What If. . . Counterfactual Explanations Were to Be Deployed?

Francesco Leofante
Department of Computing
Imperial College London
f.leofante@imperial.ac.uk

Abstract

Explainable AI is now a mature and rapidly expanding field, with a wide range of methods for interpreting models and explaining their behaviour. Despite this progress, many of these methods have yet to make their way into the engineering pipelines where AI is actually deployed. This gap suggests a question that is central to this paper: what would it take for explanations to be informative, useful and dependable when deployed? I approach this question through counterfactual explanations, which I view as a promising foundation for practical XAI: by showing what would need to change for an AI decision to differ, they are easy to interpret and carry actionable information. In their standard form, however, counterfactuals come with limitations of form and scope that constrain their applicability in deployment. I therefore consider three lines of work that build on standard counterfactual explanations and ask what would need to change for them to be better aligned with deployment requirements. The first concerns robustness, ensuring that counterfactuals remain valid under the perturbations that deployment introduces. The second extends them beyond one-shot decisions to capture sequential decision-making. The third takes counterfactuals as a starting point for contestability, arguing that meaningful contestation requires more than an explanation alone. I close by reflecting on energy systems as a critical testbed for explainability, one that highlights the promises and limitations of current methods and may offer a source of criteria for the next generation of explanation methods.

1 Introduction

In the span of a few years, Artificial Intelligence (AI) has evolved from a tool that advises consequential decisions into one that actively shapes them. These developments create a growing need for explanation methods that can respond to the practical and normative challenges such systems raise.

The explainable AI (XAI) community has responded to these challenges in several ways. Some researchers have pursued interpretability by design, developing models whose structure is intelligible [Rudin, 2019]. Others have sought understanding from within the model, with mechanistic interpretability aiming to identify the neurons and circuits responsible for specific behaviours [Olah *et al.*, 2020]. Still others have developed post-hoc methods that explain predictions in terms of influential features [Ribeiro *et al.*, 2016] or human-legible concepts [Kim *et al.*, 2018]. Closer to my own work, counterfactual explanations have been proposed to ask not why a model produced a given outcome, but what would need to change for it to produce a different one [Wachter *et al.*, 2017]. Each of these approaches has generated important insights, while also making clear that explainability is not a single problem with a single solution.

Despite the rapid progress of recent years, many explanatory methods have yet to translate into tools that can be safely and usefully integrated into engineering pipelines. In this paper, I investigate this gap in the context of counterfactual explanations, which I take as a natural starting point for practical XAI. I consider three lines of work that ask what would need to change, in the explanations themselves and in the systems that surround them, for counterfactuals to become deployable. First, they must be robust: a counterfactual should remain valid when deployment conditions depart from the idealised setting in which it was computed. Despite being often advocated as tools for algorithmic recourse, counterfactuals are critically fragile in their standard form. My work on robust counterfactuals identifies when this fragility arises, establishes conditions under which it can be avoided, and develops algorithms that produce explanations with stronger guarantees of reliability. Second, they must be extended beyond one-shot decisions to capture sequential decision-making. In this setting, the counterfactual questions of what would need to change differ from those arising in one-shot decision tasks. The challenge is no longer to explain why a model assigned a particular label to a particular input, but what would need to change for an agent to act differently over time: which actions would have to be selected, how a policy would have to change to induce a different trajectory, and which features of the environment would have to differ for another behaviour to become possible. Third, to be useful they must operate as part of a process that lets people act on them. Contestability and redress are one example: a counterfactual is a natural starting point for challenging a decision, but not enough on its own to support the exchange of reasons that contestation requires.

Finally, I turn to energy systems, a domain where these deployment challenges are especially pressing. The sector is increasingly seeking to bring AI into its operational pipelines. Yet it must do so under constraints that test the limits of current methods. I believe counterfactual explanations have a role to play, but not directly in their present form. Energy is a setting where their promises and limitations can be examined, and more broadly draw lessons for the next generation of explainability methods.

2 Counterfactuals and Robustness

Counterfactual reasoning has long been studied in cognitive science, where it plays an important role in how humans explain events [Byrne, 2019]. Inspired by this idea, the XAI community has proposed a technical notion of counterfactual explanation as the solution to an optimisation problem: given an input for which a model produces a particular decision, the aim is to find a minimally modified input for which the model would produce a different outcome [Wachter *et al.*, 2017]. This framing helps explain why counterfactual explanations are useful in practice: by showing what would need to change for an outcome to differ, they point users to what they could do to alter a decision that affects them.

However, this promise rests on the assumption that meaningful counterfactuals can be computed as solutions to an optimisation problem. In their standard form, counterfactual explanations are typically obtained by searching for an input minimally different from the original one. This often leads to solutions that lie close to the decision boundary of a model, or more generally in regions of the input space where the model may not generalise reliably [Pawelczyk *et al.*, 2022]. As a result, small changes in the setting for which a counterfactual was computed can be enough to invalidate the recourse it was meant to offer. A user who acts on such a recommendation may incur real cost, only to find that the decision they hoped to overturn remains unchanged.

Across a series of works, I have studied this fragility along four axes [Jiang *et al.*, 2024c], which I summarise briefly below. ① The first axis concerns *model change*: routine parameter updates can alter the behaviour of the model for which a counterfactual was generated, so that recourse that was once valid may later fail. In [Jiang *et al.*, 2024a] we addressed this by introducing abstractions over model parameters to capture possible updates, and using these abstractions to generate counterfactuals robust to bounded model changes. In [Marzari *et al.*, 2026], we extended this idea to a probabilistic setting, gaining scalability while retaining formal guarantees. ② The second axis concerns *model multiplicity*: many models with comparable accuracy may be trained for the same task, yet disagree on individual predictions [Black *et al.*, 2022]. In [Leofante *et al.*, 2023] we treated multiplicity as a relational property and gave the first method for synthesising counterfactuals valid across a set of neural networks, while in [Jiang *et al.*, 2024b] we extended this line of work using computational argumentation to jointly select predictions and recourse across competing models. ③ The third concerns *noisy execution*: users rarely implement recourse recommendations exactly, and a recommendation valid only at a single point may be of little use in practice. In [Leofante and Lomuscio, 2023] we studied this problem and developed techniques to certify whether a counterfactual remains valid within a neighbourhood of the prescribed point, thereby providing formal guarantees against this form of noisy execution. ④ The fourth and final axis concerns *procedural robustness*: similar individuals should receive similar recommendations [Slack *et al.*, 2021], yet many standard methods cannot guarantee this. In [Leofante and Potyka, 2024] we showed that this form of robustness is impossible to ensure in general for many existing approaches, and proposed instead to generate diverse sets of counterfactuals whose overlap is guaranteed for similar inputs, thereby recovering a procedural notion of robustness better aligned with fairness.

What this line of work reveals is that the properties required for counterfactuals to be useful in practice cannot be treated as modular additions to an otherwise standard counterfactual pipeline. They constrain one another, and progress on one front often introduces weaknesses on another. One key difficulty lies in the standard practice of generating counterfactuals through direct perturbations of the input. A promising approach is to work in a representation space that captures more of the structure of the space in which the data actually lie. A step in this direction is our recent work on LAPACE [Jiang *et al.*, 2026], which generates counterfactuals in a label-conditional latent space rather than through direct optimisation in the input space. This makes it possible to improve robustness and plausibility together, while still respecting practical constraints such as proximity and actionability. More importantly, this work points to a broader methodological lesson: if XAI is to support real applications, the properties that matter in practice must shape the design of counterfactual methods from the beginning, rather than being imposed on them afterwards.

3 Explaining Sequential Decisions

Another important dimension of practical XAI is the ability to explain sequential decisions [Baier *et al.*, 2026]. AI is increasingly being applied to control-intensive tasks, for instance in critical infrastructure [Vaquet *et al.*, 2024; Marchesini *et al.*, 2026], where decisions must be made over time through repeated interaction with an environment. In sequential decision-making, explanations may be required not just for isolated outputs but also for actions, trajectories, and the environmental conditions that shape them.

A useful distinction that has shaped my thinking around these problems is between *agent-centric* and *environment-centric* explanations. Agent-centric explanations focus on the agent's own decision process, asking, for instance, which states are relevant to a particular choice, or how an agent's behaviour would need to change for a desired outcome to become achievable. Environment-centric explanations focus instead on the structure of the environment, asking which features of the world constrain the agent's options, or how the environment itself would have to differ for a different behaviour to become possible. The two are complementary: the former explains behaviour by reference to the agent, the latter by reference to the world in which that behaviour is observed.

On the environment-centric side, I have explored these questions in the context of classical planning. In [Gigante *et al.*, 2025], we introduced *counterfactual scenarios* for automated planning, a framework that lifts explanation from a particular plan to the planning problem the agent is tasked with solving. The goal is to identify changes to the problem under which plans with desired properties become available. For example, an agent may be unable to reach a goal in the current environment, but could do so if some aspect of the environment were changed. This yields a novel type of *structural* explanations that characterise the conditions under which certain behaviours become possible. In practice, I see counterfactual scenarios less as a tool for recourse than as a means of *what-if* analysis: even when changes to the environment are not themselves actionable, they reveal why a desirable behaviour is unavailable and what features of the domain are responsible for this limitation.

On the agent-centric side, I moved from classical planning to Markov Decision Processes (MDPs) because my aim was to explain the object that most directly characterises an agent's behaviour over time, namely its policy, and MDPs provide the standard formal framework for doing so. In this setting, we introduced *counterfactual strategies* [Kobialka *et al.*, 2025]. The question here is no longer how a single action could have differed, but rather what minimal changes to a policy would raise the probability of a desirable outcome above an acceptable threshold. In subsequent work [Kobialka *et al.*, 2026], we began to complement this counterfactual perspective with attribution-based explanations for MDPs, asking not only how a strategy could have been different, but also which states of the MDP, under a given policy, were most influential towards determining the observed outcome.

These results are first steps in what I see as a much broader programme. The techniques I have proposed so far have been designed for models that are *explicit*, in that the model of the world and of the agent is given in closed form, and *stateful*, in that the agent's behaviour is described over well-defined states and transitions. However, most contemporary systems for planning and control, from reinforcement learning agents to emerging neuro-symbolic architectures, are often neither explicit nor stateful in this sense. Instead, their decision-making is encoded in the weights of a neural network, and the structure my methods rely on is often not there to be exploited. Extending principled explanation into this regime is in my view a central open problem in the area.

4 Explainability Beyond Explanations

The two lines of work discussed so far focus on making counterfactuals more reliable and on extending them to settings where decisions happen over time. Both are important, but both still treat explanation as the endpoint. In many practical settings, this is not enough. When AI systems make consequential decisions about people, the issue is not only whether those affected can understand what the system has done. It is also whether they can challenge a decision they believe is wrong, and whether there is any meaningful way for the system, or the process around it, to respond. Prior research and policy discussions use the terms *contestability* and *redress* to refer to these two requirements [Lyons *et al.*, 2021; UK DSIT, 2023]. Contestability is the ability of an affected party to challenge a decision by disputing the reasons, information, or assumptions on which it was based. Redress is the ability of the system, or of the wider pipeline in which it is embedded, to revise the decision or the basis on which it was made.

Counterfactual explanations have been advocated as a natural starting point for contestability [Wachter *et al.*, 2017]. By showing how changes in features would alter an AI decision, counterfactuals may serve as a prompt for questioning the assumptions behind it. This gives an affected person something concrete to challenge, though not a mechanism to question it. A contestation is not just a request for more explanation. It is a structured interaction that raises at least three questions: what can be disputed, when a challenge should be triggered, and how reasons are exchanged and assessed.

In [Leofante *et al.*, 2024], we proposed computational argumentation as a useful formal framework for modelling this interaction. It offers a language of arguments, attacks, and defences, together with semantics for determining which positions remain justified once competing claims are taken into account. It also fits naturally with multi-agent settings, where contestation may arise between a person and an AI system, or between multiple systems whose outputs must be reconciled. In that work, we began to characterise the kinds of contestation that may arise in such settings: disputes over the input, over the model, over the rationale given for a decision, or over the recourse proposed in response.

While computational argumentation can help structure the interaction, the question of what the arguments themselves should be remains open. For many classifiers, arguments can be formulated directly over input features, and existing explanation methods can often be adapted for this purpose. In generative models, however, the situation is different as the relevant objects of dispute are often not feature values but concepts. A user dissatisfied with the output of a generative model is unlikely to direct a challenge at the level of raw internal activations. The more natural object of contestation is the way concepts are represented and combined to produce the output; yet both are hidden from the user.

In recent work [Mansi *et al.*, 2026], we have taken a first step in this direction by developing methods to extract what we call *internal concept ontologies* from generative models. Focusing on text-to-image diffusion models, we defined these ontologies as graphs that capture concept representations distributed across different components of the model. Making this structure explicit gives users a more precise basis on which to contest an output: the issue is no longer only that the output is unsatisfactory, but that a concept may have a distorted internal representation. Crucially the same mechanism that makes this structure visible can also be used to intervene on it, providing a possible route from contestation to redress.

Together, these first steps suggest a broader lesson: counterfactuals can initiate a contestation, but they cannot sustain it alone. Identifying what would need to change is only the starting point of a challenge; the interaction that follows must also be modelled, including who is disputing what, on what grounds, and towards what revision, and the system must be

designed to respond. Where this modelling is absent, contestability reduces to a generic demand for further explanation, instead of the structured mechanisms of challenge and redress that meaningful contestation requires.

5 Looking Ahead: XAI in Action

I believe going further requires looking closely at the domains where AI is deployed and extracting from them the criteria that should drive the next generation of explanation methods.

One domain I find particularly exciting is energy systems. AI is increasingly discussed as part of the future of energy, but actual adoption remains uneven. Predictive AI is already used for tasks such as renewable generation forecasting, load forecasting, and predictive maintenance [UK Parliamentary Office of Science and Technology, 2023]. However, as we move from predictive AI to agentic and generative systems, applications become more sparse. A recent assessment of enterprise AI deployment identifies energy among the sectors where generative AI has produced widespread experimentation but little structural transformation [Challapally *et al.*, 2025]. At the same time, governments are increasingly considering how AI might support the transition to more resilient energy systems, while recognising the regulatory complexity such deployments would bring. For instance, Ofgem, the UK energy regulator, has set out explicit expectations around transparency and explainability in its guidance on the ethical use of AI in the energy sector, including the need for explainability and governance mechanisms that " *[...] allow users to escalate AI decisions that appear incorrect, biased, or opaque*" [Ofgem, 2026]. The themes of explainability and contestability that run through this paper thus already appear in regulatory expectations for the sector.

Meeting these expectations in practice will require concrete tools, and counterfactual explanations are a promising candidate. Indeed, several interesting applications already emerged during the AAAI 2025 Bridge on "Explainable AI, Energy and Critical Infrastructure Systems" [Leofante *et al.*, 2025], particularly for supporting what-if analysis around operational decisions. However, substantial technical challenges remain. First of all, there is no clear consensus on what a good explanation should be in this domain. Addressing this will require empirical user studies to elicit these requirements and translate them into design criteria for explanation methods. A second challenge is to develop new explanation methods that satisfy these requirements while simultaneously respecting the operational and regulatory constraints of energy systems, so that the resulting explanations remain actionable in practice. A third challenge is to embed such methods into the broader institutional and operational processes through which AI systems are actually used, so that explanations operate not in isolation but as part of the pipelines and governance structures that surround them. Achieving these objectives will require close collaboration with stakeholders, and this is one of the directions I am actively pursuing.

6 Conclusion and Outlook

This paper has argued that a gap remains between what current explanation methods offer and what deployment actually requires. I have investigated this gap through counterfactual explanations, discussing some of the technical solutions I have proposed in response. I have also argued that practical deployment requires explanatory systems as opposed to explanations alone, using contestability and redress are one example. Here, counterfactuals can serve as an initial trigger for a challenge, but a challenge only becomes meaningful when the interaction it initiates is structured explicitly, and I have discussed how such structure might be enforced. Looking further ahead, I have suggested that closing the gap with deployment will require looking more closely at application domains, so that their practical requirements can drive the research questions behind the next generation of explanation methods. To this end, I have highlighted energy systems as a domain I find especially compelling, precisely because it demands explanations of high quality.

This paper discussed some of the gaps I have encountered in my own work, but they are not the only ones I consider worth pursuing. One such gap not discussed here concerns a limitation inherent to counterfactuals themselves. By construction, they reason only over inputs and outputs. This is part of what makes them practical, but it also bounds what they can offer. The limitation is most consequential for redress, which often requires revising the model itself, and a method that sees only inputs and outputs cannot guide such an intervention. Combining what-if reasoning with a mechanistic view inside the model thus seems a particularly promising direction for supporting redress. A second gap concerns the integration of causal knowledge. While counterfactual and causal explanation are different in scope [Deighan and Byrne, 2026], in some applications the causal structure of the domain is exactly what is required to avoid problems such as the robustness failures discussed earlier. I see incorporating this kind of causal knowledge at scale, when it is available and useful, as an important open challenge. A third gap concerns security. As explanations reveal more about the system that produces them, they may also expose more of it to attackers. This is a serious concern in domains such as energy, where the systems in question are critical infrastructure. Contestability offers a concrete example: a party might initiate a contestation not to correct a genuine error, but to steer the system toward an outcome that favours them without legitimate basis. This risk becomes especially concrete with generative and agentic architectures, where the space of possible manipulations is large. Balancing explainability against these and other operational requirements is, I believe, one of the central challenges that practical XAI will have to address.

Acknowledgments

I am grateful to members of the Computational Logic and Argumentation group at Imperial College London for the discussions and collaborations that have helped shape the views and contributions described in this paper. I also thank Mansi, André Artelt, Alejandro Mercado and Francesca Toni for their feedback on earlier drafts of this paper.

References

[Baier *et al.*, 2026] Hendrik Baier, Mark T. Keane, Sarath Sreedharan, Silvia Tulli, and Abhinav Verma. Explanations for sequential decision-making - an overview. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 40948–40953. AAAI Press, 2026.

[Black *et al.*, 2022] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 850–863. ACM, 2022.

[Byrne, 2019] Ruth M. J. Byrne. Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6276–6282. ijcai.org, 2019.

[Challapally *et al.*, 2025] Aditya Challapally, Chris Pease, Ramesh Raskar, and Pradyumna Chari. The GenAI divide: State of AI in business 2025. Project NANDA preliminary findings, MIT NANDA (Networked Agents And Decentralized Architecture), July 2025. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf. Accessed: 2026-06-30.

[Deighan and Byrne, 2026] Laura N. Deighan and Ruth M. J. Byrne. On the different content of counterfactual and causal explanations. *Thinking & Reasoning*, 32(2):235–271, 2026.

[Gigante *et al.*, 2025] Nicola Gigante, Francesco Leofante, and Andrea Micheli. Counterfactual scenarios for automated planning. In *Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning, KR 2025, Melbourne, Australia, November 11-17, 2025*, 2025.

[Jiang *et al.*, 2024a] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Interval abstractions for robust counterfactual explanations. *Artif. Intell.*, 336:104218, 2024.

[Jiang *et al.*, 2024b] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Recourse under model multiplicity via argumentative ensembling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 954–963. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024.

[Jiang *et al.*, 2024c] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 8086–8094. ijcai.org, 2024.

[Jiang *et al.*, 2026] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Synthesising counterfactual explanations via label-conditional gaussian mixture variational autoencoders. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR, 2018.

[Kobialka *et al.*, 2025] Paul Kobialka, Lina Gerlach, Francesco Leofante, Erika Ábrahám, Silvia Lizeth Tapia Tarifa, and Einar Broch Johnsen. Counterfactual strategies for markov decision processes. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 412–420. ijcai.org, 2025.

[Kobialka *et al.*, 2026] Paul Kobialka, Andrea Pferscher, Francesco Leofante, Erika Ábrahám, Silvia Lizeth Tapia Tarifa, and Einar Broch Johnsen. Attribution-based explanations for markov decision processes. *CoRR*, abs/2605.09780, 2026. To appear at IJCAI 2026.

[Leofante and Lomuscio, 2023] Francesco Leofante and Alessio Lomuscio. Towards robust contrastive explanations for human-neural multi-agent systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, pages 2343–2345. ACM, 2023.

[Leofante and Potyka, 2024] Francesco Leofante and Nico Potyka. Promoting counterfactual robustness through diversity. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 21322–21330. AAAI Press, 2024.

[Leofante *et al.*, 2023] Francesco Leofante, Elena Botoeva, and Vineet Rajani. Counterfactual explanations and model multiplicity: a relational verification view. In *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023*, pages 763–768, 2023.

[Leofante *et al.*, 2024] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. Contestable AI needs computational argumentation. In *Proceedings of the 21st International*

Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024, 2024.

[Leofante *et al.*, 2025] Francesco Leofante, André Artelt, Demetrios Eliades, Anna Korre, Francesca Toni, and Tim Miller. Explainable ai, energy and critical infrastructure systems. *AI Magazine*, 46(3):e70033, 2025.

[Lyons *et al.*, 2021] Henrietta Lyons, Eduardo Velloso, and Tim Miller. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc. ACM Hum. Comput. Interact.*, 5(CSCW1):106:1–106:25, 2021.

[Mansi *et al.*, 2026] Mansi, Avinash Kori, and Francesco Leofante. Efficient bias mitigation in t2i diffusion models using concept graphs. *CoRR*, abs/2607.03397, 2026.

[Marchesini *et al.*, 2026] Enrico Marchesini, Eva Boguslawski, Alessandro Leite, Christopher Amato, Matthieu DUSSARTRE, Marc Schoenauer, Benjamin Donnot, and Priya L. Donti. MARL2grid-TR: A multi-agent RL benchmark in power grid operations. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Marzari *et al.*, 2026] Luca Marzari, Francesco Leofante, Ferdinando Cicalese, and Alessandro Farinelli. Probabilistically robust counterfactual explanations under model changes. *Artif. Intell.*, 351:104459, 2026.

[Ofgem, 2026] Ofgem. Ethical AI use in the energy sector. Guidance, Office of Gas and Electricity Markets, May 2026. https://www.ofgem.gov.uk/guidance/ethical-ai-use-energy-sector. Accessed: 2026-06-30.

[Olah *et al.*, 2020] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

[Pawelczyk *et al.*, 2022] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 4574–4594. PMLR, 2022.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[Slack *et al.*, 2021] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 62–75, 2021.

[UK DSIT, 2023] UK DSIT. AI regulation: A pro-innovation approach. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper, 2023. Department for Science, Innovation and Technology. White Paper presented to Parliament, 29 March 2023. Accessed: 2026-06-30.

[UK Parliamentary Office of Science and Technology, 2023] UK Parliamentary Office of Science and Technology. Energy security and artificial intelligence, 2023. POSTnote 735. Accessed: 2026-06-30.

[Vaquet *et al.*, 2024] Valerie Vaquet, Fabian Hinder, André Artelt, Inaam Ashraf, Janine Strotherm, Jonas Vaquet, Johannes Brinkrolf, and Barbara Hammer. Challenges, methods, data-a survey of machine learning in water distribution networks. In *Artificial Neural Networks and Machine Learning - ICANN 2024 - 33rd International Conference on Artificial Neural Networks, Lugano, Switzerland, September 17-20, 2024, Proceedings, Part IX*, volume 15024 of *Lecture Notes in Computer Science*, pages 155–170. Springer, 2024.

[Wachter *et al.*, 2017] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.