Track 8 – AI and Ethics

(May Thorseth, NTNU, Trondheim, Norway)

Thursday 29th 8:30-10:30 – Chair: May Thorseth, NTNU, Trondheim, Norway.

Abstracts

Digitalization of Powerlines Inspection Routine Using AI As an Infrastructuring Process

Zeina Othman[1], Silvia Bruzzone[1]
[1]Mälarldalen University, Västerås, Sweden

The transition to sustainable energy and the electrification of the economy necessitates the robustness of systems and activities. Maintaining powerline infrastructures as a vital component of the energy grid became increasingly critical to prevent any failure and enhance its resilience (Mistra Electrification 2021). Traditionally, powerline inspections were conducted manually, employing foot patrols or helicopter-assisted surveys, methods known for being slow, costly, and potentially hazardous (Nguyen et al., 2018). However, the trend towards greater digitalization and integration of artificial intelligence (AI) and algorithms has revolutionized various sectors, challenging traditional manual approaches. In the energy sector, advancements in AI and data analysis techniques have led to automated virtual powerline inspection possibilities (Nguyen et al., 2018). Specifically, AI-powered software and drones enable more efficient and sustainable inspections of powerlines, reducing dependence on helicopters and foot patrols. Nevertheless, challenges, such as coordinating multiple actors, reconfiguring work practices, data availability, and other related challenges accompany the adoption of AI in powerline inspections.

Infrastructures in modern societies tend to be taken for granted, invisible despite the fact they are central to our societies. In this sense, powerlines are infrastructures par excellence, as they are systems our increasingly more electrified societies rely on to make our lives and economies work. But these infrastructures – powerlines - are embedded in wider and more complex organizational systems, work practices, routines, standards, other materials, etc. Studying infrastructuring processes means giving an account of the invisible work, to what is taken for granted, and the vulnerability of things (Denis & Pontille, 2020). It also means to look at the role of materiality in producing digital information as well as to focus on power relations and ethical choices, and to convey relational and dynamic work that is necessary to make infrastructure work (Bowker & Star, 2000). In this contribution, we propose to look particularly at the process of digitalizing the powerlines' inspection routine using drones and AI algorithmic technologies as a process of "infrastructuring". This means focusing on the hidden work - and neglected things (de la Bellacasa, 2010) that sustain a service, a

technology, or an innovation – in this case, the digital virtual inspection work of powerlines, which entails an ample amount of hidden, invisible, and manual work to make the AI algorithmic software work in the intended way. Moreover, we believe that mobilizing the term "infrastructuring" would help us in terms of explaining the gap between AI algorithmic technologies and their usability and how they can shape the virtual inspection process.

In this paper, we focus on a case study that covers the efforts of two companies, an AI software company located in Northern Europe and a utility company in Southern Europe, who are both engaged in a large AI digitalization project to shift from manual to virtual inspections, using a collaborative AI software for analyzing images of powerlines captured by drones. Besides handling the virtual inspection of the customer's powerlines, the project also entails the development of additional new customized AI algorithmic models. By taking a Routine Dynamics lens, we view the powerlines inspection in utilities as a prototypical routine that features multiple actors engaging in repetitive, interconnected actions that form patterns of action (Feldman, 2015). We also adopt the Routine's Dynamic definition of algorithms as a type of organizational artifact intertwined with other organizational elements in a broader network of actors, practices, and theories in dynamic sociomaterial assemblages, that can have a substantial impact on organizational routines (D'Adderio, 2008, 2011; Glaser, Pollock, et al., 2021; Glaser, Valadao, et al., 2021). Moreover, the aim of the project which provides the empirical case study for this paper is to intentionally design and change the inspection routines using drones and algorithmic artifacts in the form of AI software, thus, we see this case for digitalizing and virtualizing the inspection process as a case of routine design (Wegener & Glaser, 2021), and we borrow the above-outlined concept of "infrastructuring" from STS studies and apply it to this case.

The new configuration implies the design and introduction of revamped organizational routines and practices involving multiple actors, located in a globally distributed setting, technologies, and other materials. Our question is how AI reconfigures the work of powerlines inspections? What is the digital inspection infrastructure about and what does it imply in terms of knowledge production? Upon collecting and analyzing the qualitative data consisting of 38 observations and 13 semi-structured interviews, we propose the adoption of virtual inspections, a process of re-infrastructuring from manual to digital virtual process supported by AI. Our finding indicates that designing the inspection routine and practice using AI algorithmic technology requires rethinking and considering the infrastructural relationship between all human and non-human actors, including human actors, standards, rules, space, time, and material technologies are connected visibly and invisibly in a complex infrastructure assemblage that is entangled sociomaterially and relationally.

References

Bowker, G. C., & Star, S. L. (2000). Sorting Things Out. The MIT Press.

D'Adderio, L. (2008). The Performativity of Routines: Theorising the Influence of Artefacts and Distributed Agencies on Routines Dynamics. Research Policy, 37, 769–789. https://doi.org/10.2139/ssrn.1309622

D'Adderio, L. (2011). Artifacts at the centre of routines: Performing the material turn in routines theory. Journal of Institutional Economics, 7, 197–230. https://doi.org/10.1017/S174413741000024X

Denis, J., & Pontille, D. (2020). Maintenance epistemology and public order: Removing graffiti in Paris. Social Studies of Science, 51(2), 233–258.

de la Bellacasa, M. P. (2010). Matters of care in technoscience: Assembling neglected things. Social Studies of Science, 41(1), 85–106.

Feldman, M. S. (2015). Theory of routine dynamics and connections to strategy as practice. In Cambridge Handbook of Strategy as Practice (pp. 317–330). Cambridge University Press. https://doi.org/10.1017/CBO9781139681032.019

Glaser, V. L., Pollock, N., & D'Adderio, L. (2021). The Biography of an Algorithm: Performing algorithmic technologies in organizations. Organization Theory, 2(2), 263178772110046. https://doi.org/10.1177/26317877211004609

Glaser, V. L., Valadao, R., & Hannigan, T. R. (2021). Algorithms and Routine Dynamics. In Cambridge Handbook of Routine Dynamics (pp. 315–328). Cambridge University Press. https://doi.org/10.1017/9781108993340.027

Jarrahi, M. H. (2018). Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making. Business Horizons, 61.

Mistra Electrification 2021. (n.d.). Retrieved January 9, 2024, from https://mistraelectrification.com/about-the-program/

Nitzberg, M., & Zysman, J. (2022). Algorithms, data, and platforms: the diverse challenges of governing AI. Journal of European Public Policy, 29(11), 1753–1778.

Nguyen, V. N., Jenssen, R., & Roverso, D. (2018). Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning. International Journal of Electrical Power & Energy Systems, 99, 107–120.

Wegener, F. E., & Glaser, V. L. (2021). Design and Routine Dynamics. In Cambridge Handbook of Routine Dynamics (pp. 301–314). Cambridge University Press. https://doi.org/10.1017/9781108993340.026

Responsible AI: Evolving Bodies of Practice

Fabio Tollon, University of Edinburgh, Edinburgh, UK

In recent years 'Responsible AI' (R-AI) has been applied to a number of contexts and research applications (Dignum, 2019; Zhu, 2019; De Laat, 2021). On the surface this seems a good thing, as of course we want the development, deployment, and use of AI-systems to be in line with certain normative principles, and it seems the 'responsible' frame can give us

just that. R-AI can ensure that AI-systems respect human rights and are aligned with democratic values. However, just what exactly R-AI means is contested, and often undefined. This raises problems for translating the 'principles' of various R-AI guidelines into meaningful 'practices' for those developing AI-systems. As noted, R-AI "has now become a brand-like umbrella term for the development of principles, approaches and methods of understanding what responsible AI development means and how it can be implemented" (Drage, McInerney and Browne, 2024).

While we might welcome this 'umbrella', we ought to be careful. Recent years have also shown that our governance and regulation of AI have, for the most part, been ineffective (Sadek et al., 2024). This paper contributes a novel perspective to this debate by outlining the major historical disciplinary orientations of what would eventually become 'R-AI'. By tracing the history of reflections on technology and responsibility from the 1960s, through STS, computer ethics and roboethics, to the present day, I will present important lessons from each phase of our reflection on technology and responsibility.

These reflections will come to bear on the '1st wave' of R-AI, which began in around 2015. It will allow me to answer questions such as: (1) what is to be included under the 'umbrella' of R-AI, and where do these participants come from? (2) What are the major historical themes and orientations that drive the field? (3) Is there really one R-AI community, or are there sets of intersecting and interconnected communities?

The key takeaway from this study is that 'Responsible AI' is not a label that has to do with some specific set of principles and values. More than that, it remains unclear whether there is one group of practitioners that we can call 'the' Responsible AI community. Instead, what we observe is that R-AI consists of many overlapping and intersecting communities, with diverse, contested, and evolving bodies of practice.

This brief tour showcases the way that ethical reflection on technology has changed over time, and how the focus on AI is a relatively new phenomenon. The practitioners who currently make up the R-AI ecosystem come from these (and many more) disciplines and sectors, and their interaction is what makes the R-AI ecosystem what it is. The hope is that this study, with its focus on the history of R-AI, can help ground these practices in a way that both showcases the dynamism of these distinct sets of practices, and provides some normative orientation for how best to enable a flourishing ecosystem.

By getting a better historical handle on R-AI, we can better promote a philosophically robust understanding of the concept. This means, among other things, acknowledging that there is no 'one' R-AI community, but rather a network of intersecting and interconnected communities. The practitioners who currently make up the R-AI ecosystem come from many different disciplines and sectors, and their interaction is what makes the R-AI ecosystem what it is. R-AI, on this framing, is not a 'problem' to be 'solved', but a process to be governed.

References

De Laat, P.B. (2021) 'Companies Committed to Responsible AI: From Principles towards Implementation and Regulation?', Philosophy & Technology, 34(4), pp. 1135–1193. Available at: https://doi.org/10.1007/s13347- 021-00474-3.

Dignum, V. (2019) Responsible Artificial Intelligence. Cham: Springer Nature Switzerland. Available at: https://doi.org/10.1007/978-3-030-30371-6.

Drage, E., McInerney, K. and Browne, J. (2024) 'Engineers on responsibility: feminist approaches to who's responsible for ethical AI', Ethics and Information Technology, 26(1), p. 4. Available at: https://doi.org/10.1007/s10676-023-09739-1.

Sadek, M. et al. (2024) 'Challenges of responsible AI in practice: scoping review and recommended actions', AI & SOCIETY [Preprint]. Available at: https://doi.org/10.1007/s00146-024-01880-9.

Zhu, W. (2019) '4 Steps to Developing Responsible AI. World Economic Forum'. Available at: https://www.weforum.org/agenda/2019/06/4-steps -to-developing-responsible-ai/. (Accessed: 16 November 2023).

Developing trustworthy social AI

Heike Felzmann, University of Galway, Republic of Ireland

Robust social AI – AI that is able to engage with humans in a sustained manner that is experienced as relationally convincing and cognitively appropriate - is becoming increasingly feasible with recent developments in generative AI. Uses for such AI applications range from increasingly capable customer service chatbots to personal AI assistants to AI companions.

The development of social AI raises a range of ethical concerns with regard to the design of humanAI interaction experiences that are relationally acceptable without deceiving users about the nature of the device, without creating experiences that are manipulative or could potentially be detrimental to human users, and without leading to problematic replacement of human contact. Ethical considerations regarding social AI mirror to some extent concerns discussed in relation to social robotics, but with the difference of having the potential of significantly scaled up deployment and substantially more flexible presentations and customisation. Social AI as a primarily digital system also provides substantially increased opportunities of seamlessly integrating surveillance capitalist design elements, both in relation to ongoing intimate data extraction from users and using extracted data for shaping users' behaviour.

The European Commission Ethics Guidelines for Trustworthy AI centre on trustworthiness as important concept to guide the responsible ethical development of AI, proposing a set of seven criteria to achieve trustworthy AI, without, however, providing a more detailed analysis of what characterises trustworthiness itself. The focus on analysing trustworthiness itself is motivated by the need to include the relational element of trustworthiness in the case of social AI - it is not just the product development and deployment that need to meet criteria of trustworthiness, but the user also needs to experience the social AI as trustworthy in their interaction. However, as will be shown, the relational experience of trustworthiness has relevance for understanding the trustworthiness of AI in general.

This paper proposes to draw on a philosophical analysis of the conceptual components of trust and trustworthiness, in order to establish trustworthiness as a suitable concept that can underpin a holistic understanding of responsible innovation in AI. The advantage of drawing on such a conceptualisation is twofold: (i) It has the conceptual resources to differentiate and clarify relational and task-related elements of trust and allow both a re-interpretation of the specific relevance of proposed criteria to the achievement of trustworthiness and the identification of missing or underemphasised relational aspects. (ii) It enables the application of the concept of trustworthiness to different stakeholders with different functions. One important element of trustworthiness is the recognition of responsibility towards the user of social AI and the societal context within which this use is taking place. Different parties have specific roles and need to engage different actions in response to this responsibility. In the case of social AI three different parties will be highlighted, with each making specific contributions to the achievement of trustworthy social AI: the social AI application itself, the developers, and the organisation that makes the social AI application available to users.

Responsible AI Implementation

Serinha Murgorgo[1], Nhien Nguyen[1]
[1]Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Artificial Intelligence (AI) has ushered in a new age of opportunities and challenges, unlocking its potential in this rapidly evolving technological landscape. It sparks excitement as organisations strive to get the most from investing in AI technology, which helps them improve decision-making, streamline operations, and solve important real-world challenges at scale (Berente et al., 2021). Generative AI has taken the world by storm, generating results based on historical data and future predictions. However, issues arise mainly from the opacity associated with organisations' adoption of AI. These systems can perpetuate biases present in data, lack interpretability in decision-making processes, and they are costly to train and maintain.

Thus, responsible AI has gained significant importance in ensuring trust in AI systems, addressing ethical and legal issues, and fostering ethical decision-making (Brumen et al., 2023; Dignum, 2019), especially with the rise of Generative AI. Responsible AI is defined as a governance framework that documents how a specific organisation addresses the challenges around artificial intelligence (AI) from an ethical and legal point of view (Dignum, 2019). Due to the novelty of responsible AI, literature is revolving and evolving (Rees & Müller, 2022), highlighting the need for continuous exploration, refinement and integration into organisational practices.

Despite the growing literature on responsible AI and numerous guidelines and initiatives, a significant gap remains in translating responsible AI principles into practice within organisational settings. This disparity between the principles and their practical application largely stems from the ambiguity of ethical principles and their perceived inadequacy in effectively addressing the full range of potential negative consequences associated with AI technologies (Jobin et al., 2019; Rakova et al., 2021). Responsible AI is more than just ticking

ethical boxes or adding features to AI systems (Dignum, 2019); it considers responsibility, regulation and control, ethics, transparency, design, and socioeconomic impact. The question of how to effectively implement and integrate responsible AI in organisations has become more important than ever, as it profoundly affects business, the environment, and society. By reviewing the literature on responsible AI, this paper will explore the 'how' of responsible AI governance in the organisational context and shed light on the implementation perspective.

The paper's findings indicate that responsible AI practices result in increased confidence and trust in decision-making. Collaborations that foster an inclusive AI ecosystem that addresses common challenges associated with AI systems. Additionally, a lack of leadership support hinders deep engagement with responsible AI issues. The paper contributes to enhancing the understanding of responsible AI implementation, revealing contextual factors and insights into cultural and organisational changes for effective AI implementation.

References

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. MIS Quarterly, 45(3).

Brumen, B., Göllner, S., & Tropmann-Frick, M. (2023). Aspects and Views on Responsible Artificial Intelligence. In G. Nicosia, V. Ojha, E. La Malfa, G. La Malfa, P. Pardalos, G. Di Fatta, G. Giuffrida, & R. Umeton (Eds.), Machine Learning, Optimization, and Data Science (pp. 384–398). Springer Nature Switzerland.

Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way (Vol. 1). Springer.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. Proc. ACM Hum.-Comput. Interact., 5(CSCW1), 1-23.

Rees, C., & Müller, B. (2023). All that glitters is not gold: trustworthy and ethical AI principles. AI and Ethics, 3(4), 1241-1254.

Algorithms down from the Moral Ivory Tower: Towards an Ideal 'Non-Ideal' Theory of AI Ethics.

Alexander M.S. Hjorth-Johansen[1], Christian Fieseler[1]

[1] BI Norwegian Business School

In the field of political philosophy, the 'ideal theory' approach to justice begins by suggesting a hypothetical 'ideal' state (derived from hypothetical ideal socio-political conditions) to which we should aspire (Valentini, 2012). AI ethics can be said to align with this paradigm (cf. Fazelpour & Lipton, 2020; Estrada, 2020), as much of the existing literature on AI ethics focuses on 'top-down' institutional guidelines and the discussions surrounding the principles that constitute them, such as *transparency*, *justice*, *fairness*, *non-maleficence*, *responsibility*, and *privacy* (Jobin et al., 2019). These principles have been criticized for being overly vague and lacking in

semantic content, providing limited specific recommendations, and failing to address the fundamental normative and political conflicts inherent to key concepts (Mittelstadt, 2019). In addition to these constraints, this approach can also lead to detrimental outcomes, such as paving the way for ethics-washing (van Maanen, 2022), and potentially triggering irresponsible behavior arising from the mismatch between algorithms and their real-world application contexts (Munn, 2023). Given political diversity and the resulting disagreements about moral questions—and sometimes even about the nature of morality itself—many traditional and conventional moral 'high' principles fall short in guiding the design of AI (Robinson, 2023). Agents adhering to various principles often favor different policies (Woodgate & Ajmeri, 2024).

The 'non-ideal theory' approach, on the other hand gives precedence to the realities and limitations of our dynamic and non-ideal world. It addresses issues such as the feasibility and fact-sensitivity of principles (Volacu, 2017), and tends to derive and base its principles on empirical facts observed in the world. When considering this in the context of AI ethics, one might suggest acknowledging and addressing the diverse and often conflicting interests within society and thereafter align algorithms accordingly (Wong, 2020). However, this approach carries its own set of challenges and can lead to substantial and unforeseen consequences; some of which could be disastrous (Baum, 2020). While considering stakeholder values and needs is crucial and acquiring descriptive information about them may prove useful, this alone cannot resolve all potential value tensions between diverse opposing groups, leaving designers and policymakers to continue having to navigate complex ethical decisions. Thus, we come full circle, finding ourselves once again dependent on more 'abstract' ethical principles to mitigate or manage potential trade-offs (Woodgate & Ajmeri, 2024). But which ethical principle(s) should we adopt to resolve such tensions and conflicting needs? Should we rely on the moral beliefs of an individual, a specific society, or a collective global consensus—assuming one exists?

In an attempt to resolve this dilemma, we propose an *ideal 'non-ideal' theory of AI ethics*—an 'ideal' approach to principles that are grounded in 'non-ideal' premises, such as fact-sensitivity and feasibility, while considering the everchanging public interest (Züger & Asghari, 2023). These 'overarching' principles are here given a *prima facie* status that is subject to ongoing reassessment, while serving also as guardrails to prevent undesirable outcomes and maintain broader global consistency. In this pragmatic approach, overarching principles are balanced with localized needs, creating a system of checks and balances between the two levels to ensure certainty, predictability, and safety. This approach aims to acknowledge the inevitable trade-offs between the practical realities of the non-ideal world we inhabit and the need for some type of higher-order principles. The framework aims to ensure that the solution remains consistent, flexible, and ethically robust, accounting for the global nature of AI while embracing the politically diverse landscape, thus reflecting the techno-social dynamics of our times.

References

Baum, S.D. 2020. Social choice ethics in artificial intelligence. AI & Soc, 35: 165–176. https://doi.org/10.1007/s00146-017-0760-1

Estrada, D. 2020. Ideal theory in AI ethics. ArXiv, abs/2011.02279.

Fazelpour, S. & Lipton, Z.C. 2020. Algorithmic Fairness from a Non-ideal Perspective. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA: 57–63. https://doi.org/10.1145/3375627.3375828

Jobin, A., Ienca, M. & Vayena, E. 2019. The global landscape of AI ethics guidelines. Nat Mach Intell, 1: 389–399. https://doi.org/10.1038/s42256-019-0088-2

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. Nat Mach Intell, 1: 501–507. https://doi.org/10.1038/s42256-019-0114-4

Munn, L. (2023). The uselessness of AI ethics. AI Ethics, 3: 869–877.https://doi.org/10.1007/s43681-022-00209-w

Robinson, P. 2023. Moral disagreement and artificial intelligence. AI & Soc. https://doi.org/10.1007/s00146-023-01697-y

Valentini, L. 2012. Ideal vs. Non-ideal Theory: A Conceptual Map. Philosophy Compass, 7: 654-664. https://doi.org/10.1111/j.1747-9991.2012.00500.x

van Maanen, G. 2022. AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics. DISO, 1(9). https://doi.org/10.1007/s44206-022-00013-3

Volacu, A. 2018. Bridging Ideal and Non-Ideal Theory. Political Studies, 66(4): 887-902. https://doi.org/10.1177/0032321717730297

Wong, P.H. 2020. Democratizing Algorithmic Fairness. Philos. Technol., 33: 225–244. https://doi.org/10.1007/s13347-019-00355-w

Woodgate, J. & Ajmeri, N. 2024. Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. ACM Comput. Surv. 56, 11, Article 289 (July 2024), 37 pages. https://doi.org/10.1145/3672394

Züger, T., & Asghari, H. 2023. AI for the public. How public interest theory shifts the discourse on AI. AI & Society, 38: 815–828. https://doi.org/10.1007/s00146-022-01480-5