Large Language Models (LLMs) challenges the Rule of Law, necessitating a thorough examination of their disruptive potential. This paper highlights the importance of adhering to these principles for responsible LLM deployment. Using a scenario-based approach, we show how specific design choices can lead to unintended consequences. We present a hypothetical case of developing an LLM, focusing on the inclusion of an opt-out option for personal data removal. Two scenarios are explored: one with and one without this option, illustrating how this decision impacts compliance with the Rule of Law. The paper emphasizes anticipating regulatory requirements and linking design choices to legal principles during research and development. By addressing these considerations early, stakeholders can better prepare for legislative changes and mitigate compliance risks. This paper aims to guide end-users, policymakers, researchers, and industry participants on mitigating risks and ensuring responsible LLM deployment.
Keywords: Large Language Models, Design-choices, Opt-out option, Global Majority, Compliance, Rule of Law, Research and Development, Deployment.

Introduction

The European Commission published the AI Act in the Official Journal on the 12th of July 2024[182]: a legal framework guiding the development and deployment of AI systems. The AI Act aims to uphold the values of the European Union and at the same time leverage the capabilities of AI. Further developments on the AI Act will make compliance a moving target. This is in the nature of (new) laws and regulations, since these need to be understandable, transparent and trustworthy but are not supposed to be set in solid bedrock. Interpretations of the meaning of the AI Act via case law (jurisprudence) is yet to commence, and this can lead to questions how to innovate with AI with the aim to deploy these new innovations. The extra complicating factor lies in the fact that AI is a moving target as well. This combination creates one of the biggest challenges for all parties who want to deploy AI aligned with laws and regulations. The nature of law complicates early-stage research initiatives which strive for, or promise, alignment with the AI Act when it is time for deployment of these models. The aim of the AI Act is not to stifle innovations, it asks for a forward-looking eye: the ability to foresee what it takes to become compliant when the time has come to deploy, in the legal reality, what has been made.

The question that becomes relevant to all (from developers to end-users) is how to deal with the moving target of compliance in the research and development process of AI models? Mitigating future compliance issues with all the ins and outs of the AI Act is difficult, but this does not mean that future compliance issues cannot be scoped, addressed and tried to be mitigated.

[182] "Today, on July 12, 2024, EU Regulation No. 1689/2024 laying down harmonized rules on Artificial Intelligence ("Regulation" or "AI Act") was finally published in the EU Official Journal and will enter into force on August 1, 2024."

First, the AI Act can be seen as a protection of the Rule of Law: *"It aims to protect fundamental rights, democracy, the rule of law and environmental sustainability from high-risk AI, while boosting innovation and establishing Europe as a leader in the field. The regulation establishes obligations for AI based on its potential risks and level of impact."* (European Parliament 2024). In this paper the lens of the Rule of Law will be presented as a lens to scope future compliance issues with the AI Act. The Rule of Law also allows for the review of the legal effect of specific design choices in LLMs from a more fundamental perspective.

An informed viewpoint about how future compliance risks could manifest can be provided via a scenario-based approach (Janssens, Lucassen, Middeldorp, Lobbezoo, & Schoenmakers, July 2024). Via this approach insightful perspectives are given which can inform decision-makers about what is at stake and what could be the best design choices in order to comply with the ambition of deploying an LLM for the public good. Part of the scenario-based approach is the lens of the Rule of Law (Stein, 2019).[183] In this paper we use this approach to analyse potential norm violations of the Rule of Law via a hypothetical use case: an LLM which is built from scratch. This use case takes as point of departure that it is necessary to gain control, as much as possible, over the data used to train an LLM. We will refer to this hypothetical LLM as 'LLM-from-scratch'.

A design choice that can be made by the development team of the LLM-from-scratch is whether an opt-out option before training needs to be included. Typically, training data of LLMs are curated beforehand by using algorithms that remove personal identifiable information. The performance of these algorithms, however, is not perfect and personal information can still reside within the training dataset after curation. Therefore, in addition to algorithmic means to remove personal identifiable information, one could opt for including an 'opt-out' option that allows the public to check the curated dataset for the presence of any of their personal information and request that this information is taken out of the dataset before it is used to train the LLM.

This paper investigates how the design choice of the implementation of an opt-out option within the development phase of an LLM is related to the protection of the Rule of Law and therewith fostering the well-being of the global majority. Two types of scenarios are investigated: one where an opt-out option is implemented within the development phase and one where the opt-out option is not present. For both scenarios, we will evaluate the hypothetical legal effect of the inclusion/exclusion of an opt-out option and how this affects the tenets of the Rule of Law. For instance, how does the design choice of presenting an opt-out option prior to training the model impact norms like respecting justice, legitimacy and transparency? In addition, what could be possible consequences if an LLM built from scratch does not implement an opt-out option before training?

This paper is structured as follows: Section 2 gives an overview of the opt-out option, the Rule of Law and explains the scenario-based method. Section 3 describes the analysis performed on the two scenarios, one where an opt-out option is implemented and one where an opt-out option is not implemented and investigates how the decision of including or excluding an opt-out option affects the tenets of the Rule of Law. We conclude this paper with Section 4 by providing recommendations

[183] "The Rule of Law refers to a principle of governance in which all persons, institutions and entities, public and private, including the State itself, are accountable to laws that are publicly promulgated, equally enforced and independently adjudicated, and which are consistent with international human rights norms and standards. It requires, as well, measures to ensure adherence to the principles of supremacy of law, equality before the law, accountability to the law, fairness in the application of the law, separation of powers, participation in decision-making, legal certainty, avoidance of arbitrariness and procedural and legal transparency." Robert Stein, What Exactly Is the Rule of Law? 2019, p. 188.

which can be inspiring for everyone who is planning to develop and deploy LLMs and wants to deal with future compliance issues. The recommendations in this paper can be used by decision-makers to make an informed decision on incorporating an opt-out option in the development phase of an LLM.

24.1. DEVELOPMENT

In this section the relation between the design choice made during development of the opt-out option and the Rule of Law is explained. Although the focus in this article is set on the Rule of Law as one of the foundational principles of the European Union (Article 2 Treaty on the European Union), the Rule of Law is also one of the foundational principles of the United Nations. The Rule of Law is meant to foster well-being of all human beings, amongst who the global majority is part, as can be read in the clarification of the United Nations.[184]

The United Nations clarifies the Rule of Law as follows:

> *"It requires measures to ensure adherence to the principles of supremacy of the law, equality before the law, accountability to the law, fairness in the application of the law, separation of powers, participation in decision-making, legal certainty, avoidance of arbitrariness, and procedural and legal transparency." (United Nations, 2024)*

Therefore, our assessment of an LLM in the EU context can be an example how the tenets of the Rule of Law (accountability, transparency, liability and contestability) can be globally applicable to foster the well-being of the global majority. Societies all over the world can learn from this hypothetical use case in shaping their own rules, regulations and policies around the development of LLMs to foster and ensure that the well-being of people, i.e. the global majority, is maintained. In case an opt-out option is implemented, it is of importance that the datasets used to train the LLM are open and can be searched by everyone in the world that may be present in the datasets. Access to the internet is a prerequisite to make this possible. However, since not everyone has access to the internet (Bradshaw, 2001) it is desirable -when the datasets contain personal information of people from countries who have no or limited access to the internet- to take auxiliary precautions. For example, independent institutions may be asked to perform a check for persons who are unable to do this.

24.1.1. THE OPT-OUT OPTION AND THE RULE OF LAW

Good governance is about accountability, transparency, (addressing) liability and contestability. The aim of the mechanisms of the Rule of Law is to produce a government that is legitimate and effective. Good governance is about legitimate, accountable and effective ways of obtaining and using public power and resources in the pursuit of legitimate goals.

When a government has the ambition to use an LLM for the public good this model needs to foster good governance. The opt-out option is an example of a design choice which underpins this ambition.

24.1.2. THE OPT-OUT OPTION

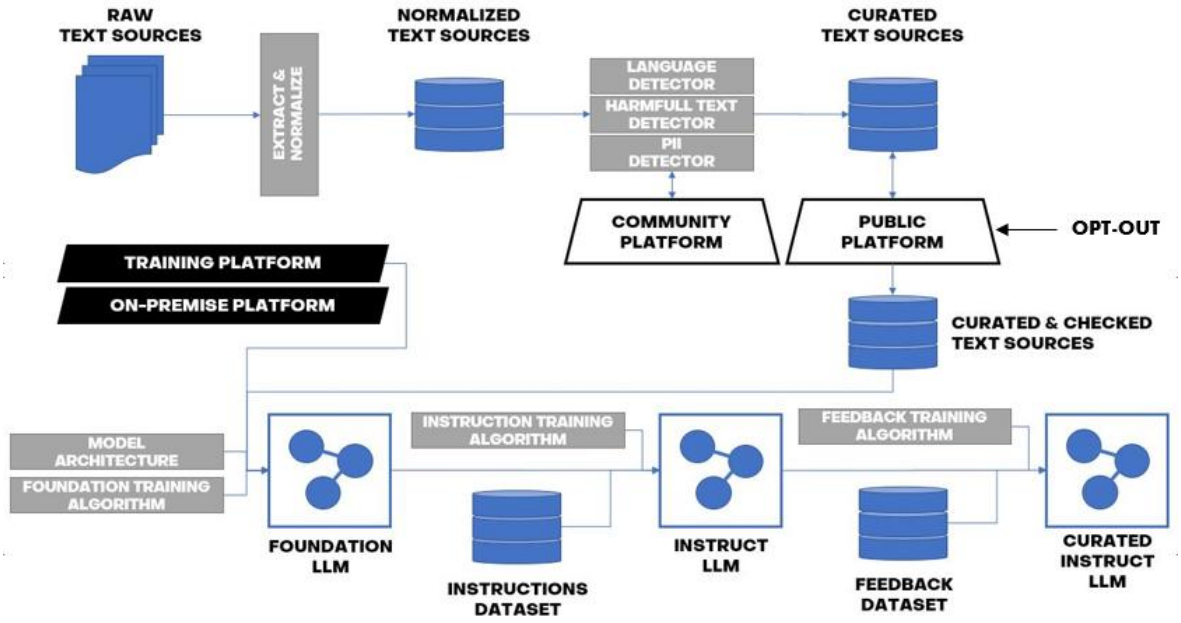
Figure 1 illustrates the pipeline of the LLM-from-scratch. The opt-out option takes place via a public platform prior to training the model. Due to privacy legislations and agreements made with the

[184] "Rule of law issues includes emerging and critical issues such as the proliferation of hate speech and incitement to violence; preventing radicalization/violent extremism; climate change and the environment impacting on the security and livelihoods of people; and the complexities of artificial intelligence and cybercrime." (United Nations, 2024)

contributors of the data sources for LLM-from-scratch, the entire database cannot be made accessible to the general public. Instead, the database can be searched by the public to examine whether their personal information is present in the database. In other words, the database is not made available but is available to be searched.

The documents checked for opt-out will be removed from the database.

Figure 1. Pipeline of the LLM-from-scratch



24.1.3. SCENARIO-BASED METHOD

LLMs can be beneficial but at the same time bring about (unintended) drawbacks that can challenge the Rule of Law. There is a need for a method that can be used to identify the tension between the Rule of Law and the consequences of design choices in the development of LLMs.

We have developed a method that identifies the tensions between the Rule of Law and emerging and disruptive technologies, of which an LLM is an example, by means of a scenario analysis. Scenarios are a useful instrument to simulate a specific environment through which an LLM can be deployed. By mapping the events in the scenario to (tenets of) the Rule of Law, advice, which is informed by European norms and values, can be shaped regarding design choices of an LLM on both technical and functional level. In addition, a scenario provides a contextualization of how the LLM will operate in practice to identify possible norm violations which need to be mitigated.

The opt-out option within the LLM-from-scratch pipeline can be regarded as a technical requirement. Is it necessary that an opt-out option is incorporated before training the model? And what are possible consequences of (not) incorporating an opt-out option? To investigate this, two scenarios will be analysed, one without and one with an opt-out option implemented. In each scenario, we will map the tenets of the Rule of Law to the events and consequences happening in the operational context of the LLM-from-scratch. Furthermore, both scenarios highlight the potential benefits and risks to maintaining the Rule of Law.

24.2. DISCUSSION

24.2.1. SCENARIO ANALYSIS: OPT-OUT OPTION IN TWO SCENARIOS

This section analyses the role of an opt-out option and the effect it has on the tenets of the Rule of Law by means of scenarios. Before the scenarios are introduced, remarks on the data curation and the relationship with an opt-out option are presented.

LLMs are typically trained on a large amount of textual data, where the data are curated before training using curation algorithms. These algorithms remove harmful texts and personal identifiable information as displayed in **Figure 1.** The accuracy rate of curation algorithms is rather good, between 90-95% (Dasgupta, Ganesan, Kannan, Reinwald, & Kumar, 2018), however given a population of, say, X million, a personal detection rate of 5-10% can still be substantive. Even when only one case 'slips through' dangers can already arise. This one case can erode legal certainty on individual level, as well as -when this is widely spread via media- on a societal level. A case like that can have a big impact on the trust of potential end users in the LLM.

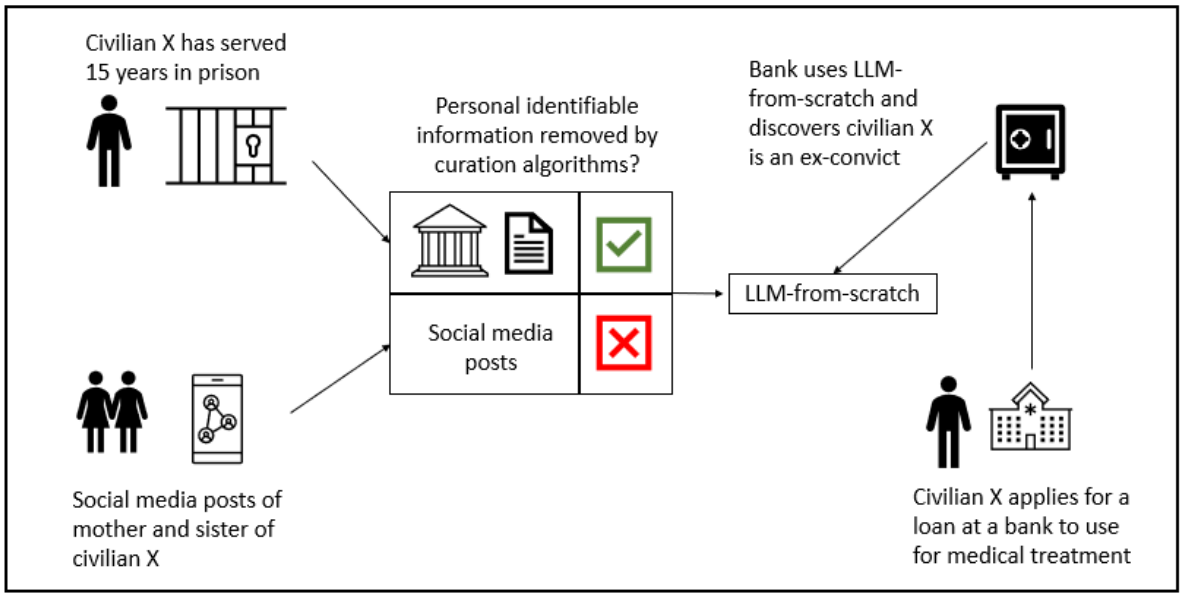
The next two sub-sections present the scenarios and showcase how the tenets of the Rule of Law may be challenged.

24.2.1.1. Scenario One: No Opt-Out Option Implemented

Figure 2 visualizes the scenario where no opt-out option has been implemented. Civilian X has served 15 years in prison. The LLM-from-scratch uses case law describing the timeline of the crime and the verdict of civilian X to train the model. The mother and sister of civilian X post on social media they are happy he has been released from prison. These social media posts are also included in the training dataset. The curation algorithms manage to remove personal information of civilian X from case law, but not from the social media posts of the mother and sister.

Civilian X has recently been diagnosed with a rare form of cancer which needs expensive treatment. Civilian X applies for a loan for the treatment at the bank which conducts a background check on civilian X using the LLM-from-scratch. The bank discovers civilian X is a convicted murderer who has been in prison for 15 years. The bank is doubting: should they accept the loan or refuse the loan because he/she has been a person who was convicted for a crime? Is the person eligible to obtain a loan?

Figure 2. Visualisation of scenario where no opt-out option has been implemented



In this scenario, the presence of the personal data of civilian X in the LLM-from-scratch has a negative effect on acquiring a loan.[185] **Figure 3** analyses how the events in the scenario are challenging the tenets of the Rule of Law.

Figure 3. Analysis Of Scenario One: No Opt-Out Option Implemented

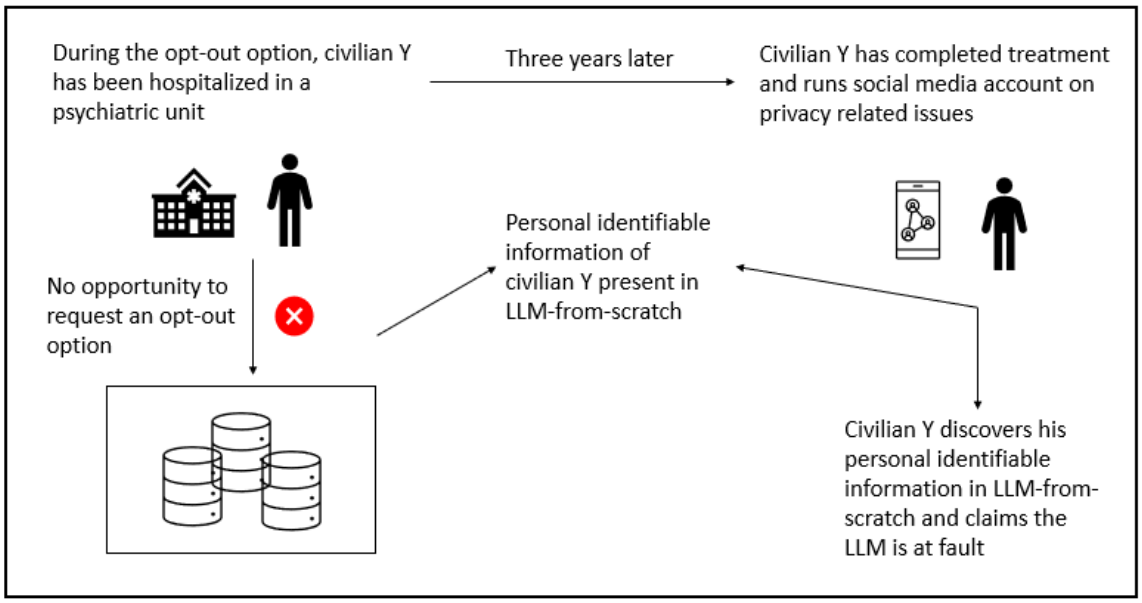
Tenet	
Accountability	When it comes to the tenet **accountability**, the bank has to make decisions which may sort legal effects that are based on (juridical) facts. When this decision, as stated in the scenario description, is made without a legal basis this can lead to unjust situations such as not approving a loan which -if there is no other ground on which this disapproval stands- is unjust. The bank's decision to not grant a loan to the ex-convicted must be based on a legal basis. If this is not the case, it is illegal which is contradictory to the tenet of legitimacy.
Transparency	It can be questioned if the output given by the LLM-from-scratch provides factual information and whether it is **transparent** how the data is collected, curated and trained. The decision to make use of data of court cases is an important decision in itself whose result should be clearly communicated to the general public to mitigate risks of people being unaware that their court cases could be included within the model.
Contestability	**Contestability** implies that the ex-convicted should be given the possibility to contest the decision which is made by the bank, also in front of court. Therefore, it is of importance that, if the bank uses the LLM-from-scratch to base their decision upon, the bank is transparent about the fact it is using a LLM and also provides context on how it has been used. For example, what is the prompt that they used?
Liability	If the LLM-from-scratch is not transparent about the fact that it has used court cases to train the model and the bank is unaware that this type of data has been incorporated, the tenet **liability** is challenged. If the bank does not know that court cases have been included, the question raises if they can be held liable for errors, mistakes or damages that are consequences of their (wrongful) decisions. Or is the LLM-from-scratch liable?

[185] Other areas where a negative impact may be found are being rejected for a job application since you are being a family member of a (ex)-criminal or not getting a rental apartment/mortgage because you are connected to a (ex)-convicted killer.

24.2.1.2. Scenario Two: The Opt-Out Option Is Implemented

Figure 4 visualizes the scenario where an opt-out option has been implemented. We focus on civilian Y, who has been hospitalized in a psychiatric unit during the opt-out option and therefore did not have the opportunity to request an opt-out option. Three years later, civilian Y has completed treatment and uses social media to create awareness on privacy issues. Civilian Y discovers that his personal identifiable information is present in the LLM and decides to claim the LLM team. The question raises if the team developing the LLM is responsible for removing personal information of persons who did not have had the opportunity (e.g. due to mental health issues, age or disabilities) to make use of the opt-out option.

Figure 4.Visualisation of scenario where an opt-out option has been implemented



Regardless of the answer to the question whether the LLM team is responsible, the tenets of the Rule of Law are challenged as reflected in **Figure 5**.

Figure 5. Analysis Of Scenario Two: Opt-Out Option Is Implemented

Tenet	
Transparency	It is important to create awareness and **transparency** amongst both civilians and family members or caretakers of incapacitated persons about the possibilities of the opt-out option. A possible solution to deal with the situation of incapacitated persons could be to provide an indirect opt-out option to family members or caretakers of the incapacitated persons. The family member or caretaker can access the database for the incapacitated person and request an opt-out option in his/her name. The opt-out option is thus requested indirectly via relatives of the incapacitated person. However, how can it be ensured that the incapacitated person gives consent to the family members or caretakers to perform such an action? If an opt-out option is implemented, the team building the LLM-from-scratch needs to carefully consider how the opt-out option should be implemented and presented to the public. **Transparency** plays an important role in this. It is important that LLM-from-scratch is transparent about all the architectural choices made, including the decisions made on the implementation of the opt-out option. Furthermore, policies and communication strategies can be used to inform the public about the opt-out option. The logging of the architectural choices including reasoning why each choice is made gives LLM-from-scratch the opportunity to inform the public and meet their expectations.
Accountability and liability	Transparency in the architectural choices also contributes to the other tenets of the Rule of Law: a transparent way of working is necessary to give the powers (legislative, executive and judicial power), who have legitimatized decision power the ability to check if the architectural decisions are **legitimate**. This is also important when you take **accountability** questions into account: without the possibility to address responsibility about the architectural choices made, the accountability cannot be addressed. This can lead to problems when errors occur which raises **liability** issues and it becomes important to contest if the architectural choice made have led to these errors and therewith liabilities. Another important question is linked to being future proof: how can it be ensured that the opt-out option is made future proof as in e.g. compliance with the AI Act and related legislation such as the GDPR and connected jurisprudence? It is currently not a legal obligation to provide an opt-out option to the public before training a LLM. Moreover, if one decides to implement an opt-out option, no guidelines are given on how to do this. Beyond current questions connected to how compliance is conceived at this moment and time, it can be helpful to take the tenets of the Rule of Law when shaping the architecture choices, such as the opt-out option, to strive for future proof compliancy. It is therefore important to closely monitor the AI Act and related legislation for possible changes.

CONCLUSION

In this paper we have investigated how the design choice of incorporating an opt-out option during the development of a hypothetical LLM, LLM-from-scratch, can contribute to a fair deployment of the LLM and how the tenets of the Rule of Law may be challenged by this decision. The scenario analysis aids in making an informed decision whether an opt-out option should be included to ensure a fair and just deployment of an LLM. The scenario analysis has resulted in the following recommendations:

In the case that an opt-out option has not been implemented:

- If court cases or other sensitive data are included in the training phase of an LLM, this should be made transparent and clearly communicated to the users of that LLM. As the public did not receive an opt-out option, the curated data still (potentially) contains personal identifiable information which may cause negative outcomes for individuals in society.
- The developers of the LLM should be aware that claims could follow from individuals who feel discriminated or disadvantaged by the LLM.

In the case that an opt-out option has been implemented:

- Providing an opt-out option to the public before training the model can reduce the possible chance of injustice in legal effects. This option can also provide a form of human oversight via

the check of the public on the dataset before training. It can even be seen as a form of citizen participation.

- Access to the internet is an important enabler to make a check on the data and opt-out option possible. If the data contains personal information of people who have less accessibility to the internet, the data could be (double) checked by independent institutions to foster the well-being of the global majority.
- It is important that the LLM developers investigate how to deal with incapacitated persons in the opt-out option. For example, by exploring the possibility to provide an indirect opt-out option to family members or caretakers of incapacitated persons. The design of such an indirect opt-out option should be carefully thought through and can be quite challenging since incapacitated persons may be unable to give their consent.
- Developers of LLMs should be aware that, in the case that they do not provide a special opt-out option for incapacitated persons, they can receive claims from persons who, at the time of the opt-out option, were incapacitated.
- The workings of the opt-out option should be transparent. In order to achieve this, the design choices made regarding the opt-out option and the implementation of the opt-out option should be clearly logged and documented.

References

Bradshaw, A. C. (2001). Internet users worldwide. Educational Technology Research and Development, 111-117.
Dasgupta, R., Ganesan, B., Kannan, A., Reinwald, B., & Kumar, A. (2018). Fine grained classification of personal data entities. arXiv preprint arXiv:1811.09368.
European Parliament, 13th of March 2024 - 12:25, Official Press Release, Artificial Intelligence Act: MEPs adopt landmark law.
Janssens, L., Lucassen, O., Middeldorp, L., Lobbezoo, L., & Schoenmakers, O. (July 2024). Responsible AI and the Rule of Law. TNO.
NATO. Retrieved August 2023, accessed at https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html.
Stein, R. A. (2019). What exactly is the rule of law. Hous. L. Rev., 57, 185.
United Nations, (2024), United Nations and the Rule of Law. What is the Rule of Law, Retrieved October 2024, from https://www.un.org/ruleoflaw/what-is-the-rule-of-law/#:~:text=It%20requires%20measures%20to%20ensure,and%20procedural%20and%20legal%20transparency.

Legislation

Article 2 of the Treaty on European Union: *"The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail."*
AI Act Draft Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))

Appendix A

List of Key Definitions

AI	Artificial Intelligence

AI Act	European legislation on harmonised rules on Artificial Intelligence
EDT	Emerging Disruptive Technology
LLM	Large Language Model
NATO	North Atlantic Treaty Organization
OSINT	Open Source Intelligence
Requirements	Technical requirements which can become tools of good governance Rule of Law. Is shaped by various sources, such as: case law; legal doctrine; legal interpretation methods; positive law; rules and regulations; draft rules and regulations and legal theory.
AI Technology	AI technologies and/or systems with applied AI applications
Rule of Law tenets	accountability, transparency, contestability mechanisms, processes of rules and regulations; case law; policies; etc.

25. Addressing Gender Data Gaps in the Global Majority: Opportunities and Challenges of Synthetic Data

Ronald Musizvingoza

Abstract. This paper explores pervasive gender data gaps affecting the global majority, highlighting their negative impact on health, particularly among women and girls. When these data gaps persist, the rapid use of AI can exacerbate existing inequalities by failing to fully incorporate the global majority's experiences. We argue that synthetic data can be a powerful tool for addressing these gaps by generating representative datasets that reflect diverse gender experiences. While acknowledging the risks associated with synthetic data, including potential biases and cybersecurity threats, this paper emphasises the need for robust methodologies, ethical frameworks and guidelines to ensure its responsible use. By integrating real-world data and fostering collaboration with gender experts, we advocate for a multifaceted approach to AI development that prioritises gender equality. Ultimately, we call for policies that promote inclusive research and data practices, ensuring synthetic data contributes to equitable health outcomes for the global majority.
Keywords. AI, Synthetic Data, Gender Data Gaps, Health, Global Majority

Introduction

Gender data gaps, or the lack of data on diverse gender experiences, are widespread and disproportionately impact people from the global majority(Musizvingoza & Lopes, 2022). Women comprise half the world's population, but data on their status, health, and well-being is lacking. Closing these gaps is crucial to reflect the global majority's experiences and needs, especially within artificial intelligence(AI)(Musizvingoza, 2024). Despite global commitments to Sustainable Development Goal 5, only 48% of the necessary data to assess progress is available. With a 3% annual growth rate, collecting all required gender-specific data will take 22 years, missing the 2030 deadline by over a decade(Encarnacion, Emandi, & Seck, 2022).

Gender data gaps perpetuate inequalities in education, work, and healthcare by failing to support effective programs and overlooking marginalised groups(Paris21, 2024). AI tools can worsen this issue by embedding and amplifying gender biases if not trained on comprehensive, representative data(O'Connor & Liu, 2023). Since these tools learn from their training data, excluding or misrepresenting the global majority, especially women can have serious consequences especially in healthcare. For example, AI tools for liver disease screening were found to be less accurate for women(Straw & Wu, 2022), and delays in diagnoses for Black patients have been linked to biased datasets(Williams, 2023). Additionally, AI in judicial sentencing were found to be discriminatory towards black offenders(Lippert-Rasmussen, 2022).

One potential solution to address gender data gaps is using artificially generated synthetic data to mimic the original dataset's characteristics or meet predetermined criteria(Deng, 2023; Marwala, Fournier-Tombs, & Stinckwich, 2023). By simulating the properties of original datasets, synthetic data can be pivotal for training AI tools, especially in contexts where data is sensitive, scarce, or biased. This paper examines the risks and opportunities associated with using synthetic data to address gender data gaps in health, particularly from the perspective of global majority populations frequently underrepresented in AI and data governance discussions. It will provide insights into how synthetic data can address gender data gaps and enable gender and health equity.

25.1. Discussion

25.1.1. Gender Data Gaps

Gender data gaps are globally prevalent, particularly in developing countries and regions such as Southeast Asia, Latin America, Sub-Saharan Africa, and the Pacific, impacting millions of vulnerable women and girls(Kathleen Grantham, 2020). Despite commitments from 193 countries to the 2030 Agenda, comprehensive data on gender-specific SDG indicators is still lacking(Encarnacion et al., 2022). These gaps are especially pronounced in healthcare, with only 21.8% of gender health indicators available in 2023(World Bank, 2024). This lack of data hampers effective public health responses, particularly in the global majority, especially in African countries(Adebisi & Lucero-Prisno, 2022). For example, during the COVID-19 pandemic, 76% of high-income countries reported COVID-19 case data by sex, compared to only 37% of low-income countries (Hawkes et al., 2021).

AI development and decision-making are controlled mainly by the Global North, particularly North America(Anthony, Sharma, & Noor, 2024). Nevertheless, the impact of AI on the global majority is substantial(Norori, Hu, Aellen, Faraci, & Tzovara, 2021). AI models are often trained on data generated online, which excludes experiences from connectivity-limited environments, making the models unrepresentative of the global context. With 2.6 billion people offline, representing 37% of the global population, this issue is further exacerbated in developing countries, where 96% of those offline reside and where, on average, 21% of women have internet access compared to 32% of men(ITU, 2022).

Moreover, AI models trained on biased data amplify gender bias. Medical studies have historically excluded female participants, leading to research data collected from males being generalised to females(Merone, Tsey, Russell, & Nagle, 2022). A significant portion of the datasets used for training AI algorithms in healthcare is derived from such biased research, resulting in persistent gender bias. This gap has far-reaching implications for healthcare, particularly in disease prevention, diagnosis, and treatment(di Lego, 2023; Norori et al., 2021). To address these issues, the World Health Organization (WHO) AI for health guidance(World Health Organization, 2021) highlights the importance of ethical, legal, and human rights considerations, focusing on data governance, algorithmic transparency, inclusiveness, equity, and accountability(Lopes, Saitabau, Rustagi, & Khosla, 2023).

25.1.2. Synthetic Data

Synthetic data can address imbalances and underrepresentation, helping fill gender data gaps in AI model training(Deng, 2023). It helps overcome data scarcity, sensitivity, and bias challenges by providing a flexible and safe alternative to real-world data(Deng, 2023). For example, in healthcare, synthetic data is used as a proxy for real data to support medical research while ensuring confidentiality(Giuffrè & Shung, 2023; Gonzales, Guruswamy, & Smith, 2023; Kokosi & Harron, 2022; Laderas et al., 2017; Reiner Benaim et al., 2020). Other notable uses of synthetic data include estimating the benefits of healthcare policies and interventions, pre-training models for specific patient populations, and improving public health models for predicting disease outbreaks (Giuffrè & Shung, 2023; Gonzales et al., 2023; Kokosi & Harron, 2022; Laderas et al., 2017; Reiner Benaim et al., 2020). Furthermore, synthetic data supports the creation of digital twins, simulating real-time behaviour, including gender-specific health patterns(Giuffrè & Shung, 2023).

Synthetic data's key features—privacy preservation(Gonzales et al., 2023; James, Harbron, Branson, & Sundler, 2021; Tiwald, Ebert, & Soukup, 2021), scalability(Almirall et al., 2022), realistic

generation(Dahmen & Cook, 2019), representativeness(James et al., 2021; Tiwald et al., 2021), and reproducibility—are critical in addressing gender data gaps. Synthetic data helps balance datasets, augment limited data, and generate high-dimensional data, improving reliability for accurate gender analysis(Juwara, El-Hussuna, & El Emam, 2024). The UZIMA-DS project in Kenya exemplifies the use of AI-ready synthetic datasets to create early warning systems, addressing data gaps while promoting open access in health research(Thuku, Baker, Mwigereri, Waljee, & Siwo, 2024). Another example is the World Bank's Synthetic Data for an Imaginary Country, a hierarchical simulation and training dataset covering demographic, education, and health variables(World Bank, 2023). These examples highlight how synthetic data can be used to close gender data gaps in health by enhancing the representation of underrepresented groups, improving access to gender-specific data, enabling more accurate simulations, and facilitating broader sharing of gender-sensitive information, particularly in resource-limited settings, leading to more equitable health interventions and research. While synthetic data offers opportunities for training AI models, it risks oversimplifying complex human experiences, perpetuating biases, and neglecting the realities of the global majority, underscoring the need for ethically grounded approaches that integrate real-world data.

25.1.3. Methods for Generating Synthetic Data

Methods for synthetic data generation can be categorised into statistical and probabilistic approaches, machine learning techniques(ML), and deep learning methodologies(DL) (Hernandez, Epelde, Alberdi, Cilla, & Rankin, 2022). Statistical methods generate synthetic data by sampling from existing datasets(Kaur et al., 2021; Pourshahrokhi, Kouchaki, Kober, Miaskowski, & Barnaghi, 2021). ML approaches use models like decision trees and regression to generate new data points that mimic the statistical properties of the original data, especially when real data is scarce or sensitive(Gonzales et al., 2023; Lu et al., 2023). DL-based methods use neural networks to generate synthetic data(Achuthan et al., 2022; Mohamed & Frank, 2024; Nikolentzos, Vazirgiannis, Xypolopoulos, Lingman, & Brandt, 2023). In healthcare, synthetic data generation techniques produce valuable datasets, including synthetic patient records for research and analysis(Nikolentzos et al., 2023), synthetic time-series health records to capture dynamic patient health trajectories(Li, Cairns, Li, & Zhu, 2023), realistic medical images such as MRI and CT scans for model training(Skandarani, Jodoin, & Lalande, 2023), and simulations of imbalanced clinical variables in HIV antiretroviral therapy datasets(Giuffrè & Shung, 2023; Kuo et al., 2023). These datasets can help address gender data gaps by providing more comprehensive and representative data for gender analysis, enabling more inclusive and effective healthcare solutions.

These synthetic data generation techniques can address gender data gaps by reflecting real-world diversity, even when actual data is limited or biased. They enhance fairness, equity, privacy, and inclusivity—key elements of gender data—while improving representation, intersectionality, and bias mitigation in underrepresented groups, thus fostering more accurate gender analysis. For instance, DL-based methods can generate synthetic data that closely mimics real-world diversity(Ali et al., 2022) and balance gender-skewed datasets for more representative outcomes(Makhlouf, Maayah, Abughanam, & Catal, 2023). Statistical and probabilistic approaches can be used to enhance data privacy — a critical aspect of gender data while maintaining the statistical properties of original datasets and protecting sensitive information(Skandarani et al., 2023). When gender-specific data is scarce, synthetic data can augment the dataset, providing richer insights and improving analysis for underrepresented genders(Motamed, Rogalla, & Khalvati, 2021; Yu et al., 2020). Additionally, synthetic data can create high-dimensional datasets that capture complex relationships among variables,

including gender, enabling sophisticated analyses(Sun, van Soest, & Dumontier, 2023). These techniques can also identify and mitigate biases in real-world datasets by balancing underrepresented gender categories and promoting equitable representation(Paproki, Salvado, & Fookes, 2024; van Breugel, Kyono, Berrevoets, & van der Schaar, 2021). Ensuring privacy, fairness, and equity is crucial for comprehensive gender data. Synthetic data can support these principles by generating diverse datasets that reflect real-world complexities, ensuring demographic parity, and improving the accuracy and fairness of analyses and decisions(Rajabi & Garibay, 2021).

25.1.4. CHALLENGES

Synthetic data in healthcare poses challenges, with risks of bias amplification, low interpretability, and insufficient methods for auditing data quality(Giuffrè & Shung, 2023). Furthermore, synthetic data entails multifaceted risks, including cybersecurity threats, model inaccuracies, data integrity, misuse, intellectual property infringement, and data contamination, which can exacerbate gender data gaps if not adequately addressed(Marwala et al., 2023). For example, cybersecurity threats could lead to the exposure of sensitive data related to underrepresented gender groups. At the same time, model inaccuracies might perpetuate existing biases by generating flawed or incomplete gender data. Data misuse and contamination can distort gender representation in datasets, further skewing analysis. Additionally, intellectual property infringement could limit access to diverse datasets, thus hindering efforts to ensure fairness, inclusivity, and equitable representation of gender data. For instance, a synthetic dataset based on facial images of predominantly men or a specific racial group will reflect this imbalance if not addressed, perpetuating gender biases(Hao et al., 2024).

The success of synthetic data in healthcare depends on ensuring diversity, transparency, bias mitigation, privacy, fairness, and robust evaluation to advance AI responsibly and represent underrepresented genders(Gonzales et al., 2023). Countries like Singapore have developed guidelines to harness synthetic data's potential while balancing utility and protection risks by defining its purpose, preparing data thoughtfully, following best practices, and managing re-identification risks(Personal Data Protection Comission, 2024). Similarly, the United Nations University recommends ethically using synthetic data in AI, emphasising diverse data sources, various generative models, transparency, and quality metrics(de Wilde et al., 2023). These guidelines help close gender data gaps and promote inclusive AI solutions by ensuring synthetic data is designed to benefit diverse populations, particularly in the global majority.

CONCLUSIONS

In conclusion, while synthetic data offers significant promise for addressing gender data gaps and promoting equitable AI, its application must be cautious due to risks like bias amplification and misuse. The methods for generating synthetic data provide potential solutions by enabling the creation of more representative datasets and mitigating biases that disproportionately affect underrepresented groups, particularly the global majority. Since AI reflects real-world gender biases, addressing these is key for equitable AI. We recommend comprehensive gender data collection, inclusive definitions, well-trained data collectors, collaboration with gender experts, and adopting ethical guidelines and best practices widely recognised in data governance to ensure the responsible use of synthetic data. Furthermore, we recommend supplementing synthetic datasets with real-world gender data to ensure a more accurate and inclusive portrayal of people from the global majority. Prioritising gender equality in AI development, evaluating data for bias, diversifying teams, and enforcing ethical guidelines will mitigate

potential biases. The Global Digital Compact is a critical opportunity to embed gender perspectives into digital governance. Without such efforts, AI may widen existing gender gaps. Future research should refine synthetic data techniques, develop auditing frameworks, and address socioeconomic factors to capture gender disparities. Policy recommendations should prioritise the implementation of ethical guidelines for synthetic data usage, foster transparency in data practices, and promote inclusive research agendas that consider the needs and experiences of the global majority to ensure that synthetic data contributes to meaningful and equitable outcomes.

References

Achuthan, S., Chatterjee, R., Kotnala, S., Mohanty, A., Bhattacharya, S., Salgia, R., & Kulkarni, P. (2022). Leveraging deep learning algorithms for synthetic data generation to design and analyze biological networks. *J. Biosci.*, *47*.

Adebisi, Y. A., & Lucero-Prisno, D. E. 3rd. (2022). Fixing Data Gaps for Population Health in Africa: An Urgent Need. *Int. J. Public Health*, *67*, 1605418. https://doi.org/10.3389/ijph.2022.1605418

Ali, H., Biswas, Md. R., Mohsen, F., Shah, U., Alamgir, A., Mousa, O., & Shah, Z. (2022). The role of generative adversarial networks in brain MRI: a scoping review. *Insights Imaging*, *13*(1), 98. https://doi.org/10.1186/s13244-022-01237-0

Almirall, E., Callegaro, D., Bruins, P., Santamaría, M., Martínez, P., & Cortés, U. (2022). *The use of Synthetic Data to solve the scalability and data availability problems in Smart City Digital Twins*. Retrieved from http://arxiv.org/abs/2207.02953

Anthony, A., Sharma, L., & Noor, E. (2024). *Advancing a More Global Agenda for Trustworthy Artificial Intelligence*. Carnegie Endowment for International Peace.

Dahmen, J., & Cook, D. (2019). SynSys: A Synthetic Data Generation System for Healthcare Applications. *Sensors*, *19*(5). https://doi.org/10.3390/s19051181

de Wilde, P., Arora, P., Buarque, F., Chin, Y. C., Thinyane, M., Stinckwich, S., … Marwala, T. (2023). *Recommendations on the Use of Synthetic Data to Train AI Models*. United Nations University.

Deng, H. (2023). *Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems: A Primer*. Geneva, Switzerland.: UNIDIR.

di Lego, V. (2023). Uncovering the gender health data gap. *Cad. Saude Publica*, *39*(7), e00065423. https://doi.org/10.1590/0102-311XEN065423

Encarnacion, J., Emandi, R., & Seck, P. (2022). It will take 22 years to close SDG gender data gaps. Retrieved from https://data.unwomen.org/features/it-will-take-22-years-close-sdg-gender-data-gaps

Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digit. Med.*, *6*(1), 186. https://doi.org/10.1038/s41746-023-00927-3

Gonzales, A., Guruswamy, G., & Smith, S. R. (2023). Synthetic data in health care: A narrative review. *PLOS Digit. Heal.*, *2*(1), e0000082. https://doi.org/10.1371/journal.pdig.0000082

Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., … Tang, H. (2024). *Synthetic Data in AI: Challenges, Applications, and Ethical Implications*. Retrieved from http://arxiv.org/abs/2401.01629

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, *493*, 28–45. https://doi.org/10.1016/j.neucom.2022.04.053

ITU. (2022). *The State of Broadband 2022: Accelerating broadband for new realities*.

James, S., Harbron, C., Branson, J., & Sundler, M. (2021). Synthetic data use: Exploring use cases to optimise data utility. *Discov. Artif. Intell.*, *1*(1), 15. https://doi.org/10.1007/s44163-021-00016-y

Juwara, L., El-Hussuna, A., & El Emam, K. (2024). An evaluation of synthetic data augmentation for mitigating covariate bias in health data. *Patterns (New York, N.Y.)*, *5*(4), 100946. https://doi.org/10.1016/j.patter.2024.100946

Kathleen Grantham. (2020). *Mapping Gender Data Gaps: An SDG Era Update*. (March). Retrieved from https://data2x.org/resource-center/mappinggenderdatagaps/

Kaur, D., Sobiesk, M., Patil, S., Liu, J., Bhagat, P., Gupta, A., & Markuzon, N. (2021). Application of Bayesian networks to generate synthetic health data. *J. Am. Med. Informatics Assoc.*, *28*(4), 801–811. https://doi.org/10.1093/jamia/ocaa303

Kokosi, T., & Harron, K. (2022). Synthetic data in medical research. *BMJ Med.*, *1*(1), e000167. https://doi.org/10.1136/bmjmed-2022-000167

Kuo, N. I.-H., Garcia, F., Sönnerborg, A., Böhm, M., Kaiser, R., Zazzi, M., … Barbieri, S. (2023). Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. *J. Biomed. Inform.*, *144*, 104436. https://doi.org/10.1016/j.jbi.2023.104436

Laderas, T., Vasilevsky, N., Pederson, B., Haendel, M., McWeeney, S., & Dorr, D. A. (2017). Teaching data science fundamentals through realistic synthetic clinical cardiovascular data. *BioRxiv*. https://doi.org/10.1101/232611

Li, J., Cairns, B. J., Li, J., & Zhu, T. (2023). Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digit. Med.*, *6*(1), 98. https://doi.org/10.1038/s41746-023-00834-7

Lippert-Rasmussen, K. (2022). Algorithm-Based Sentencing and Discrimination. *Sentencing Artif. Intell.*, 0. https://doi.org/10.1093/oso/9780197539538.003.0005

Lopes, C. A., Saitabau, A., Rustagi, N., & Khosla, R. (2023). A digital health governance agenda for sexual and reproductive health and rights. *Sex. Reprod. Heal. Matters*, *31*(4), 1–6. https://doi.org/10.1080/26410397.2024.2372865

Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., & Wei, W. (2023). *Machine Learning for Synthetic Data Generation: A Review*. *14*(8), 1–19.

Makhlouf, A., Maayah, M., Abughanam, N., & Catal, C. (2023). The use of generative adversarial networks in medical image augmentation. *Neural Comput. Appl.*, *35*(34), 24055–24068. https://doi.org/10.1007/s00521-023-09100-z

Marwala, T., Fournier-Tombs, E., & Stinckwich, S. (2023). *The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development*. (1), 1–11.

Merone, L., Tsey, K., Russell, D., & Nagle, C. (2022). Sex Inequalities in Medical Research: A Systematic Scoping Review of the Literature. *Women's Heal. Reports (New Rochelle, N.Y.)*, *3*(1), 49–59. https://doi.org/10.1089/whr.2021.0083

Mohamed, S., & Frank, L. (2024). *Generative Adversarial Networks (GANs) for Synthetic Test Data*. (June).

Motamed, S., Rogalla, P., & Khalvati, F. (2021). Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images. *Informatics Med. Unlocked*, *27*, 100779. https://doi.org/10.1016/j.imu.2021.100779

Musizvingoza, R. (2024). Bridging the Gender Data Gap: Harnessing Synthetic Data for Inclusive AI. Retrieved from https://unu.edu/macau/blog-post/bridging-gender-data-gap-harnessing-synthetic-data-inclusive-ai

Musizvingoza, R., & Lopes, C. A. (2022). *Limited gender data deepens inequalities*. https://doi.org/10.54377/916A-61EC

Nikolentzos, G., Vazirgiannis, M., Xypolopoulos, C., Lingman, M., & Brandt, E. G. (2023). Synthetic electronic health records generated with variational graph autoencoders. *Npj Digit. Med.*, *6*(1), 83. https://doi.org/10.1038/s41746-023-00822-x

Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

O'Connor, S., & Liu, H. (2023). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI Soc.* https://doi.org/10.1007/s00146-023-01675-4

Paproki, A., Salvado, O., & Fookes, C. (2024). Synthetic Data for Deep Learning in Computer Vision & Medical Imaging: A Means to Reduce Data Bias. *ACM Comput. Surv.*, *56*(11). https://doi.org/10.1145/3663759

Paris21. (2024). *Improving co-ordination to move the gender equality agenda forward in Africa*. Retrieved from https://www.paris21.org/news/improving-co-ordination-move-gender-equality-agenda-forward-africa

Personal Data Protection Comission. (2024). *Privacy Enhancing Technology (PET): Proposed Guide On Synthetic Data* (No. 1; pp. 1–42). Singapore: Personal Data Protection Comission-Singapore.

Pourshahrokhi, N., Kouchaki, S., Kober, K. M., Miaskowski, C., & Barnaghi, P. (2021). *A Hamiltonian Monte Carlo Model for Imputation and Augmentation of Healthcare Data*. 1–9.

Rajabi, A., & Garibay, O. O. (2021). Towards Fairness in AI: Addressing Bias in Data Using GANs. In C. Stephanidis, M. Kurosu, J. Y. C. Chen, G. Fragomeni, N. Streitz, S. Konomi, … S. Ntoa (Eds.), *HCI Int. 2021—Late Break. Pap. Multimodality, Ext. Reality, Artif. Intell.* (pp. 509–518). Cham: Springer International Publishing.

Reiner Benaim, A., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashiach, T., … Beyar, R. (2020). Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med Inf.*, *8*(2), e16492. https://doi.org/10.2196/16492

Skandarani, Y., Jodoin, P.-M., & Lalande, A. (2023). GANs for Medical Image Synthesis: An Empirical Study. *J. Imaging*, *9*(3). https://doi.org/10.3390/jimaging9030069

Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Heal. Care Informatics*, *29*(1). https://doi.org/10.1136/bmjhci-2021-100457

Sun, C., van Soest, J., & Dumontier, M. (2023). Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *J. Biomed. Inform.*, *143*, 104404. https://doi.org/10.1016/j.jbi.2023.104404

Thuku, N., Baker, J. A., Mwigereri, D. G., Waljee, A. K., & Siwo, G. (2024). UZIMA-DS AI-Ready Synthetic Data.

Tiwald, P., Ebert, A., & Soukup, D. T. (2021). *Representative & Fair Synthetic Data*. (2002), 1–5.

van Breugel, B., Kyono, T., Berrevoets, J., & van der Schaar, M. (2021). DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Adv. Neural Inf. Process. Syst.* (Vol. 34, pp. 22221–22233). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper{_}files/paper/2021/file/ba9fab001f67381e56e410575874d967-Paper.pdf

Williams, P. (2023). Retaining Race in Chronic Kidney Disease Diagnosis and Treatment. *Cureus*, *15*(9), e45054. https://doi.org/10.7759/cureus.45054

World Bank. (2023). World—Synthetic Data for an Imaginary Country, Sample, 2023 A synthetic hierarchical dataset for simulation and training purposes. World Bank.

World Bank. (2024). *Data Availability | World Bank Gender Data Portal*. Retrieved from https://genderdata.worldbank.org/en/data-availability?indicator=SH.DTH.NCOM.ZS{\&}year-bucket=0{\&}country=LUX{\&}topic-year-bucket=0

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance* (pp. 3–22). https://doi.org/10.1142/9789811238819_0001

Yu, N., Li, K., Zhou, P., Malik, J., Davis, L., & Fritz, M. (2020). Inclusive GAN: Improving Data and Minority Coverage in Generative Models. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Comput. Vis. – ECCV 2020* (pp. 377–393). Cham: Springer International Publishing.