Towards a system-theoretic approach to algorithmic (un)fairness

EVA DE WINKEL[*], Delft University of Technology, The Netherlands
JACQUELINE KERNAHAN[*], Delft University of Technology, The Netherlands
ROEL DOBBE, Delft University of Technology, The Netherlands

Most scholarship on algorithmic fairness understands fairness as a static problem that is addressed in the design of the algorithm and its components, overlooking its embedding in complex contexts of use and governance. This static framing limits the applicability of existing approaches to algorithmic fairness in new domains, where stakeholders lack established fairness norms and analogies to other fields may fall short. This paper examines the challenges of operationalizing algorithmic fairness in new contexts through a system-theoretic lens. Using a case study on algorithmic systems for managing grid congestion in electrical distribution grids, we identify three core challenges: (1) anticipating *situations of unacceptably unfair outcomes*, (2) localizing contributing *factors*, and (3) identifying *interventions and associated responsibilities* to prevent such outcomes. Drawing on system safety, a discipline that has dealt with complex safety problems in algorithmic systems for decades, we propose concepts and tools to support a system-theoretic approach to fairness.

Keywords: fairness, system safety, algorithmic systems, artificial intelligence, electrical distribution systems

Reference Format:
Eva de Winkel, Jacqueline Kernahan, and Roel Dobbe. 2025. Towards a system-theoretic approach to algorithmic (un)fairness. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 7 pages.

1 Introduction

As algorithmic systems become more integrated into society, concerns about fairness have grown. Over the past decade, the algorithmic fairness community has aimed to address fairness concerns, such as racial discrimination in criminal justice systems and hiring practices [15]. These efforts have revealed that many competing notions of fairness cannot all be satisfied simultaneously [4, 13]. Consequently, scholars increasingly advocate for mitigating "unfairness" in algorithmic systems rather than striving for "algorithmic fairness" [2, 10].

Mitigating unfairness in algorithmic systems first requires its identification. Much of the research in algorithmic fairness focuses on sectors such as criminal justice, predictive policing, and hiring decisions where many sources of unfairness and insight into historical forms of structural injustice are already identified through well-known case studies [15]. Beyond these recurring case studies, potential issues of unfairness appear to be relatively underexplored,

[*] Both authors contributed equally to this research.

Authors' Contact Information: Eva de Winkel, Delft University of Technology, Delft, The Netherlands, e.dewinkel@tudelft.nl; Jacqueline Kernahan, Delft University of Technology, Delft, The Netherlands, j.a.kerhanan@tudelft.nl; Roel Dobbe, Delft University of Technology, Delft, The Netherlands, r.i.j.dobbe@tudelft.nl.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

perhaps reflecting the inherent difficulties in identifying their sources within these emerging domains. Developers often cannot anticipate all existing prejudices, social biases, and protected groups that may interact with the model and control system, making it difficult to address these issues explicitly during the design phase [9].

Furthermore, the constitutive factors that contribute to unfairness, as well as the possible levers to intervene, are typically spread across a broader set of processes, stakeholders, and responsibilities [8, 21]. Some of these lie outside the visibility and control of algorithm developers. The influence of these broader factors appears to be relatively understudied in the algorithmic fairness literature, which tends to emphasize the developer's role in designing fairness metrics and technical bias mitigation strategies (e.g., see [18, 19]). While these efforts offer valuable contributions, they may lack a systematic framework for identifying, addressing, and monitoring unfairness in emerging contexts. Additionally, they often provide limited guidance on the roles and responsibilities of stakeholders beyond the developer in efforts to promote fairness.

In this paper, we ask: how can unfairness in algorithmic systems be identified and addressed in domains where fairness concerns are not yet well understood? Our objective is to develop a system-theoretic perspective on algorithmic fairness that accounts for the distributed nature of responsibility and control. We do so by exploring a novel problem domain for algorithmic fairness: electrical distribution systems, where algorithmic systems are deployed to make more efficient use of the electricity grid amid growing renewable integration and electrification.

Our paper is structured as follows. We begin by introducing a system-theoretic perspective on safety and fairness in algorithmic systems. This conceptual foundation informs our subsequent case study on managing grid congestion, a domain where algorithmic fairness concerns are still emerging. In this context, we examine the challenges of applying algorithmic fairness in such a new domain, where established norms, metrics, and institutional practices may be lacking. Finally, we explore how anticipatory tools from system safety can help address these challenges.

2 A system perspective on safety and fairness

Recent scholarship has emphasized the importance of adopting a sociotechnical systems perspective to understand and address fairness in algorithmic systems [3, 9, 24]. This perspective recognizes that algorithmic outcomes are not solely the result of technical design choices but emerge from complex interactions between technical components, institutional structures, and social practices. As such, evaluating fairness requires examining the broader system in which algorithms are embedded, rather than focusing exclusively on the algorithmic artifact itself [5, 9].

This need for a broader systems perspective mirrors developments in the field of safety engineering. Historically, three major paradigms have shaped safety thinking. The Safety I perspective, which became dominant from the 1960s onward in domains such as aviation and nuclear energy, focused on identifying and eliminating component failures as the primary cause of accidents [12]. The Safety II perspective expanded this view by shifting attention from preventing failures to understanding and supporting how systems succeed under varying conditions [12].

Building on and critiquing earlier approaches, the system safety (or Safety III) perspective introduced a systems-theoretic approach to managing hazards [16]. In system safety, designing for safety focuses on preventing unsafe situations that may lead to loss or harm by controlling factors within the engineered system and environmental conditions that may give rise to such situations. To prevent unsafe situations, interventions must be implemented at the appropriate level by stakeholders capable of influencing the relevant factors. This includes the engineering of the algorithmic system, the design of processes (such as norms under the AI Act), institutional mechanisms (like regulatory oversight to ensure standards are met), and the associated responsibilities within an organization. A key contribution is its use of anticipatory tools to identify and mitigate potential hazards before they occur.

Fairness and safety share important conceptual parallels. Rather than trying to create 'safe' systems, system safety focuses on eliminating 'unsafe' situations [16]. This parallels the called-for approach in the algorithmic fairness literature to achieve "fairness" through the elimination of "unfairness". Like fairness, safety is not a static property; new unsafe situations can arise due to changes in the designed system, its domain, or interdependent systems. Furthermore, like fairness, safety is not only an engineering problem; it requires a comprehensive understanding of how design, maintenance, management, policy, regulation, and oversight impact the overall system. From this perspective, a systems-theoretic approach offers valuable tools for anticipating and addressing unfairness.

3 Case study: grid congestion in electrical distribution systems

To ground our approach, we present a case study of algorithmic systems developed for managing grid congestion in electrical distribution systems. Grid congestion occurs when power demand or supply exceeds the technical limits of cables and equipment [1], which is increasingly common in countries such as the Netherlands. To address this challenge, grid operators are developing algorithmic systems that actively distribute grid capacity among households during operations to keep the load within allowed limits [20]. Fairness in this context refers to the equitable distribution of grid capacity among electricity users, ensuring that no group is systematically disadvantaged. For example, households with rooftop solar panels may be prioritized differently than those without, potentially leading to disparities in access to grid capacity.

Ensuring that the distribution of grid capacity is *fair*, is particularly complex due to the fragmented responsibility and control across a wide range of stakeholders. Within grid operators, responsibilities are divided among departments such as operations, grid planning, and asset management. Externally, stakeholders such as energy regulators, technology providers, national and local governments, and end-users either shape how algorithmic systems are designed, or influence how the burdens and benefits of the broader energy system are distributed. These stakeholders often operate with differing objectives and limited coordination, making it difficult to align on shared fairness goals in both system design and governance [7, 23].

4 Challenges of algorithmic fairness in a new domain

Applying algorithmic fairness to new domains presents several challenges, particularly when fairness concerns are not yet well understood or embedded in institutional practices. In this section, we highlight three core challenges in this process: (1) anticipation of *situations of unacceptably unfair outcomes*, (2) localizing *factors* that might lead to such situations, and (3) identifying *interventions and associated responsibilities* for preventing unacceptable unfairness caused by an algorithmic system. Figure 1 provides a conceptual illustration of how multiple factors, which can be within the visibility and control of various stakeholders, might contribute to an unfairness situation in a given domain.

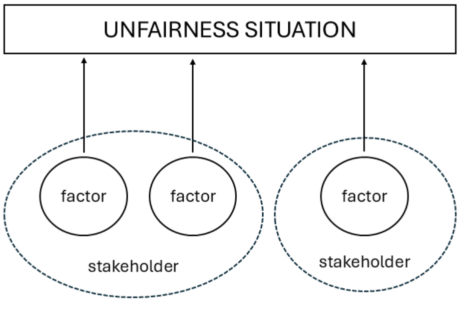

Fig. 1. A combination of factors, controlled by various stakeholders, can contribute to unfairness

4.1 Anticipation of unfairness situations

Identifying what constitutes “unfairness” in new contexts can be challenging. Often, unfair situations are not recognized until they occur, meaning we must wait for these scenarios to arise before addressing them (e.g. [6, 14]). In emerging domains, stakeholders, including system designers, regulators, and operators, may lack historical precedents or established fairness norms to guide them. While analogies to other domains (e.g., criminal justice or hiring) may offer some insight, they are not always fully transferable.

For example, in the case of grid capacity distribution, the axes along which unfairness might occur can differ significantly from those in the COMPAS or Amazon cases, where unfairness emerged due to differences in race and gender [6, 14]. Here, disparities may arise between households and businesses, between those with and without rooftop solar panels, or across socio-economic groups. Anticipating such situations requires a proactive, context-sensitive approach that considers how technical decisions interact with social and institutional structures. This is further complicated by the dynamic nature of the domain, where outcomes can shift over time due to policy changes, infrastructure developments, or changing user behavior [17, 21].

4.2 Localizing factors contributing to unfairness situations

Fairness concerns often stem from a complex web of decisions made by multiple actors. While technical teams may focus on data and model behavior, broader societal impacts are also shaped by how algorithmic systems are embedded in institutional processes and policy environments [10]. In many cases, no single actor has full visibility into all the factors that contribute to unfair outcomes.

In the electrical distribution system, for instance, different departments within grid operators, such as operations, planning, and asset management, make decisions that affect how grid capacity and associated costs are distributed. At the same time, government agencies influence these distributions through subsidies for technologies like heat pumps, electric vehicles, and solar panels. These subsidies often fail to reach low-income households due to high upfront costs, lack of home ownership, or limited access to financing, thereby reinforcing existing inequalities [22]. A systematic method is needed for identifying potential factors that could lead to unfairness within algorithmic systems, the broader domain, and interacting systems from other domains.

4.3 Defining responsibilities for preventing unfairness situations

Intervention methods for eliminating factors that contribute to unfairness do not necessarily have to be within the algorithmic model; they can also be introduced through governance mechanisms such as policy [10]. In some cases, addressing unfairness through technical methods may even be impossible, making it necessary to mitigate unfairness through other parts of the system. While algorithm designers may identify problematic patterns, they often lack the authority to implement changes at the policy or organizational level. Conversely, policymakers may not be aware of how their decisions interact with algorithmic systems to produce unintended consequences.

In the case of grid capacity distribution, for example, fairness may be affected by how regulators set tariffs or how municipalities prioritize infrastructure upgrades. These decisions can create structural advantages or disadvantages for certain groups, which algorithmic systems may then reinforce. Ensuring fairness thus requires a coordinated effort among stakeholders to define responsibilities, establish monitoring mechanisms, and implement interventions at multiple levels of the system.

5 System-theoretic tools for anticipating algorithmic (un)fairness

To address the challenges of operationalizing fairness in new and underexplored domains, we draw four types of tools from system safety that support the identification, analysis, and mitigation of unfairness. These tools are designed to anticipate and prevent unsafe outcomes in complex sociotechnical systems and can be adapted to help stakeholders navigate fairness concerns in contexts where established norms, metrics, and case studies may be lacking. The four types of tools are as follows.

Scenario analysis: System safety provides concepts and tools, such as scenario analysis, to anticipate situations of unacceptably unfair outcomes before they occur [11, 16]. This can include qualitative modeling of potential unfair outcomes through workshops or serious gaming, as well as quantitative modeling, such as Monte Carlo simulations, to analyze the distribution of burdens and benefits in various scenarios.

System modeling: System safety also provides tools for tracing direct and indirect factors contributing to unfair outcomes within complex systems. System modeling and process mapping methods can be used to trace and investigate the underlying interactions that may lead to unfair outcomes [11, 16]. These methods can help stakeholders understand how various interactions within the system may lead to unfair situations.

Control structures: System safety emphasizes the importance of clear responsibilities and accountability structures for ensuring the proper behavior and outcomes of complex systems. Safety control structures can be adapted to "fairness control structures" which define roles and responsibilities for addressing unfairness in algorithmic systems [16]. Control structures visualize the coordination required among stakeholders to mitigate unfairness, including what information must be shared and what conditions should trigger action.

Governance tools: In system safety engineering, safety management systems provide a structured framework for proactively identifying and mitigating hazards throughout the lifecycle of complex systems. These safety management systems formalize processes such as hazard analysis, documentation, and monitoring, and are widely used in high-risk domains like aviation and nuclear energy. A similar logic is emerging in the governance of algorithmic systems through algorithmic impact assessments, which aim to institutionalize anticipatory practices by requiring structured procedures for risk identification and stakeholder consultation before deployment.

6 Discussion and way forward

This paper has explored the challenges of applying algorithmic fairness in new domains, using a case study in electrical distribution systems. We argued that concepts and tools from system safety offer a promising foundation for addressing these challenges through a structured and anticipatory approach. However, these tools also present limitations: they do not resolve the normative ambiguity of fairness, which is context-dependent and shaped by diverse stakeholder values, and they can be resource-intensive and require adaptation to specific domains. Despite these challenges, system-theoretic tools offer a valuable lens for engaging with fairness in complex sociotechnical systems, particularly where established fairness norms and metrics are lacking.

While algorithmic impact assessments have gained traction as a governance mechanism in the algorithmic fairness field, other system-theoretic tools have received comparatively less attention. Our work contributes to an emerging conversation about how such tools, traditionally used in safety-critical domains, might support fairness in complex sociotechnical systems. Future research should investigate how these methods can be tailored to algorithmic fairness contexts and integrated with existing technical and governance approaches.

References

[1] Karim L. Anaya and Michael G. Pollitt. 2015. Options for allocating and releasing distribution system capacity: Deciding between interruptible connections and firm DG connections. *Appl. Energy* 144 (2015), 96–105. https://doi.org/10.1016/j.apenergy.2015.01.043

[2] Stefan Buijsman. 2024. Navigating fairness measures and trade-offs. *AI and Ethics* 4 (2024), 1323–1334. https://doi.org/10.1007/s43681-023-00318-0

[3] Maarten Buyl and Tijl De Bie. 2024. Inherent Limitations of AI Fairness. *Commun. ACM* 67, 2 (jan 2024), 48–55. https://doi.org/10.1145/3624700

[4] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[5] Efrén Cruz Cortés and Debashis Ghosh. 2020. An Invitation to System-wide Algorithmic Fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 235–241. https://doi.org/10.1145/3375627.3375860

[6] Jeffrey Dastin. 2018. Insight - Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/idUSKCN1MK0AG

[7] Eva de Winkel, Zofia Lukszo, Mark Neerincx, and Roel Dobbe. 2025. Adapting to limited grid capacity: Perceptions of injustice emerging from grid congestion in the Netherlands. *Energy Res. & Soc. Sci.* 122 (2025), 103962. https://doi.org/10.1016/j.erss.2025.103962

[8] R. Dobbe and A. Wolters. 2024. Toward Sociotechnical AI: Mapping Vulnerabilities for Machine Learning in Context. *Minds & Machines* 34 (2024), 12 pages. https://doi.org/10.1007/s11023-024-09668-y

[9] Mateusz Dolata, Stefan Feuerriegel, and Gerhard Schwabe. 2022. A sociotechnical view of algorithmic fairness. *Information Systems Journal* 32, 4 (2022), 754–818. https://doi.org/10.1111/isj.12370

[10] Ben Green. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35 (2022), 90. https://doi.org/10.1007/s13347-022-00584-6

[11] Erik Hollnagel. 2017. *Hollnagel, Erik. FRAM: the functional resonance analysis method: modelling complex socio-technical systems*. CRC press, London, UK.

[12] Erik Hollnagel. 2018. *Safety-I and safety-II: the past and future of safety management*. CRC press, London, UK.

[13] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[14] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[15] Daphne Lenders and Anne Oloo. 2024. 15 Years of Algorithmic Fairness – Scoping Review of Interdisciplinary Developments in the Field. arXiv:2408.01448 [cs.CY] https://arxiv.org/abs/2408.01448

[16] Nancy Leveson. 2012. *Engineering a safer world: applying systems thinking to safety*. MIT press, Cambridge, MA.

[17] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 3150–3158.

[18] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. https://doi.org/10.1145/3457607

[19] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (Feb. 2022), 44 pages. https://doi.org/10.1145/3494672

[20] Bharath Varsh Rao, Mark Stefan, Thomas Brunnhofer, Roman Schwalbe, Roman Karl, Friederich Kupzog, Gregor Taljan, Franz Zeilinger, Peter Stern, and Martin Kozek. 2022. Optimal capacity management applied to a low voltage distribution grid in a local peer-to-peer energy community. *Int. J. Electr. Power Energy Syst.* 134 (2022), 107355. https://doi.org/10.1016/j.ijepes.2021.107355

[21] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[22] Benjamin K. Sovacool, Max Lacey Barnacle, Adrian Smith, and Marie Claire Brisbois. 2022. Towards improved solar energy justice: Exploring the complex inequities of household adoption of photovoltaic panels. *Energy Policy* 164 (2022), 112868. https://doi.org/10.1016/j.enpol.2022.112868

[23] Benjamin K. Sovacool, Sanya Carley, and Lynne Kiesling. 2024. Energy justice beyond the wire: Exploring the multidimensional inequities of the electrical power grid in the United States. *Energy Res. & Soc. Sci.* 111 (2024), 103474. https://doi.org/10.1016/j.erss.2024.103474

[24] Anna-Lena Theus. 2023. Striving for Affirmative Algorithmic Futures: How the Social Sciences can Promote more Equitable and Just Algorithmic System Design. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (, Chicago, IL, USA,) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 558–568. https://doi.org/10.1145/3593013.3594022