CATAPULT Energy Systems

Urban Frontiers

AI Risks for Energy Networks: Challenges, Management and Regulation

David Altabev
Director, Urban Frontiers

Stephen Haben
Senior Data Science Consultant,
Energy Systems Catapult

June 24

Contents

1. Executive Summary

The opportunity and the challenge

Delivering a net-zero carbon energy system is both the biggest challenge and opportunity that the UK energy sector faces, one that will require the deployment of new technologies at a scale and pace not seen before. The energy network of 2050 is rapidly changing to an increasingly digitised, decentralised, low carbon system. Managing this increasing complexity and its interactions efficiently will be critical to achieving our net-zero carbon targets. It will need to navigate a balancing act in terms of maximising the deployment of low and zero carbon technologies, maintaining the security and stability of the electrical grid, and minimising costs for consumers.

Artificial intelligence (AI) will be a key tool in managing this transition and enabling the necessary system of the future. Barriers to entry are being lowered and the technology is more readily available to an increasing array of practitioners, accelerating its adoption into the energy system and creating new opportunities for decarbonisation and energy efficiency.

Addressing the knowledge gap

Whilst much of the debate to date has rightly focussed on the opportunities that AI brings to the energy sector, there has been a lack of a counterbalance about what risks might arise from the deployment of increasingly powerful algorithms and the potential for conflicting demands across the energy system. The need to explore these adoption risks is pertinent given the rapid technological developments in AI and increasing deployment on our critical national infrastructure. At the same time, there is a need to ensure that risk mitigations are not overly stringent so as to not hinder powerful and effective AI innovations.

As algorithms and AI become more widely used for automation by different actors in the energy sector, from flexibility aggregators to the system operator, there is an increasing risk that these automated systems will interact in a way that is detrimental to the whole network. In a highly connected system, decisions made to optimise narrow objectives may have unintended consequences on the wider system. Algorithms applied in different parts of the energy network may have conflicting objectives that lead to undesirable oscillations, creating suboptimal behaviour (increasing costs for consumers). More concerningly, errors in one algorithm could trigger a reinforcing feedback loop that could lead to a wide scale system failure.

Scope & Objectives

In addition to the literature review and desk research, we interviewed 25 people from across key organisations in the energy sector to draw together a broad range of viewpoints. Our stakeholder engagement centred around several core themes;

1. AI use cases in the energy sector: Where is AI currently used and what are the likely future applications?

2. AI related risks in the energy sector and other sectors: What are they and what are the possible impacts, how might we mitigate them?
3. Risk management for AI: How do we manage risk, what risk management frameworks can we adapt?
4. Regulation and governance: Where does responsibility lie, what does good governance look like?

This report does not seek to downplay the economic, social and environmental opportunities and benefits that AI can bring to a multitude of use cases in the energy sector, but rather recognises that effective risk management can unlock innovation and ensure that it meets the needs of all stakeholders in the ecosystem. It should be seen as a primer for a much wider conversation that needs to happen across industry, to address the concerns raised through the stakeholder engagement.

AI Risks and the challenges of AI adoption

Over the coming years we are going to see AI algorithms deployed onto the energy system in potentially high-risk spaces in a way that we have not seen before. In some cases AI may create new risks, accentuate current risks, or it may create new indirect risks where the effect from multiple automated devices accumulate or exacerbate risks across the network.

This report has explored instances of risk from other industries where AI has been a contributing or lead factor, to understand how these risks might play out in the energy sector. Risks from over automation as seen with the two Boeing plane crashes in 2019, or the algorithmic racial bias of Google's Gemini AI tool are examples of the types of risks that could easily carry over into the energy sector with damaging consequences.

The cascade risks from AI and automation for example could arise when EV penetration reaches a tipping point in terms of load that can be switched on relative to local grid capacity. The deployment of multiple algorithms, primarily designed to ensure the lowest costs to their consumers, rather than grid stability, could create conflicts and overload the grid causing cascading network faults. Some of these risks are more near-term than others which may become risk scenarios once we reach a tipping point regarding decentralisation of the energy system and penetration of AI.

The focus in this report has been on where the use of AI in a singular or interconnected setting actuates risks on the energy networks. This could be through enhancing known risks, introducing new potentially unforeseen risks, changing the risk profile in certain applications, or rendering existing mitigation strategies ineffective or in need of adaptation.

The stakeholder engagement and desk research highlighted eight core areas where there could be significant challenges and emerging risks in adopting AI;

1. AI strategy & Visibility
2. Collaborative AI Approaches
3. Organisational culture
4. National security & AI Assurance

5. Data & Modelling
6. Skills for AI and related disciplines
7. Consumer Protection
8. Regulation & Governance

Managing AI risk

As critical national infrastructure, the energy sector is traditionally risk adverse, and risk management for electricity networks needs to be rigorous and robust. A serious fault on a network can lead to extensive blackouts and loss of supply.

AI risks have not been considered in detail for power systems, but the effects of automation and AI are starting to be considered. One of the first areas of this with respect to electricity networks is the consideration of risks associated to the design of autonomous systems. Whilst early work is ongoing in this area, our research highlighted significant knowledge gaps. As part of this project, we considered an investigation into some of the risk management frameworks that are, or could be, potentially applied to AI. We focused on three general risk frameworks, computer security, MLOps and Safety Engineering.

Regulation and Governance

The future energy sector is going to need a revised regulatory framework from what currently exists to ensure that the market operates as intended, maximising the benefits to consumers, whilst supporting the adoption of new technologies that will enable and support the sector's rapid pivot to Net Zero. Existing regulation and governance procedures will need to be reviewed and updated to account for the new system's operating environment, technologies and outcomes. We argue that this work is best done proactively to set the future direction for the industry and engage stakeholders in a collaborative process to co-create regulation that will support the market conditions that we will need.

Recommendations and next steps

Our research and analysis have outlined the scale of the challenge in designing and delivering an AI-enabled future energy sector, as we move towards a more decentralised and automated energy network. The challenges identified in this report represent an opportunity for the sector to take a proactive and collaborative approach to AI risk management, creating a regulatory market that nurtures the innovations that will support rapid decarbonisation, efficiency improvements and new market entrants.

The following recommendations and proposals for future areas of work, expanded upon in Section 8, will require engagement and validation from the wider stakeholder community to ensure buy-in.

FOCUS AREAS FOR FUTURE WORK		RECOMMENDATIONS
1. Culture & Skills	A. Create cross-industry consensus on AI risks	Initiate a cross-sector AI special interest group
	B. Leverage the engineering safety culture to embed an AI enabled safety culture	
	C. Skills training and requirements review	
2. Systems Modelling	D. Data sharing and availability	Develop a sandbox testing environment
	E. Developing an outcomes-based AI roadmap and strategy	
	F. AI risk foresight mapping and planning	
3. Regulation & Governance	G. Map the regulatory gaps and priority areas	Create best practice AI risk guidance
	H. Regulatory oversight and enforcement	

About Energy Systems Catapult

Energy Systems Catapult was set up to accelerate the transformation of the UK's energy system and ensure UK businesses and consumers capture the opportunities of clean growth. We are an independent, not-for-profit centre of excellence that bridges the gap between industry, government, academia and research. We take a whole system view of the energy sector, helping us to identify and address innovation priorities and market barriers to decarbonise the energy system at least cost.

2. Introduction

2.1 Overview

The energy sector is undergoing a rapid transition as we move from an analogue, centralised fossil-fuel driven system to a digitised, decentralised zero carbon system. Recent advancements in the field of artificial intelligence and related technologies will accelerate this transition, as barriers to entry are lowered and access to the technology is opened to more practitioners, accelerating its adoption into the energy system and creating new opportunities for decarbonisation and energy efficiency.

The UK Government has demonstrated national leadership on the world stage with the adoption of the UK 's National Strategy for AI[1], a 10-year plan to make the UK an AI superpower and harness the benefits of this technology for all areas of society. The AI Safety Summit[2] in 2023 and the establishment of the AI Safety Institute[3] at the heart of government demonstrate an understanding of the balance that must be struck between opportunity and risk. These national level endeavours set the culture and the framework from which industry specific action should follow to create sector specific guidance and playbooks.

However, the debate to date within the energy sector has largely focussed on the opportunities for AI, absent the current and future emerging risks that this rapid adoption of a new technology could pose to the energy sector, the UK National Grid and Net Zero targets.

The need to explore these adoption risks is pertinent given the rapid technological developments in AI and increasing deployment on to our critical national infrastructure. The energy sector should seek to take a proactive stance in setting the right market conditions to enable the appropriate balance between optimising for innovation and safeguarding against legitimate security concerns.

It must be stated that this report does not seek to downplay the economic, social and environmental opportunities and benefits that AI can bring to a multitude of use cases in the energy sector, but rather recognises that effective risk management can unlock innovation and ensure that it meets the needs of all stakeholders in the ecosystem. It should be seen as a primer for a much wider conversation that needs to happen across industry, to address the concerns raised through the stakeholder engagement.

This report is part of a wider body of work by a cross-Catapult consortium funded by Innovate UK, looking at AI Preparedness across sectors. The cross-Catapult consortium comprises Energy Systems Catapult, Digital Catapult, High-Value Manufacturing Catapult, Cell & Gene Therapy, Offshore Renewables Catapult, Medical Discoveries Catapult and Satellite Applications Catapult. The views set out here do not represent the views of the

[1] https://www.gov.uk/government/publications/national-ai-strategy
[2] https://www.gov.uk/government/topical-events/ai-safety-summit-2023
[3] https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

other Catapults, only the views of the Energy Systems Catapult and those collated from stakeholder interviews (unless otherwise stated).

2.2 Objectives

This work seeks to explore how and where AI risks might present themselves across the energy system, framed through specific use cases. It seeks to understand how risk management frameworks can be applied to the energy sector and how to build good governance and regulation to create an innovative marketplace that exploits AI whilst mitigating potential downsides.

The goals of this report are threefold:

- Support the sector to utilise useful/effective AI in the safest possible way for those using the network
- Develop an evidence-based assessment of the potential risks and what their impacts and mitigations may be
- Identify a series of recommendations for government and key stakeholders to support the development of responsible and safe AI deployment in a pro-innovation regulatory environment

It is aimed at practitioners, decision makers, and researchers across the energy sector. Tackling the challenges raised in this report will require significant effort and resources from industry, academia, and government. For industry, it offers the opportunity for further engagement to help formulate the right market and regulatory conditions that will support AI deployment, decarbonisation and innovation. For practitioners and researchers, it offers new energy specific areas of research and collaboration. For the public sector, the report highlights gaps in our collective current thinking and efforts, offering a direction for future investment and attention. Early efforts now will yield multiple benefits by setting the correct culture and frameworks whilst AI is still in its relative infancy, compared to playing catchup once AI becomes a dominant force on our energy system. It is with this proactive mindset that the report should be viewed.

2.3 Scope and Methodology

As algorithms and AI become more widely used for automation by different actors in the energy sector, from flexibility aggregators to the system operator, there is an increasing risk that these automated systems will interact in a way that is detrimental to the whole system. In a highly connected system, decisions made to optimise narrow objectives may have unintended consequences on the wider energy network. Algorithms applied in different parts of the energy network may have conflicting objectives that lead to undesirable oscillations, creating suboptimal system behaviour (increasing costs for consumers). More concerningly, errors in one algorithm could trigger a reinforcing feedback loop that could lead to a wide scale system failure.

Currently, there is no assessment or understanding of the potential for cascading failures within the future energy system caused by multiple interacting AI technologies and applications. This also means the thinking around effective mitigations is very immature.

This report therefore seeks to advance our understanding by exploring and presenting various frameworks which may be the basis for identifying and managing high risk algorithm interactions. It then examines mitigations and contingencies which could form the basis for making future policy and best practice recommendations.

Whilst the scope is relevant to both the electrical and gas networks, the focus of the work will look at the electrical grid as this is where the risks are likely to most pronounced in the future due to the current direction of travel for electrification and decarbonisation.

To achieve these goals, we've focussed our research and stakeholder engagement around several core themes;

1. **AI use cases in the energy sector:** Where is AI currently or likely to be used?
2. **AI related risks in the energy sector and other sectors:** What are they and what are the possible impacts?
3. **Risk management for AI**: How do we manage and mitigate risk, what risk management frameworks can we adapt?
4. **Regulation and governance**: Where does responsibility lie, what does good governance look like?

A stakeholder map of key organisations and individuals involved in AI related projects and startups across the energy sector was developed, and from this a prioritised list of people was developed to contact for interview. A total of 25 people (See Appendix 9.2) were interviewed throughout the project using a semi-structured interview process, to gather as wide a range of views and perspectives on the risks of AI deployment, risk management, regulation and governance. In addition, Energy Systems Catapult held a webinar on AI for Decarbonisation: Energy System Flexibility[4] that yielded insights into participants concerns around the risks, that were included. Thematic analysis was then conducted, to cluster emerging perspectives and issues around AI risks, risk management frameworks, and regulation and governance. The thematic analysis and clustering of key risk areas and recommendations forms the basis of the following sections of this report, supplemented through the desk research and opinions of the Energy Systems Catapult team.

Key stakeholders involved in the research included;

- Government departments
- Industry associations
- Network operators
- Regulators
- Academics
- Digital technology consultants
- SMEs including flex aggregators and EV charging

[4] https://www.youtube.com/watch?v=vMc4-JD2-lk

2.4 Artificial Intelligence in the context of this project

There are several global approaches to, and definitions of, AI. For this report, we refer to The Alan Turing Institute (the UK's national institute for data science and AI, and a key partner on the ADViCE programme – see below) which defines AI as:

"The design and study of machines that can perform tasks that would previously have required human (or other biological) brainpower to accomplish. AI is a broad field that incorporates many different aspects of intelligence, such as reasoning, making decisions, learning from mistakes, communicating, solving problems, and moving around the physical world.[5]."

Artificial Intelligence for Decarbonisation's Virtual Centre of Excellence

Energy Systems Catapult is part of the Artificial Intelligence for Decarbonisation's Virtual Centre of Excellence (ADViCE) initiative, delivered in partnership with Digital Catapult and The Alan Turing Institute.

It has already published two reports on the AI decarbonisation opportunities;

1. AI for decarbonisation - Assessing the UK landscape for artificial intelligence and its use in decarbonisation
2. ADViCE: AI for Decarbonisation Challenges

ADViCE is part of the UK government's larger AI for Decarbonisation Innovation Programme, which accelerates the development of innovative artificial intelligence (AI) technologies for decarbonisation applications, to support the UK's transition to Net Zero.

[5] https://www.turing.ac.uk/news/data-science-and-ai-glossary

3. Current AI Landscape

3.1 Market analysis and insight, adoption

AI is at the forefront of the UK's strategic development priorities in science, innovation and technology, and its adoption is rising significantly across multiple sectors. It has been identified as a key enabling technology for decarbonisation efforts, particularly in its capacity to rapidly capture, manage and assess large datasets.

An analysis of many aspects of market insights and adoption have been explored as part of the AI for Decarbonisation's Virtual Centre for Excellence led by Digital Catapult, Energy Systems Catapult and The Alan Turing Institute (see Ecosystem and Challenges reports[6] for further details). However, some of the more salient points for the energy sector are repeated in the arguments below to highlight how AI can support the sector, and the opportunities it provides.

3.2 Main Applications

Up until relatively recently, the main applications of AI have been within the domain of time series forecasting, for variables such as renewable generation, electricity price and demand. Demand and generation forecasts have been deployed at the transmission level for decades but more recently they have been incorporating machine learning into their models, for example, National Grid ESO have used AI to produce a 33% improvement in the accuracy[7] of their solar forecasts. Further, due to the increased volatility of network load, there is also some consideration of moving away from single value point forecasts, towards probabilistic forecasts to better model the uncertainty of the system.

Another recent development has been the application of demand and generation forecasts for distribution network operators (DNOs) to enable more localised management and planning. With the move towards localised energy markets there will likely be a requirement for energy market forecasting tools soon.

Network operators also deploy AI for maintenance and anomaly detection. One of the main responsibilities of transmission and distribution network operators is the maintenance of their assets, and AI is used in condition monitoring and to help identify faults. Faults can be detected through application of AI on imagery data (e.g. from drones or helicopters) or from analysing the data from the equipment sensors.

Another main area AI is applied is within flexibility applications. There are many flexibility providers who are utilising forecasting and optimisation algorithms to understand when to schedule their flexibility assets such as storage or demand side response. This includes companies such as Modo Energy, Arenko, Habitat Energy, Kraken Flex, who are all considering AI/machine-learning within their offerings and platforms. A particular subset of AI applications for flexibility has been through scheduling of electric vehicle charging.

---

[6] https://www.turing.ac.uk/research/research-projects/advice

[7] https://www.nationalgrideso.com/news/eso-and-alan-turing-institute-use-machine-learning-help-balance-gb-electricity-grid

Organisations such as CrowdCharge, and OtaskiES deploy smart charging of connected EVs to reduce costs, emissions, and can be used to reduce the loads on the grid.

Energy suppliers are also providing products and services to their consumers and utilising the smart meter and advanced metering infrastructure to understand occupants and their preferences. This could involve offering new tariffs or paying them to play within different markets (see for example the recent ESO Demand Flexibility Service trial[8]).

Generators, especially for renewables, also use machine learning models to estimate their outputs. A notable example is Open Climate Fix who are using deep learning methods applied to satellite data to improve solar PV generation forecasts.

Finally, building management systems[9] are also deploying AI, either through utilising monitor and/or smart devices within the building, or utilising smart plugs which can make standard devices smart. Grid Edge is an example of the former and combines energy usage from within the building with other data sets, such as external weather conditions, to shift demand to times of reduced carbon and/or lower costs. Measureable.energy, utilise a smart plug which allows them to monitor the usage of a device but also control them so as to use them when electricity is cheaper, or in times of lower demand.

A whole host of individual problems where AI can be applied to energy problems (either currently or in the future) have been collected as part of the ADViCE programme and can be found on the Advice Challenge webpage[10].

3.3 Data

Since the Energy Data Taskforce[11] was published in 2019 there has been a move to *presumed open* datasets which has since been included within Ofgem's Energy Data Best Practice[12] guidance. In subsequent years, each DNO has been releasing energy data on their own hubs including network maps, substation demand data, and curtailment data. Most recently Ofgem has declared aggregated smart meter data[13] to be presumed open, providing many more opportunities for innovation since this is one of the most valuable datasets due to the increased network visibility it provides.

Despite this, there are still some limitations with smart meter data availability. The aggregated data is generated from the combination of a minimum of five smart meters, but most likely will be the aggregation up to a secondary substation feeder level which consist of well over 100 smart meters. Individual smart meter data, except from static (and increasingly dated) anonymised innovation trials, is largely inaccessible due to privacy

[8] https://www.nationalgrideso.com/industry-information/balancing-services/demand-flexibility-service-dfs
[9] https://www.iea.org/articles/case-study-artificial-intelligence-for-building-energy-management-systems
[10] https://es-catapult.github.io/advice-challenge/
[11] https://es.catapult.org.uk/report/energy-data-taskforce-report/
[12] https://www.ofgem.gov.uk/sites/default/files/2021-11/Data_Best_Practice_Guidance_v1.pdf
[13] https://www.ofgem.gov.uk/publications/decision-updates-data-best-practice-guidance-and-digitalisation-strategy-and-action-plan-guidance

concerns. Therefore, many consumer applications and products can only be developed by selected companies (typically suppliers).

There is also a challenge with ensuring all the datasets follow similar standards and therefore it is easy to understand, and connect the individual data sets. Different standards have been suggested, with Ofgem adopting the Common Information Model, for example, for their data related license requirements[14].

Open low carbon technology data is another challenge. There are gaps in the availability of known uptakes in technologies such as heat pumps, electric vehicles, storage devices, and residential PV, and there is even less available telemetry data, recording their demand/generation levels. These technologies will be a vital component of a future energy systems, and the data will be necessary in order to build realistic energy systems models, and to better understand risks and management strategies.

3.4 Deployment

The ADViCE programme did a study of which organisations were deploying AI in the energy sector. Although it cited evidence that the UK is ranked third in terms of AI readiness, *"The use of AI technologies is still limited to a minority of businesses, with only 2% of businesses piloting AI"*.

Further, utilising the Crunchbase platform and searching the terms *"AI"* and *"Energy"* returned 90 results/companies. The ADViCE report also stated that 70% were based in London, and the others in major cities including Manchester, Derby, Bristol, Edinburgh and Glasgow. Further 44% of companies were categorised as seed, followed by 24% as early-stage venture.

There are only small amounts of information about what specific AI is being deployed across the energy sector. This was also expressed by the stakeholders we interviewed which provides challenges in understanding exactly what the risks are, and how to manage them.

3.5 Current AI investment levels

Investment in AI for decarbonisation was also considered as part of the ADViCE Ecosystem report, which also includes a look at Agriculture and Manufacturing applications which are out of the scope of this work. For the energy specific applications there is many sources of investment. Innovation projects are funded through a number of sources including UKRI funding calls (notably the EPSRC), Strategic Innovation Fund (SIF), Low Carbon network Funds (LCNF) and the NIA/NIC projects. Many of these consider network-based projects looking at everything from smart fuses, flexibility control, smart local energy systems, and advanced forecasting methodologies.

[14] https://www.ofgem.gov.uk/publications/common-information-model-cim-regulatory-approach-and-long-term-development-statement

There is also the most recent BridgeAI[15] programme, and of course the AI for Decarbonisation Programme[16] which, in addition to funding the ADViCE project, has also funded several streams of AI projects across organisations working in the energy sector. Stream 3 has recently funded projects focusing on projects such as optimising behind-the-meter residential renewables, commercial fleet electrification, demand flexibility services, and solar forecasting using satellite data.

In the private investment space, the ADViCE report notes that *"Between 2020 and 2022, investment into cleantech experienced a 50% increase, totalling £945 million"*, and *"According to Crunchbase data, the total amount of equity funding raised by AI companies building specifically for decarbonisation and sustainability initiatives, to date, totals £405.3 million"*.

A more complete list of funding sources and investment in AI for decarbonisation can be found in the ADViCE reports.

3.6 Skills issues

Skills gaps are a major concern for many organisations, and this has been demonstrated through the interviews as part of this project and through the results of a small-scale survey[17] conducted through Energy Systems Catapult in 2023. This survey found that there are significant gaps in data science skills in the energy sector, specifically in coding, domain expertise and advanced techniques. At least 40% of respondents found it difficult to hire a data scientist with the required skills, and the departments are relatively new with (at the time of the survey) over two-thirds of teams having only formed within the last five years.

It was reported in the ADViCE Ecosystem report that in 2022 "40% of businesses developed AI inhouse; 40% purchased off-the-shelf solutions; 20% outsourced development of AI applications to external providers". This also suggests that many businesses may not have the sufficient skills or finances to fully deploy AI solutions internally, which may have implications for the understanding of the AI risks. More widely it means that a large percentage of businesses will have very little transparency regarding the algorithms that are being deployed, which may hinder risk management strategies.

3.7 Opportunities for AI

There are several applications that have not been deployed and/or have only been considered in innovation projects. The ADViCE programme produced a challenges report listing around 70 challenges[18] where AI could help impact decarbonisation, including an assessment of the maturity of research and testing for each problem. These individual tasks led into "grand challenges" and the way to help solve these will be explored in further work as part of ADViCE, including through working groups and a knowledge base.

---

[15] https://iuk.ktn-uk.org/programme/bridgeai/

[16] https://www.gov.uk/government/publications/artificial-intelligence-for-decarbonisation-innovation-programme

[17] https://es.catapult.org.uk/report/data-science-skills-in-the-energy-sector-survey-results/

[18] https://es-catapult.github.io/advice-challenge/

The three grand challenges where AI can play a significant role in the energy sector, as defined in the ADViCE AI for Decarbonisation Challenges Report are outlined below;

1. Unlocking Domestic Decarbonisation

Decarbonising homes requires changes to both heating systems and consumer behaviours in every home in the UK. Engaging consumers in that process, financing it, and delivering it at pace are all major challenges.

2. Enabling Net Zero Infrastructure

Electrification of heating and transportation, combined with increased renewables mean the UK needs significant expansion of our electricity networks. Delivering at the required scale – and pace – is a real challenge, with lots of renewable generation being held up due to delays or uncertainty in network connections.

3. Maximising Flexibility in Energy Networks

An electrified, high renewables future requires energy demand to flex, so users consume and store energy when the wind is blowing, and the sun is shining. This requires a radical change in how networks, markets and end users operate, which also requires an introduction of new technology, presenting a host of new challenges.

Lack of data is one of the major blockers to innovation but also presents an opportunity. Generative AI may be one way to fill in the gaps in the most important data sets, especially smart meter data, and low carbon technologies, such as heat pumps and electric vehicles. However, there must be sufficient data to train these models to ensure they have the veracity of nature and realism required so that users can trust the outputs from model and tests. For example, Centre for Net Zero are developing a Generative AI tool, Faraday, which creates household load profiles from their training on the Octopus Energy smart meter datasets. A key measure for their tool is fidelity[19] so that the outputs most likely resemble real household energy behaviour.

Validating synthetic data is another challenge since only those developing the tools have access to the original raw data. Therefore, there may also be an opportunity for AI developers to also create tools to test the validity and representativeness of the generative AI outputs.

There are opportunities to explore techniques such as privacy enhancing technologies[20] using AI techniques, such as federated learning and differential privacy, to enable the sharing of modified or anonymised datasets such as smart meter data.

Finally, data sharing can be improved by the utilisation of new infrastructure. For example, there has been a recent investigation into a data sharing fabric[21] to enable a "digital spine" through which defined governance roles and responsibilities will allow secure and interoperable data exchange. There are also further investigations into data markets for

[19] https://arxiv.org/abs/2404.04314
[20] https://royalsociety.org/news-resources/projects/privacy-enhancing-technologies/
[21] https://es.catapult.org.uk/project/digital-spine-feasibility-study/

energy. AI may be a useful tool to ensure that these platforms and tools operate fairly and efficiently.

Due to the relative immaturity of AI being deployed across the energy sector, there is also many potential risks, as has been investigated in this project. This provides many opportunities for exploring how AI can be used to identify, manage and mitigate these risks. In addition to helping identify the areas of highest risk and their impacts, there is opportunities to look at explanatory AI methods[22] to provide better understanding of the models which are deployed, especially with respect to the needs for power system and network operators. This should be explored in conjunction with regulators and government to better understand the key responsibilities and owners across the sector, and how to coordinate the reporting, and regulation of these risks.

[22] https://www.ibm.com/topics/explainable-ai

4. AI Risks & Use Cases

In this section, we outline examples of risks that have materialised in the energy sector where the risk profile could be altered by AI, and examples of AI risks from other sectors that could have parallel applications in the energy sector. It then sets out two use cases of cascade risks that could occur on the electrical grid because of AI.

4.1 AI Risks

The following examples are of incidents from the energy sector, or other sectors where AI or automated algorithms have been a leading factor and are demonstrative of the types of risks that could translate across to the energy sector.

Tacit collusion: The development of automated decision-making raises the potential for price collusion between companies and gaming the system. This could occur without any human involvement, just through AI algorithms working with each other. Such is the potential risk that Ofgem recently launched a consultation[23] for plans on rules for AI before publishing a formal framework later this year. There are similar parallels with the Demand Flexibility Service[24] (DFS) where households are encouraged to shift energy consumption at peak times to smooth out the peak. Around 1.6 Million households and businesses took part in the 2022/23 DFS. It was later found out that a very small minority of consumers anecdotally engaged in attempts to game the service. They tried to artificially inflate energy consumption through load-shifting prior to the saving period, to generate increased payments[25]. This within-day adjustment was removed for the 2023/2024 version of DFS.

Flash Crash: In 2010, the US stock market suffered a $1 trillion-dollar flash crash[26] that lasted for 36 minutes, causing some of the biggest ever drops in market indices, leading to a surge in trading volumes as people tried to protect their positions, or take advantage of the chaos. Whilst ultra-high speed algorithmic trading accounts for the majority of transactions, a subsequent investigation by the U.S. Department of Justice found that spoofing algorithms designed to distort the market were in part to blame[27].

Algorithmic pricing signals: The use of algorithmic pricing among retailers is not particularly new or innovative. However, if those algorithms don't have built-in limits or safeguards or come up against another algorithm that is designed to work against it, feedback loops can occur with interesting results. In 2011 two Amazon book sellers, profnath and bordeebook used automatic algorithmic pricing to set the price of each other's books relative to each other[28]. One set the cost of their books slightly higher than

---

[23] https://www.cityam.com/ofgem-energy-watchdog-to-consult-over-ais-risk-to-industry-collusion/
[24] https://www.nationalgrideso.com/industry-information/balancing-services/demand-flexibility-service-dfs
[25] https://inews.co.uk/inews-lifestyle/money/bills/power-rinsing-gaming-national-grid-energy-saving-scheme-british-gas-customers-2253691
[26] https://assets.publishing.service.gov.uk/media/5a7c284240f0b61a825d6d18/11-1226-dr7-crashes-and-high-frequency-trading.pdf
[27] https://en.wikipedia.org/wiki/2010_flash_crash
[28] https://www.michaeleisen.org/blog/?p=358

the other, whilst the other would set theirs marginally lower. The automatic mischief that went unnoticed by most, saw Peter Lawrence's "The Making of a Fly", an otherwise unknown biology text about flies that typically sold for $35, reach an astronomical price of $23,698,655.93 as the automatic algorithms outcompeted each other for a few days.

Automation: The aviation sector has embraced automation to make flying safer over the years with notable success. However, the two Boeing 737 Max crashes within 6 months of each other in 2019 brought the risks of automation into sharp focus. There is growing concern that there is an increasing reliance on automation on what are extremely complex systems, and that this reliance is creating problems of automation transparency and automation complacency. In automation transparency, the issue is the pilots not being educated on the latest systems, how they work, their operating parameters, etc. Automation complacency is when pilots' skills degrade because automation means they no longer need to use the skills they were trained for, and so aren't as easily able to recover systems when the automation fails or does not operate as expected.

Algorithmic Bias: Google recently released its Gemini AI tool but had to quickly pull the product after it was found to be generating images which were viewed to be racist and historically inaccurate[29]. The AI model had been trained to ensure a wide range of people were included in its results, presumably to counter previous examples of racial bias and stereotyping in previous AI models, but failed to account for instances where ranges of people should not occur. Google's stock fell 4.5% as a result, knocking $90Bn off its value[30].

Chatbots: The use of chatbots is one of the most common consumer-facing applications of AI and there are many examples of them going rogue or being manipulated, such as when a DPD chatbot was enticed into making up a poem about how DPD is useless[31]. What happens though when the chatbot functions as expected and provides a customer with incorrect information? In 2022 Air Canada's chatbot gave out incorrect information about the airline's bereavement policy and would not honour the reduced price until a Civil Resolution Tribunal found in favour of the customer, citing negligent misrepresentation and that the chatbot was part of the company's website and they were responsible for the information it gave out[32]. Elsewhere scammers used AI technology and WhatsApp to impersonate Greg Jackson, CEO of Octopus Energy to scam customers[33].

National Security: In May 2021, ransomware attackers infected the US Colonial Pipeline's digital systems, shutting it down for several days[34]. To date, it's the largest publicly

---

[29] https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical
[30] https://www.forbes.com/sites/dereksaul/2024/02/26/googles-gemini-headaches-spur-90-billion-selloff/
[31] https://www.theguardian.com/technology/2024/jan/20/dpd-ai-chatbot-swears-calls-itself-useless-and-criticises-firm
[32] https://bc.ctvnews.ca/air-canada-s-chatbot-gave-a-b-c-man-the-wrong-information-now-the-airline-has-to-pay-for-the-mistake-1.6769454
[33] https://utilityweek.co.uk/ofgem-warns-of-ai-discrimination-in-energy-sector/?regwall=success&advance_login=success
[34] https://www.techtarget.com/whatis/feature/Colonial-Pipeline-hack-explained-Everything-you-need-to-know

disclosed cyber-attack against critical national infrastructure in the US, meaning there are likely other incidents that have not been disclosed. The pipeline which moves oil from refineries to markets affected other industries such as aviation, as oil could not be moved. In the end, Colonial Pipeline had to pay the hackers to regain control of their systems. AI systems can increase cyber security risks by creating new avenues for attacks (e.g. large language models can be used to produce prompt injection attacks[35]), or by manipulating the effectiveness and reliability of the AI models themselves (say by modifying the training data).

System Instabilities: The rapid increase of solar photovoltaics installations in Germany created new risks on the networks. The complications arose from regulation which required *"immediate shut-down of the PV inverter if the grid frequency should at any point in time reach or exceed 50.2 Hz."*[36] This was appropriate when levels of connected PV were relatively low, but as installations increased this created a massive network imbalance and system instability. This example highlights the potential unintended consequences of regulation but also the requirement for whole systems impact assessments. It also shows the need for the development of appropriate testing environments which can consider as many of the real-world conditions which occur or will occur in the future. The future mass deployment of AI systems on the energy network could exacerbate and create similar types of instabilities and frequency issues by rapidly propagating synchronised control actions. For example, the smart control of EVs could lead to issues of ramping on the network[37].

4.2 Use Cases

Throughout the project research and stakeholder engagement, several real and hypothetical use cases where significant AI risks may present themselves in the future have emerged. Some of these are more near-term than others which may become risk scenarios once we reach a tipping point regarding decentralisation of the energy system and penetration of AI. The use cases presented here are the most common use cases where we feel significant AI risks may present themselves.

4.2.1 Co-ordinated EV charging

The number of electric vehicles on the road in the UK is increasing rapidly, forecasted to rise from ~1 million in 2024[38] to 12-26 million in 2035 according to the 2021 Future Energy Scenarios[39]. A large proportion of these will be charged overnight using domestic 7kW chargers. Smart charging control systems aggregate EVs and respond to half-hourly time-of-use tariffs to provide customers with either lowest cost or guaranteed cost of charging which creates steep ramp rates from EVs all coming online at once. Project REV[39] (Resilient Electric Vehicle Charging) estimated that just 2% of these chargers switching on at the

[35] https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you-need-to-know
[36] https://www.dnv.com/cases/the-german-50-2-hz-problem-80862
[37] https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you-need-to-know
[38] https://www.smmt.co.uk/2024/02/uk-reaches-million-ev-milestone-as-new-car-market-grows/
[39] https://es.catapult.org.uk/report/resilient-electric-vehicle-charging/

same time would generate a load step of between 1.7 and 3.6 GW, significantly more severe than the August 2019 loss of supply incident[40].

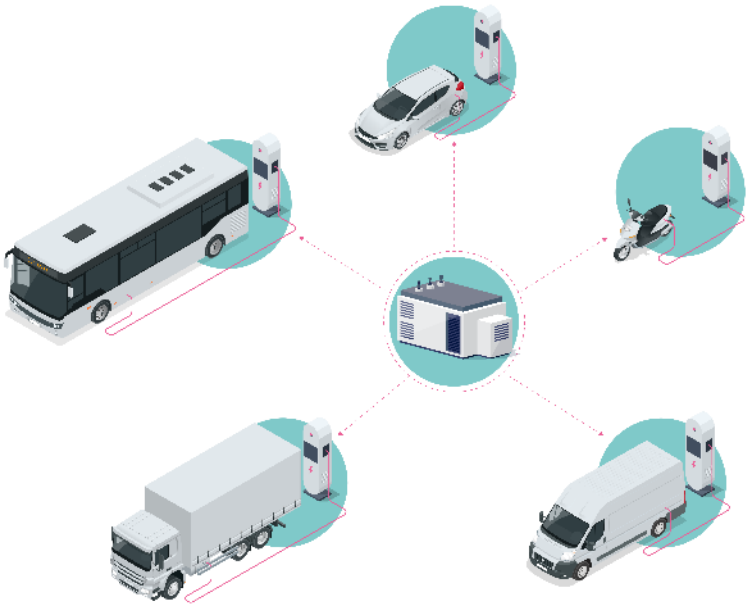

Figure 1. Schematic illustrating the diversity of electric vehicles which could theoretically be connected to a single local substation in the near future.

Products such as Intelligent Octopus Go[41] help smooth the curve from these assets coming on at once by spreading the charging load as most EVs only need to be charged for part of night. However, this is only one product currently controlling a relatively small load, albeit one that is growing rapidly with 150,000 EV batteries under management as a virtual power plant that could in theory power two UK cities such as Leeds and Birmingham for an entire evening[42]. A cascade risk arises when EV penetration reaches a tipping point in terms of load that can be switched on relative to local grid capacity. The deployment of multiple algorithms, primarily designed to ensure the lowest costs to their consumers, rather than grid stability, could create conflicts and overload the grid causing cascading network faults. How these networks interact with markets, and how those markets are designed (e.g. volume-based pricing where aggregators book in capacity at a fixed price) such that the right economic signals are communicated to participants to ensure grid stability and resilience is still an under-explored area. This could be resolved with expensive storage assets, or flexibility assets from some of the bigger aggregator providers who have access to assets with enough capacity.

[40] https://www.ofgem.gov.uk/publications/investigation-9-august-2019-power-outage
[41] https://octopus.energy/smart/intelligent-octopus-go/
[42] https://www.businessgreen.com/news/4207810/power-birmingham-leeds-octopus-passes-1gw-ev-battery-management-milestone

"The focus of EV charging / V2G technology design is customer needs and cost; it will do "just enough" to meet grid-related regulations such as fault ride through and high/low voltage withstand. Present regulations were not designed for a zero-carbon future so will need revision."[39]

There are six scenarios highlighted in the Project REV report where EV charging could create risks for the grid security. These are;

1. Step: Too many chargers switching on or off at the same moment.
2. Ramp: Too many chargers switching on or off within a few minutes.
3. Oscillations: A group of chargers switching on and off.
4. Degraded stability: Increases risk of post-fault collapse.
5. Demand control: Network defences are eroded.
6. Restoration: Erratic behaviour after restart will hinder the process of restoration.

4.2.2 Network Management & Price Signalling

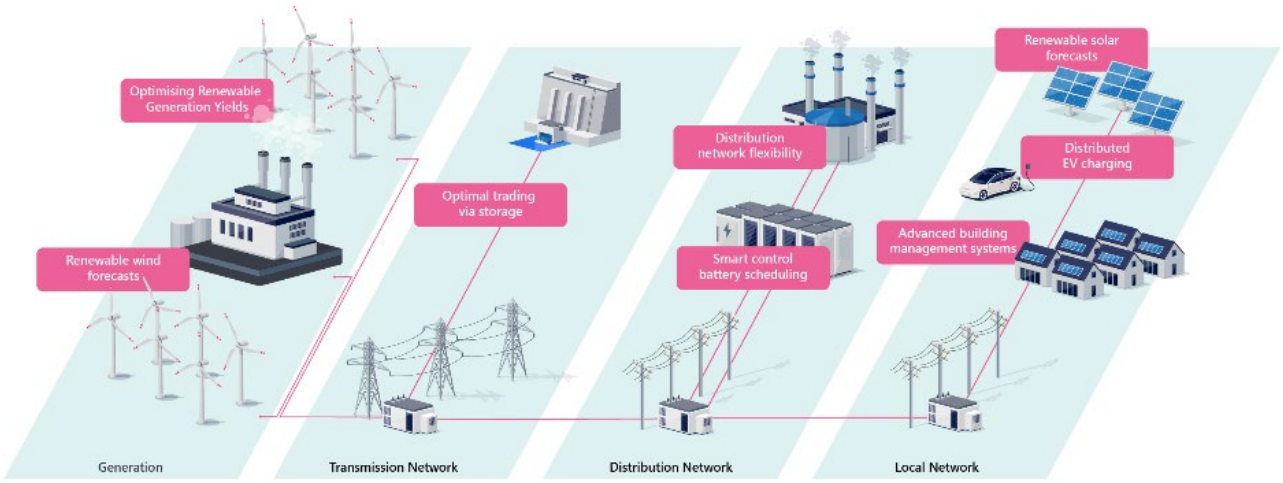

Figure 2. Illustration of the future electricity network, including areas of possible applications of AI, from Generation down to individual homes.

The increased electrification of transportation and domestic home heating, driven primarily through EVs and heat pumps uptake respectively, are going to create significant new loads onto an already constrained local and national grid network. This is turn is going to drive a significant increase in the numbers of smart devices controlling how these assets consume energy (EVs, batteries, heat pumps etc) and feed into local supply (solar panels), see Figure 2. The expected penetration of heat pumps into domestic heating settings could require significant local grid reinforcements to deal with flexible demand[43].

[43] https://www.regen.co.uk/wp-content/uploads/Building-a-GB-electricity-network-ready-for-net-zero.pdf

Flexible demand asset aggregators will utilise smart assets to provide services to the grid, supporting liquidity and system stability, whilst maximising income for their customers based on price signals. AI-based interoperable solutions will play a key role in automation, to balance the complexities of supply and demand in real-time. Whilst larger providers are required to be part of the balancing mechanism with dedicated communications to manage this, smaller providers aren't required to do this as it is seen as cost prohibitive. However, should smaller providers proliferate in the same way that smaller energy suppliers did in the last 10 years, this could create significant flexible aggregator assets that aren't part of the balancing mechanism but still have the collective capacity of larger providers.

The complex interaction of millions of these devices, with many interdependencies and confounders, driven by highly stochastic variables (e.g. price and weather) could potentially lead to cascade risks on the network. Multiple devices trying to optimise to similar drivers can lead to compounded and potentially unpredictable responses. For example, scenarios could feasibly arise whereby future Distribution System Operators (DSOs) publish price schedules which influence various AI-based flexible demand aggregators to shift all demand to one period, overloading the network in those periods. In response this may cause DSOs (perhaps through AI driven algorithms) to shift their price schedule to account for the response to the new demand pattern that they see, at which point the AI-based flexible demand aggregators moves all the demand into the new slot, potentially causing stability issues for the grid as it tries to control the power flow, frequency and voltage with all these competing assets.

5. Challenges of AI Adoption

In the coming years we are going to see high levels of AI algorithms deployed onto the energy system in potentially high-risk spaces. This creates a known unknown, where we know that we are potentially introducing singular or cascade risk fault points but without a detailed understanding of the risks themselves. We may not know the impact of the risks, how they propagate across the system, or the best mitigation strategies to get the network back up and running.

It should be stated from the outset that the use of AI does not automatically increase the risk profile to the energy sector, and whilst it may accentuate the risks or change the risk profile, it may not be the direct cause. There will be many instances where the use of AI and automation may reduce risks for example, from human error in choosing the wrong setting or dispatching the wrong load, or better utilisation of weather forecasting. AI is just a new tool that we can use to solve problems and we need to carefully balance the need to accelerate its adoption into the energy system so that it delivers on its promise, whilst ensuring a smooth transition that minimises harm.

Therefore, the focus is on where the use of AI in a singular or interconnected setting actuates known risks in a new way, introduces new potentially unforeseen risks, changes the existing risk profile in certain applications, or render existing mitigation strategies ineffective or in need of updating to account for a change in conditions.

The following sections present a synthesis of the key issues raised during the stakeholder engagement and desk research.

5.1 AI Strategy & Visibility

We are in the "inflated expectations" period of the AI hype cycle, firmly buoyed by the phenomenal success and adoption of generative AI tools such as ChatGPT that has brought the power and potential of AI applications to the forefront of public awareness. Across the energy sector, there is an increasing understanding of the opportunities that AI presents for the energy sector and its various components and stakeholders, as demonstrated by Energy System Catapult's previous work through ADViCE amongst many other industry initiatives.

Depending on where you sit in industry, there is a wide range of experiences and AI readiness, and the energy sector currently lacks a coherent direction of travel for how and where we would like to see AI deployed across the energy network. The general sentiment heard across industry is that barring some rapid transformational shift in a short period of time, that we're quite a long way off to really being able to develop and deploy AI algorithms at scale across the National Grid, DNOs, etc.

The ESO Digitalisation Strategy Action Plan[44] published in December 2023 sets out the future direction for ESO to become a Digital Leader and drive collaborative digitization of the whole energy system (Figure 3). This includes a principle of being "AI Driven" and

[44] ESO Digitalisation Strategy Action Plan

details of a Generative AI Deep Dive. The step from Modernise Tech to Digital First and beyond, outlined in "RIIO-2 (Revenue = Incentives + Innovation + Outputs) BP2-3 period 2023-2026 ("Interoperability and resilience across the energy system is possible through greater digitalization") creates an immediate opportunity to define the AI space driving what it means for the sector.

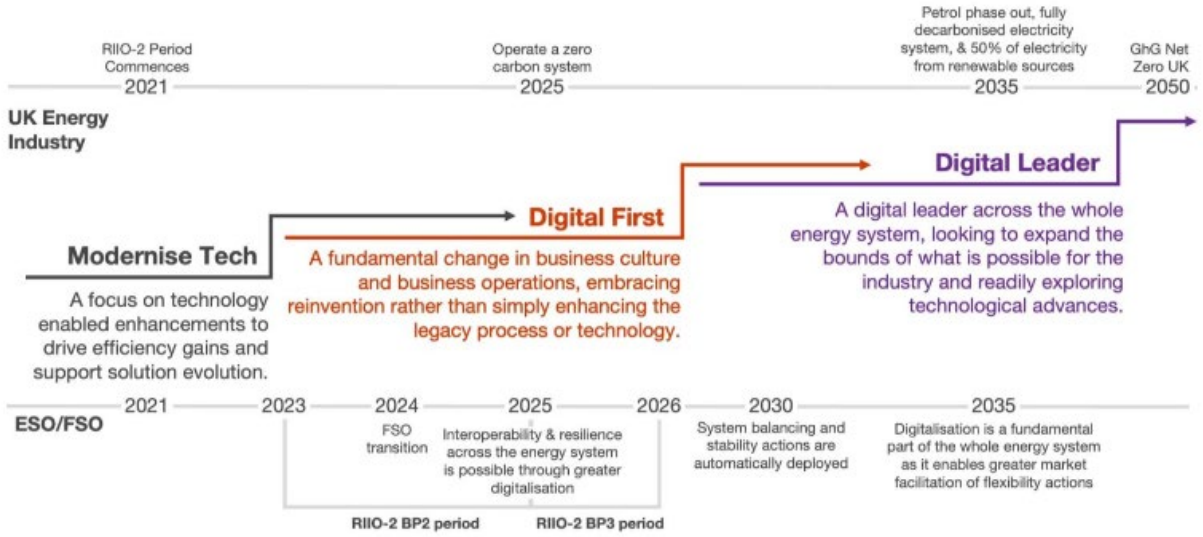

Figure 3. ESO Digitalisation Strategy Action Plan timeline of digitalisation transitions[44].

Across the sector, there is concern that there is a lack of awareness of where AI is being deployed and for what purpose, and that therefore emergent risks that might arise from these deployments are not even recognised, never mind being managed and mitigated. Further it was noted that whilst everyone was talking about deploying AI, there was a lack of a common definition of AI and if they were deploying ML or some other command functions under the umbrella of AI. You can't manage what you can't see and at present the current lack of visibility of AI deployments into different aspects of the energy system are hampering efforts to understand the current and emerging risk profile and develop effective mitigation strategies.

We also need to define what we mean by AI, and what it should do. Whilst programmes such as ADViCE have set out around 70 use cases, there needs to be a greater understanding of what we want AI to do for the energy sector to guide the industry, as opposed to a long list of all the potential use cases.

There's a lack of understanding between what's needed to develop a proof-of-concept to demonstrate the value, to deploying it safely and robustly in the system. There is a risk that without a burden of proof on a model's ability to safely deploy, that incorrect or poorly modelled algorithms with weak safety mitigations are already interacting with the energy system.

5.2 Collaborative AI Risk Assessment

There is a necessary and understandable risk averse approach in industry that has been rightly developed over the years as part of the wider safety culture. However, this safety culture has not necessarily been translated to our thinking on AI as our understanding and

appreciation of the risks is immature at present. We are currently preparing to deploy AI onto systems and infrastructure that has not been designed for this type of technology and we risk some negative unintended consequences. We need to be proactively thinking about, and preparing for, individual and cascade risks now to avoid getting exposed by some unforeseen catastrophic event in the future.

The vast majority of thinking about AI and cascade risks is happening in isolation by individual organisations and not across the sector. Despite a recognition of the increasing potential for cascade risks to occur in the future, it remains a theoretical risk and therefore not near-term, and there were few identifiable research studies looking at their potential and impact. Currently the sector has a blind spot when planning for and mitigating network cascade effect risks, and what little work is being done does not have enough cross-sectoral visibility. There is the potential to learn from other sectors such as finance and aviation where risk management frameworks and approaches are more advanced and embedded into the sectors, to learn how cascade risks could be modelled and mitigated.

Risk is an inherent feature of any inter-connected system and cannot and should not be designed out, as to do so would inhibit overall functionality, efficiency and innovation. The question that the energy sector will have to grapple with, is what level of risk are we willing to accept in our future interconnected system? That question can only be answered once we have a clear picture of the risks at a practical modelled level where likelihood and impact can be assessed.

The energy system is becoming increasingly complex in response to the need to decarbonise and increase flexibility. With this increased complexity comes an increased risk that 'normal accidents'[45] will occur due to the very nature of the system and an inability to model for every scenario or outcome. The deployment of AI models onto an increasingly complex system does however raise the possibility of epistemic accidents[46], those that occur when a model behaves exactly as its designers intended but comes up against misunderstood parameters or limits of the system that it is being deployed on.

Initiatives such as the Accelerator for Systemic Risk Assessment[47] have developed a set of Principles for Systemic Risk Assessment and Response[48] to accelerate awareness of systemic risks and to guide collective action, that could inform approaches in the energy sector.

5.3 Organisational Culture

As we move from a centralised to a decentralised energy system, our organisational and leadership structures need to reflect this decentralised approach and foster greater cross-sector collaboration and learning. The challenges, opportunities and risks of the future

[45] https://en.wikipedia.org/wiki/Normal_Accidents
[46] https://research-information.bris.ac.uk/files/168151260/Epistemic_Accidents_JC_Chapter.pdf
[47] https://www.asranetwork.org/
[48] https://assets-global.website-files.com/64650eb651139221ff45c0c3/660fe30f9fba10fe06b9f744_ASRA_Principles_2024.pdf

inter-connected system can't be realised by organisations working in siloes and this will necessitate a significant culture shift across the sector.

Although many organisations in the energy sector already have strong backgrounds in risk management due to the nature and dangers inherent in operating energy networks, the risks of AI are relatively new and not fully understood. It may require some cultural changes and ways of thinking to properly operate. This includes:

Systems thinking: Interdisciplinary teams and skills will be required to properly deploy and manage AI systems, but a wider whole systems view will be needed to identify all the risks from an increasingly connected interdependent energy network.

Horizon Scanning: The domain of AI and the associated risks are likely to continue to change rapidly as we learn more about these systems and how to appropriately manage them. It is therefore necessary for actors across the energy sector to try and keep up to date with the changing landscape and look at adjacent sectors to learn from their approaches. This should also consider the possible external regulations and legal requirements.

Community building: Technical and non-technical domain experts can help identify risks, but also there should be the building and engagement with the wider AI risk community and stakeholders. There may be the need for further review boards to help better understand risks and how to mitigate them.

Internal culture: AI risks and management need to be understood across the organisations from the technical staff creating the algorithms and tools, to the C-suite determining the direction and focus of the organisation.

AI sufficiency: It can seem desirable to solve everything with AI, however, depending on the application and risks it may be inappropriate to use AI for certain problems. Assessments may be required to understand whether AI is the most appropriate way to solve a particular problem, or part of a problem. It may be better to keep a "human in the loop" rather than fully automate.

5.4 National Security & AI Assurance

Security of AI is a wide-ranging topic that can't be adequately dealt with in this project but needs to be acknowledged. When we talk about security, we are addressing many facets, from national security of critical national infrastructure, AI assurance and safeguarding, and to consumer protection (discussed in section 5.7). The cross-sector challenges highlighted present a risk of delivering critical national infrastructure that lacks the resilience to achieve an equitable net-zero energy system.

National Security

Critical national infrastructure is defined as systems that are necessary for a country to function, of which the energy system is one of 13 in the UK, and whose loss or compromise

could impair national security[49]. At the time of writing, the UK is the third most targeted country for cyber-attacks behind the US and Ukraine[50].

The risks to the energy sector are wide ranging and beyond the scope of this report but include:

- Targeted cyber-attacks to knock out parts of the system.
- Backdoors being installed on chips sourced from other companies and nations.
- Malicious data or algorithms being deployed to disrupt supply.
- Injecting bad data to skew the market or falsify forecasts that AI algorithms depend on.

As we deploy new technologies and systems across our inter-connected system, the attack vectors from bad actors increase along with the potential for cascade effects across the system if one part is affected.

AI Assurance and Safeguarding

There are significant risks with AI assurance due to a lack of skills and capacity to understand the models that we are building. This risk increases as the power (and lack of transparency) of the models increases and the ability to control them diminishes. AI safety is a rapidly emerging field, particularly at the national level in the UK but needs to translate down to the energy sector. Current tools and techniques to mitigate the risk of AI systems are limited and there has been little R&D into approaches that can provide quantitative safety guarantees for AI systems, which is a major omission given that this is happening on critical national infrastructure.

We will need to develop the tools that allow us to verify and control these models and early work is beginning in this area with the Safeguarded AI[51] programme from ARIA. This £59m programme will seek to develop the safety standards needed for transformational AI and such fundamental research should also include critical national infrastructure such as the power grid.

5.5 Data & Modelling

Data is the fundamental fuel for AI models that require extensive datasets to train their algorithms to reflect real world scenarios. Yet in nearly every interview, several risk areas around data were raised that need consideration:

Data quality and availability: AI models are only as good as the data that they are trained on. Poor data infrastructure results in either poor models with poor outcomes and optimisations, or an inability to deliver AI at all. Significant risks exist around both the quality and availability of data sets for companies to train their AI models on, both at an organisational and sector level, with further differences between traditional and greenfield organisations.

---

[49] https://www.npsa.gov.uk/critical-national-infrastructure-0
[50] https://www.ncsc.gov.uk/files/NCSC-Annual-Review-2022.pdf
[51] https://www.aria.org.uk/programme-safeguarded-ai/

The energy sector is highly fragmented with differing levels of digital maturity and data readiness creating an uneven baseline from which to work from, both from the perspective of current data collection, but also historical data availability which can be used to train on a wider range of scenarios. Business confidentiality is often cited as a reason to protect individual organisations data sets which whilst understandable, when presented as a blanket response inhibits appropriate and best practice sharing of data. It is often unclear what data sets are currently being used to train AI models, and whether these were being extrapolated to create new data sets which could compound errors and biases from initial poor quality data sets. This then raises the questions of whether a model has reached the edge of its capabilities, what operating scenarios are outside of its ability to function, and how does it deal with those scenarios (by deferring to a human operator or carrying on as normal and potentially break something)?

Data sharing: To make any kind of AI model work and predict something with decent accuracy it needs lots of sufficiently representative data, but the energy sector has been traditionally hesitant to share data across organisations and more openly. Data is seen as proprietary, an asset to be protected or sold. This reduced system visibility between organisations and increases the risks of cross-boundary conflicts. There has been significant progress in this area following the development of Ofgem's Data Best Practice (DBP) and the utilisation of the *presumed open* data principle. Most recently as a results of a DSP consultation (August 2023) aggregated smart meter data has been designated presumed open enabling the sharing of more granular consumption data whilst protecting consumer privacy. DNOs have already started to release this data[52, 53].

There are strong parallels between data sharing in the energy sector and the local authority public sector that has gone through a similar digital transformation, in terms of breaking down silos, building trust between organisations to make data sharing the default, and standardising data-sharing agreements to facilitate and speed up the process. Paraphrasing Eddie Copeland, Director of the London Office of Technology and Innovation, *"If you're not ready to share data with the public sector, then you're not ready to work with the public sector."* As we move towards an inter-connected data-driven energy system this should become the default approach to working across the energy sector.

Data modelling: Many AI models are often referred to as "black boxes" because of the inability to see and interrogate the internal mechanisms and transparently assess how the inputs relate to the outcomes. There's a lot of commercial confidentiality bound up in these models which the manufacturers are very reluctant to breach, but from the perspective of the system operator there are justifiable concerns about how they will operate, particularly on scenarios that may fall outside of the data training set. Inference of how AI models will behave and how they will interact with each other is becoming increasingly important due to their numbers, scale and potential impact on the system.

[52] https://ukpowernetworks.opendatasoft.com/pages/smart-meters/
[53] https://www.ssen.co.uk/news-views/2024/SSEN-first-DNO-publish-smart-meter-half-hourly-consumption-datasets/

Requiring some form of certification to demonstrate meeting a minimum standard was raised as a potential mitigation, though needing to certify every algorithm would require significant resources and potentially act as a brake on innovation. A possible middle-ground would be to have transparency on the decisions that will be made by that model so that if it did start performing outside of expected norms, then there would be a way to conduct some initial analysis. There are understandable commercial concerns around this that would need to be addressed, and it needs to be determined if current regulatory powers are sufficient to require a company to disclose the inner workings of an AI model which could be deemed as commercially sensitive IP. A balance will need to be found where AI models can be stress tested and interrogated if needed (potentially requiring new AI assurance tools to do this) and maintaining a pro-innovation approach.

Data validation: Data, especially streaming data, may change (suddenly or slowly) over time (also referred to as "concept drift"). This can invalidate models, and their outputs. AI can be used to automate the checking of datasets, especially useful for large numbers of data sources. Opening data sets up for training algorithms is imperative if models are to be both correctly trained in the first place, but also for retraining to account for system evolution, or to develop new models when previous models become outdated or obsolete. Further, without data on fringe use cases and near misses, there is an inherent risk that AI models cannot be trained effectively on rare occurrence events, and therefore may inhibit incorrect behaviour in these instances, by operating exactly as programmed and trained for the most generalisable use cases only. In these fringe use cases AI should only be used for augmentation, supporting control engineers with domain system knowledge.

Data ethics: There are additional risks created from the push to increase data availability for AI modelling around consumer acceptance, privacy concerns and GDPR. As we increase data sharing, consumers normally have little choice but to accept terms and conditions which may cover use of their data by the contracting party and potential third parties. A lack of understanding of how their data is being used could risk a consumer backlash, limiting the uptake of new, AI-driven technologies, regardless of the benefits they may bestow.

There are also risks associated with ensuring that the training data provided has limited and/or identifiable biases which can be mitigated for[54]. Currently available smart meter data is likely skewed towards certain groups/demographics and excludes others, for example, those on credit meters. This creates a built-in bias which if implemented across modelling algorithms will only further entrench the bias that we currently see, leading to unfavourable social outcomes such as not delivering an equitable and affordable supply to all customers.

[54] https://es.catapult.org.uk/report/data-ethics-and-bias/

5.6 Skills for AI and Related Disciplines

At a practical level, this is perhaps one of the easiest risk areas to understand and mitigate as it relates to easily identifiable gaps with known remediations. There are multiple key areas in the skills agenda which need consideration including:

Sector Leadership: The skills debate is an area that generates widely differing opinions, in part because no-one person or organisation wants to admit that they lack the skills to keep up with or set the agenda. This creates a risk of bias towards over-stating skills and understanding, instead of curiosity and leadership to truly understand what is needed at both an organisational and sector level. This risk appears to be borne out by the current lack of AI specific strategies and guidance, along with an absence of risk management frameworks that have been developed specifically to meet the needs of the energy sector. This should not be read as a criticism but an acknowledgement of perceptions across industry, and a recognition that with a new technology that is seeing incredibly fast adoption rates, a degree of urgency is necessitated to upskill and fill the leadership void. This is particularly crucial to enable collaborative working to understand inter-connected cascade risks and the fragmented nature of skills across industry was raised as a concern.

Interdisciplinary teams: Data science and AI teams need to be closely connected to risk management teams to ensure competing perspectives are considered and factored into model development. It's one thing to have the skills to develop your models and another to understand the potential implications and how they might impact across a much wider inter-connected system. In addition, AI deployed at scale in the energy sector will likely interact and affect multiple systems, including communications networks, transport networks, and multiple energy vectors. Teams which have previously been focused on, for example, the electricity networks alone will find they require knowledge and experience in multiple areas.

Human oversight: Automation should support and augment existing decision-making processes and not override the human in the control room. Many interviewees acknowledged that many AI systems will require some degree of "human-in-the-loop" especially in applications where risks and their impacts may be more pronounced.

There is also concern that automation creep could occur with the sector losing institutional knowledge and core competencies as we lean into relying on automation. Regular training will be necessary to ensure upkeep and retaining of human skills and proficiency for when they are required.

Recruitment: At an organisational level, being able to recruit for and retain data science, machine learning and other related disciplines is proving increasingly challenging[55], particularly for energy network and public sector organisations who are competing against the private sector where salaries are higher. This presents the risk of an imbalance between energy network organisations and traditional private sector organisations, where the energy network organisations are not able to effectively manage the interactions into the

[55] https://es.catapult.org.uk/report/data-science-skills-in-the-energy-sector-survey-results/

system, or where organisations are not able to collaborate effectively due to differing levels of digital maturity. It was suggested by some interviewees that the AI use cases for traditional private sector organisations working in the energy sector were generally more obvious than those for energy network organisations, and therefore people would be more attracted to organisations where they could see how their skills would be applied.

There is also a potential risk about skills availability in discouraging new entrants into the market and seeding innovations. Larger companies such as Octopus who are well known and actively champion their AI credentials and products find it easier to attract and retain talent compared to say, newer startups.

Retaining skills & knowledge: The skills debate also reaches into the skills and knowledge to truly understand how the energy system works. As automation creeps in, we risk losing institutional knowledge that originally would have been passed down by working on the system itself and creates a knowledge gap around managing fringe events that will be very hard to fill.

Domain knowledge: An additional risk that could arise as AI applications proliferate, is when newer entrants into the market lack an energy systems background. To apply AI safely means understanding the system where they will be deployed. Domain expertise is essential to avoid unintentional consequences and understand the potential interdependencies and core risks. Companies that have been working in the sector for a long time will have an inherent domain expertise and be cognisant of how their models may interact with existing systems and design in effective mitigation strategies. Newer entrants who may lack such domain expertise could risk introducing new risks simply through a lack of understanding of how their AI model might interact beyond its single intended application use. In addition, a lack of domain knowledge may also reduce the effectiveness of the automation systems themselves since knowledge of the underlying system and its characteristics are often key in developing high performing AI models.

Specific Risk Expertise: As AI systems become more ubiquitous and complicated, there may be a requirement for specific AI risk engineers with skills tailored to the deployment of advanced algorithm and machine learning. These people will need to know what questions need to be asked to enable remediation of issues.

5.7 Consumer Protection

When we talk about AI applications, it's easy to overlook risks to consumers as so many of the use cases are intended to benefit them.

AI models require large datasets to be trained on, and as the energy system changes, updated datasets to be retrained on, making data increasingly valuable and necessary for new actors seeking to enter the market. This raises GDPR and consent risks regarding how consumers data is used and potentially sold to third parties for them to train their own models. Consumers are likely to have little choice but to accept data usage terms and conditions with little to no understanding of how their data may be used. Ideally consent should demonstrate awareness of what you're consenting to, be voluntary and revocable, particularly as the power of AI tools and what they are being used for progresses.

Ofgem recognised the consent principle and the need for consumers to be able to share their energy data securely with trusted markets participants, as presented in their *"Call for Input – Data Sharing in a Digital Future"*[56] published in January 2024.

The use of imperfect and unrepresentative consumer data raises further risks about embedding biases in the AI models and serving customers with impartiality. This was picked up by Ofgem in a call for input on the use of AI within the sector[57];

"Done well, AI has the potential to improve service to consumers and identify when unfair outcomes could occur. Done poorly, AI has the potential to exclude certain customers causing discrimination and inequality, create bias resulting in disturbances in the marketplace to the detriment of customers or result in collusive processes or reducing the stability of the marketplace."[58]

The benefits to consumers need to be explained to encourage data-sharing consent, and to learn the lessons from the smart meter rollout. If we are to reach our net zero and decarbonisation targets, effective engagement of consumers will be critical, and we can't afford a backlash from a misuse of data or a lack of transparency of the risks and benefits. This risk links regulatory needs for ensuring that there is a clear process to seek redress should there be negative impacts on consumers. At present, if a consumer is unfairly impacted because of an AI algorithm, the process for that impact to be addressed and who would be held accountable is quite opaque from a consumer perspective.

5.8 Regulation & Governance

Appropriate regulation and governance are key to encourage and enable the safe deployment of AI in the energy sector. This will be discussed in detail in Section 7.

[56] https://www.ofgem.gov.uk/publications/data-sharing-digital-future
[57] https://www.ofgem.gov.uk/publications/use-ai-within-energy-sector-call-input
[58] https://utilityweek.co.uk/ofgem-warns-of-ai-discrimination-in-energy-sector/

6. Approaches to Managing AI Risk

As critical national infrastructure, the energy sector is traditionally risk adverse, and risk management for electricity networks needs to be rigorous and robust. A fault on network can lead to extensive blackouts[59] and loss of supply. The electricity and gas networks feature strongly in the National Risk Register[60] with major risks including:

- Failure of National Electricity Transmission
- Regional Failure of the electricity network, and
- Failure of the gas supply infrastructure

However, these are by no means the only energy-based risks, and there are some extensive risks, especially from AI, which are not covered within the NRR. A recent paper has reflected on the limitations of the NRR[61], in that it doesn't take a whole systems perspective, and that there may be "(non-risk) precursor contexts and events which might amplify many acute risks and/or nullify planned responses". The paper also highlights that risks may be co-dependent on other conditions which may increase their chance of occurring. This is particularly relevant to the power systems since many assets and applications are driven by markets and/or weather despite not being directly dependent on each other.

The report also notes that risks can be conditionally dependent on each other which may increase the chances of cascades. This is also relevant to power systems since networks are naturally highly connected and hierarchical. With increased connected devices used to enable flexibility, there is a chance for cascading, domino effects which propagate through the grid.

The recommendation from the report is the need to consider pairs of risks and second order interactions, and thus an important requirement is a whole systems approach to risk identification and management.

6.1 Current Risk Considerations in Power Systems

AI risks have not been considered in detail for power systems but of course risk is still a key component of network management. In terms of non-AI risk management there is various standards that the network operators must abide. For example, National Grid consider Security and Quality supply standards[62] for managing and planning the network. Any AI

---

[59] https://www.ofgem.gov.uk/publications/investigation-9-august-2019-power-outage

[60] https://assets.publishing.service.gov.uk/media/64ca1dfe19f5622669f3c1b1/2023_NATIONAL_RISK_REGISTER_NRR.pdf

[61] https://www.researchgate.net/publication/378697497_The_National_Risk_Register_2023_Some_Reasoned_Reflections

[62] https://www.nationalgrideso.com/industry-information/codes/security-and-quality-supply-standard-sqss/sqss-code-documents

implementation would likely have to ensure that there are no conflicts. More generally anyone who wishes to generate, distribute or supply electricity must comply with various technical codes and standards[63].

The effects of automation and AI are starting to be considered. One early area of focus with respect to electricity networks is the consideration of risks associated to the design of autonomous systems. For example, there is the IEEE P7009 Working Group which considers fail-safe design of autonomous systems[64]. These focus on safe functional design and operation, such as identifying fail-safes to prevent harm from autonomous systems. In the UK there are already some regulations[65] on the use of EV smart charging points[66]. The regulation is relevant for private EV charge points and smart cables, and covers, smart functionality, interoperability and requirements for cyber security standards with the scope being extended as per a 2022 consultation[67].

6.2 Risk Frameworks

As part of this project, we considered an investigation into some of the risk management frameworks that are, or could be, potentially applied to AI. In particular we focused on three general risk frameworks, computer security, MLOps and Safety Engineering. We describe a brief summary of each of these frameworks in the following sections.

6.2.1 Computer and Cyber Security

Many approaches exist in the areas of cyber security and computer security, which although are not a focus area of this report do share some of the same features, especially around connectedness and complexity. Therefore, there are some principles which may be particularly relevant for AI safety in energy networks.

Two common computer security frameworks are STRIDE[68] which focuses on sources of risk such as authorisation and non-reputability (confirming whether someone was or was not responsible), and DREAD[69] which focus on the size of the potential impact from a risk including the number and type of affected users.

Attack trees[70] are also primarily used in cyber security, but they could also be used for root cause analysis more generally, mapping out potential risks, interactions and their causes.

---

[63] https://www.ofgem.gov.uk/energy-policy-and-regulation/industry-codes-and-standards/standards/technical-standards
[64] https://sagroups.ieee.org/7009/
[65] https://www.legislation.gov.uk/uksi/2021/1467/contents/made
[66] https://www.gov.uk/guidance/regulations-electric-vehicle-smart-charge-points
[67] https://assets.publishing.service.gov.uk/media/6425a2d23d885d000fdadfc0/smart-secure-energy-system-government-response.pdf
[68] https://en.wikipedia.org/wiki/STRIDE_model
[69] https://en.wikipedia.org/wiki/DREAD_(risk_assessment_model)
[70] https://www.ncsc.gov.uk/collection/risk-management/using-attack-trees-to-understand-cyber-security-risk

Google also has a cyber security framework SAIF[71]. A useful concept which could transfer to AI risk management is the red teaming concept, where team members can pretend to be an adversarial organisation to test the safety and risk management of a system or company.

Cyber security, the risks and the mitigations are covered extensively in the literature, so this is not the focus however, many of the risk principles apply to other applications and sectors, and they also align with the pro-innovation principles from the UK government AI safety summit.

6.2.2 MLOps and AI-pipeline Based Frameworks

Another set of frameworks and approaches we found focus on the pipeline of developing and deploying AI, and therefore provide practical identification of where and what type of risks may enter different parts of the process, from collecting data, to training the algorithm.

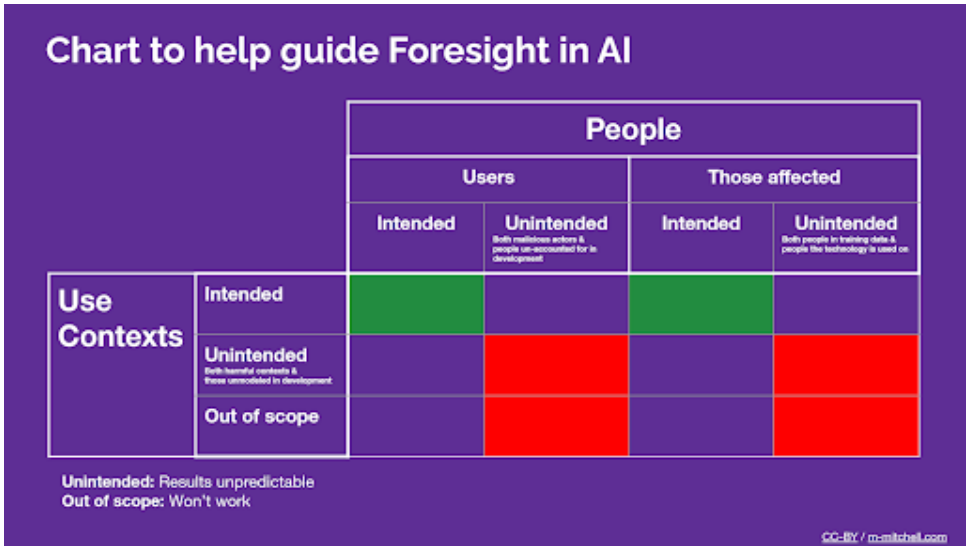

Figure 4. Model Cards[72] approach for users and use contexts. Figure from Hugging face[73]

One approach developed by Margret Mitchell, is the "Pillars of Rights"[74] focusing on the four "pillars" of a machine learning pipeline:

- Data Collection: Who produces the data, and who is represented
- Training Process: Who develops the models, number of parameters, etc.
- Model Evaluation & Analysis: How is the model evaluated, is it interpretable, how long does it take to compute
- System Deployment: what it effects, and what are the potential harms.

[71] https://safety.google/cybersecurity-advancements/saif/
[72] https://huggingface.co/docs/hub/model-cards
[73] https://huggingface.co/blog/ethics-soc-2
[74] https://www.techpolicy.press/the-pillars-of-a-rightsbased-approach-to-ai-development/

Different approaches can be taken at different stages, for example, at the model evaluation stage harms in terms of intended and unintended consequences can be considered in terms of people and use contexts (see Figure 4).

A similar breakdown of potential risks and risk controls is considered in a report by McKinsey[75]. Here they also consider an additional initial step "conceptualization" which considers the misuse and unethical use cases from its initial conception. There is a focus also on the unintended consequences for organisations, and in particularly mentions interaction issues, and the need for gap analysis in existing risk management frameworks to identify needs for improvement or enhancements.

One of the most comprehensive materials on AI risk is via the National Institute of Standards and Technology (NIST), which has been developing an AI risk Management framework[76], including a Taxonomy and Terminology of attacks and mitigations[77]. This also considers the AI lifecycle, which overlaps many of the above frameworks looking at application context, data, the AI model and deployment. They echo many of the same principles as noted in the cyber security frameworks, including accountability, transparency, fairness, explainability, and safety.

NIST have developed a Map-measure-manage AI risk framework in addition to a governing function:

- Govern: Concerns the development of a culture of risk management within organizations. This enables the other functions of map, measure and manage.
- Map: context to frame risks of an AI system. Whether AI is appropriate or warranted. Understand limitations in processes, identifying constraints in real-world applications.
- Measure: employs techniques, and methodologies to analyze, assess, benchmark, and monitor AI risk and their impacts.
- Manage: Prioritizes risks and allocates risk management resources to map and measure risks.

The report also considers cyber security attacks which directly affect the algorithm. Hence although cyber security isn't a focus of our report, these particular risks can create secondary effects such as cascades by reducing the integrity of the model. For example, they mentioned data injection attacks, data poisoning attacks (which modify training data samples), and model poisoning where the model and/or its parameters are affected.

These frameworks are useful for classifying the different functional elements of the ML/AI pipelines and therefore simplify the task of isolating some possible areas of risk. However, they don't necessarily identify the system wide risks which may not be present within any

[75] https://www.mckinsey.com/capabilities/quantumblack/our-insights/confronting-the-risks-of-artificial-intelligence
[76] https://www.nist.gov/itl/ai-risk-management-framework
[77] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

single element of the pillars. Further they do not strongly consider the interconnected risks, or their potential for cascades.

6.2.3 Safety Engineering Frameworks

Safety engineering frameworks focus, as the name suggest, on how to ensure the safety for engineering systems. For example, self-driving/autonomous vehicles, nuclear power plant safety, offshore oil and gas, spacecraft design etc. Frame works that exist in this category include

- STPA: Systems Theoretic Process Analysis[78]
- FMEA: Failure Mode and Effects Analysis[79]

FMEA provides a top-down, step-by-step process for identifying failures in a design, and the potential impact of those failures. Information is collected on the different ways something could fail, the causes of failure, as well as recommended actions. These components are all rated which are combined to provide a ranking of risks. A limitation of FMEA is that by breaking the problem into components some risks could be missed which do not arise from the failure of a single component, but from multiple components.

In contrast STPA takes a systems approach but considers a control problem rather than failures. System boundaries, losses and hazards are first defined, and then a control process is modelled. The next step is to identify which of the control actions are unsafe and identify the causes which could lead to unsafe control actions.

There have been some considerations to apply these safety engineering frameworks to machine learning algorithms. One paper suggests these two approaches could be complementary, with FMEA investigating particular machine learning processes or components (e.g. a model, or set of data), and STPA could be used to look at specific risks and how the algorithms could be integrated within a larger system[80].

6.3 Evaluation of Risk Management Frameworks and Future Needs

The frameworks explored above all have useful features and procedures for the development of a future risk frameworks for those deploying AI on power system, including the use cases focused on in this report. The cyber security provides many principles such as transparency, accountability, and fairness which are vital for producing safe outcomes and limiting risks. They also align with the principles outlined in the recent governments pro-innovation approach to AI regulation.

The MLOps frameworks breakdown the risk identification approach according to the main components of the AI development pipeline process, from data collection down to

[78] https://www.gate.energy/the-brainery/stpa
[79] https://asq.org/quality-resources/fmea
[80] https://dl.acm.org/doi/fullHtml/10.1145/3544548.3581407

deployment. However, these approaches do not necessarily consider interdependent risks across the systems or potential cascade effects. In contrast, we found that safety engineering frameworks, in particular FMEA and STPA do offer complementary bottom-up and top-down approaches, and therefore a whole systems approach for identifying risks, in particular interdependencies, and looking for causes which may not be risky events themselves.

It therefore appears that the principles from all three approaches are worthwhile in minimising risks, but the whole systems approach from safety engineering frameworks appears to be the most promising and perhaps should serve as the foundation for further investigation. They are already applied in many engineering applications, and therefore could be easily understood within current processes, and assimilated with the power systems engineering culture. The testing of frameworks such as outlined above could be one use case for the sandbox which we include in our recommendations in Section 8.5.2.

7. Regulation & Governance for AI

The future energy sector is going to need a revised regulatory framework from what currently exists to ensure that the market operates as intended, maximising the benefits and increasing personalised requirements to consumers, whilst supporting the adoption of new technologies that will enable this transition, supporting the sector's rapid pivot to Net Zero. Existing regulation and governance procedures will need to be reviewed and updated to account for the new system's operating environment, technologies and outcomes. We argue that this work is best done proactively to set the future direction for the industry and engage stakeholders in a collaborative process to co-create regulation that will support the market conditions that we will need.

This section looks at regulatory and governance for AI in the energy sector, including frameworks, principles for regulation and governance, and the current regulatory landscape.

7.1 Regulatory Frameworks for the Future Energy Sector

There's a strong awareness that the industry is probably not sufficiently mature in understanding and managing AI risks in the energy sector, in part due to a greater focus on the bigger challenge of decarbonisation, but also due to a lack of skills and capacity to properly address and understand AI risk issues. This is not viewed as an immediate challenge and lacks the prioritisation that is needed to think far ahead on a speculative issue. Regulators will need to address potential market mechanism vulnerabilities since future AI applications could exploit market failures quicker and more cost effectively than humans currently can. There is a concern that unless this is prioritised soon that the sector will be playing catch up. A lack of awareness and visibility of what is actually being deployed on the system also inhibits action as you cannot regulate what you cannot see, and an initial landscape review is needed to understand where the regulatory gaps are[81].

The UK Government has adopted a pro-innovation approach to AI and this ethos should filter down into the emerging regulatory space for AI in energy and set out the type of market that we want, choosing where we want to let people compete, using reward mechanisms where appropriate and avoiding dictating, for example, what types of models or technologies could be used. Aligning incentives to desired outcomes was highlighted as an existing constraint preventing the development of flexibility opportunities at the local distribution level in contrast to the national markets. Crafting regulation to encourage DNOs and private sector companies to collaborate with a whole system view to support the stability of the grid would be welcome.

Regulation and governance should be proportionate to risks. Therefore, a key question arising from the research and stakeholder discussions concerned how to assess the actual levels of risk from AI and how much do we need to start scenario planning and red teaming to ensure that it doesn't present too much risk in the future? The

[81] Note at the time of finalising this report Ofgem announced a timely call for input into the use of AI in the energy sector: https://www.ofgem.gov.uk/publications/use-ai-within-energy-sector-call-input

acknowledgement of qualitative risk and the need for safeguarding measures highlighted the absence of quantified risk measurement and that without such knowledge, any attempts at regulation could be premature and inhibit innovation and deployment. This is particularly true for smaller providers where onerous regulation could inhibit market entry. However, should the number of small flexibility/aggregator providers proliferate, then they could create a bigger risk than those posed by some of the larger providers due to the cumulative assets under their management. We need regulation and governance that operates effectively for a decentralised system where risks may arise from the co-ordination of actors rather than from the actions of individual actors. This also points to the need for a dynamic regulatory framework approach that can adapt, evolve and retest assumptions as the energy system evolves. It also creates a focus point for stakeholders where they can raise risks they see in the system, to bring accountability into the system. Beyond regulation, according to an AI Powerplay event by Osborne Clarke, "*Businesses in the energy sector will need to ensure they are alive to very novel legal and regulatory issues that will emerge over the coming years*[82] and ensuring that they can demonstrate how they are effectively managing risks. This raises further questions around algorithmic governance and the need to proactively demonstrate compliance before methods are deployed on the energy grid. The Energy Digitalisation Taskforce recommended a Register of Algorithms[83] to mitigate these risks, which could be combined with some form of certification, though this would need to be balanced with the pro-innovation approach mentioned earlier.

7.2 A Pro-Innovation Approach

A lot of the sentiment around regulation is centred around the risks that AI poses to society, and its potential for misuse. These sentiments are understandable, and risks and misuse should be addressed. However, a balance needs to be struck between the need to innovate, develop new solutions and fulfil the promise of AI, with a regulatory framework that nurtures fair competition alongside the mitigation of unintended consequences, malpractice and general risks. The UK government has taken a pro-innovation approach[84] to AI risks driving the recent Bletchley Declaration[85] on AI safety at the recent AI safety summit. This also considered the risks and capabilities[86] from frontier AI systems such as generative AI.

The focus has been on *outcomes* to avoid specifying exactly how these should be managed. The core principles of Pro-Innovation document are:

1. Safety, Security & Robustness – risks should be continually identified, assessed and managed.
2. Appropriate Transparency & Explainability

[82] https://www.osborneclarke.com/insights/how-navigate-emergence-ai-energy-sector
[83] https://es.catapult.org.uk/report/algorithm-governance/
[84] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach
[85] https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park
[86] https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf

3. Fairness – should not undermine the legal rights of individuals or organisations, discriminate unfairly or create unfair market outcomes.
4. Accountability & Governance – ensure effective oversight, with clear lines of accountability.
5. Contestability & Redress – an AI decision, which is harmful or creates material risk of harm, should be contestable by users, affected third parties, and actors across the AI lifecycle.

Alongside these principles, the UK Government will be exploring the use of innovation sandboxes and whether regulation should be considered at a sectoral level or other

In addition, to support the implementation of the core principles the Department for Science, Innovation and Technology (DSIT) released an introduction to AI assurance including potential ISO and IEEE international standards that could be adapted[87]. The AI Assurance mechanisms fall into the following brackets:

- Risk assessment.
- Algorithmic impact assessment: wider effects on environment, and data protection.
- Bias audit: Unfair bias in input data, and the outcome of decisions.
- Compliance audit: adherence to internal policies, external regulations, and legal requirements.
- Conformity Assessment: meets requirements prior to being placed on market.
- Formal Verification: formal maths methods and proofs to verify if system satisfies specific requirements.

There also shared a portfolio of AI assurance techniques which presents case studies from a variety of sectors and applications.

However, as noted by Faculty AI[88], AI Assurance is still under explored from a regulatory perspective in that we don't know enough about how these models work even at this stage, and that this issue will only grow as the power and complexity of AI models grow;

"How do we build the technologies and the tools that allow us to interrogate, understand and control these models? There's a gap in the regulatory space for mandating that the people who are building deep foundation models, are investing their time and resources into this research and in proportion to the amount of money they spend on building the underlying technology. That would help them work according to government standards and make sure that we have the technological means to stay in control."

John Gibson, Chief Commercial Officer of Faculty AI

In April 2024, the UK and US, signed a Memorandum of Understanding[89] to work together to develop tests for the most advanced AI technologies and models, building on

---

87 https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf

88 https://faculty.ai/blog/artificial-intelligence-can-it-be-regulated/

89 https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety

commitments made at the AI Safety Summit in November 2023. It's early days in this area of AI regulation, but one that the energy industry should shadow and learn from to ensure safety and assurance of AI models as they are increasingly deployed in more complex environments across critical national infrastructure.

7.3 Role of Regulation in AI Risk Management

There are multiple ways in which strong regulatory and governance frameworks play a critical role in risk management that should be considered when developing and deploying AI technologies and algorithms. The following where some of the roles of AI regulation raised by stakeholders in our investigation:

1. **Risk identification**: Identifying where potential risks might occur through detailed risk assessments and scenario modelling enables regulators to address impacts on various stakeholders and propose regulatory or governance mitigation strategies.
2. **Standards and requirements**: Establishing standards and requirements for the development, deployment, and use of AI systems, covering safety, reliability, security, privacy, and ethical conduct, ensures that AI meets quality and performance benchmarks.
3. **Compliance and enforcement**: Regulatory bodies need to have the powers and capabilities to enforce relevant legislation, mitigating risks due to non-compliance, privacy breaches or unfair practices.
4. **Mitigation of harms**: Frameworks to provide mechanisms to mitigate potential harms associated with AI technologies, including safeguarding measures to prevent misuse or abuse. They should also cover means of addressing negative outcomes of unintended consequences.
5. **Best practice dissemination**: By encouraging transparency, accountability, fairness, and ethical conduct, regulators help mitigate risks associated with irresponsible or unethical use of AI technologies.
6. **Monitoring and adaptation**: The AI landscape is moving rapidly, and regulations and frameworks need to be monitored and adjusted to address emerging risks and challenges to remain effective and relevant.

7.4 Principles for AI Regulation and Governance

When considering building the emergent frameworks that will govern the new energy sector, DSIT released some guidance for implementing the AI regulator principles[90]. In addition, the following principles for regulation and governance for AI were raised by stakeholders that closely aligns with the core principles in the UK's pro-innovation framework:

1. **Ethical AI**: Regulation and governance frameworks should promote AI systems that adhere to basic principles regarding ethics, fairness, transparency, accountability,

[90] https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety

privacy and human rights. Ethical guidelines such as the EU's Ethical Guidelines for Trustworthy AI and the OECD's AI Principles provide frameworks for ethical AI development. Within the government framework these aspects will likely fall under the core principle of fairness.

2. **Transparency**: The AI models being deployed need to demonstrate openness and transparency in their operations, what they are optimised for and the decision-making process that governs them. This should cover the data that they were trained on, and future data that might be needed to retrain as the energy system evolves, their capabilities and limitations to understand how and where they can be relied upon, particularly about extreme use cases.
3. **Accountability**: Organisations who develop and deploy AI models should have a duty of care and be able to be held accountable for their models' actions and outcomes, including singular and cascade failures. Governance should look at mechanisms for understanding who is responsible, addressing said harms, and mechanisms to allow those who are harmed to seek redress.
4. **Bias and non-discrimination**: Although bias can never be eliminated, AI models should be deployed with minimal bias, without discriminating against different demographic groups, and actively promoting non-discrimination. Linking to transparency, this involves demonstrating how bias has been reduced, ensuring that bias within the data training sets is identified or mitigated for any pre-existing structural inequalities. If necessary, outcomes should be calibrated with respect to any harmful biases against individuals or demographic groups.
5. **Safety and reliability**: AI systems should be safe, reliable, and robust against errors and failures. This includes establishing standards for testing and certification of AI systems, implementing safeguards to prevent unintended consequences, and developing protocols for managing risks associated with AI technologies.
6. **Privacy and data protection:** Regulation should ensure that an individual's privacy and personal data is protected. AI models need to comply with relevant data protection laws and regulations, including GDPR, with consent mechanisms in place for data use, and safeguarding measures to protect sensitive information.
7. **International standards and collaboration**: AI technologies are by their nature global, and models deployed on the UK system may have been developed by international companies. Regulatory frameworks should promote best practices, harmonised frameworks across diverse markets, and create collaborative governance mechanisms to facilitate dialogue.

In addition to the above principles for deploying AI systems, there is a need for anticipatory risk planning to help forecast scenarios to avoid unforeseen circumstances. Planning control and mitigation measures to bring systems back online is seen as imperative in our collective approach to managing future risk. Such approaches will not eradicate risk but should seek to mitigate the most obvious and impactful.

Nesta set out six principles for anticipatory regulation that the energy regulator could emulate[91]:

1. Inclusive and collaborative
2. Future-facing
3. Proactive
4. Iterative mindset
5. Outcomes-based
6. Decentralised experimentation

Inclusive and collaborative working is critical to enable a mapping of responsibilities enabling a wide range of stakeholders to contribute to future market design, creating engagement and buy-in from a wide range of actors. Startups were wary of potentially being made to declare every algorithm, or for being responsible for grid stability, noting that current regulation doesn't encourage them to optimise for this.

[91] https://www.nesta.org.uk/project/anticipatory-regulation/

8. Recommendations and Next Steps

As discussed in the introduction, the energy sector should seek to take a proactive stance in setting the right market conditions to enable the appropriate balance between optimising for innovation and safeguarding against legitimate security concerns. These recommendations therefore should be seen as a primer for a much wider conversation that needs to happen across industry, to address the issues raised through this research.

Our research and analysis have outlined the scale of the challenge in designing and delivering an AI-enabled future energy sector, as we move towards a more decentralised and automated energy network. It has highlighted the emerging trends and use cases, with the emergent risks and subsequent potential impacts on the network, in what is a rapidly developing space. The very nature of the problem of interconnected systems and risks necessitates industry-wide collaborative initiatives that bring stakeholders together to develop a shared understanding of the risks and learning around mitigations that can guide the sector into the future.

These challenges whilst significant, also present an opportunity for the sector to take a proactive and collaborative approach to AI risk management, building on encouraging foundations. The aim should be to create a regulatory market that nurtures the innovations that will support rapid decarbonisation, efficiency improvements and new market entrants, whilst smoothing the future risk profile and preventing a hard-handed reactive approach.

It is with this proactive ethos in mind that the proposals for future areas of work and recommendations have been developed. Drawing from insights from stakeholder engagement, they represent a synthesis of where we feel future action is most warranted to address some of the challenges raised, and where collaborative stakeholder engagement could have the most impact.

Proposals for future areas of work and recommendations have not yet been tested across industry and so will require engagement and validation from the wider stakeholder community to ensure buy-in and support. Energy Systems Catapult as a neutral mediator in the energy sector may be well placed to lead the industry engagement to shape the next phases of work. As a first step, we would propose organising a roundtable with stakeholders engaged as part of this project to validate the recommendations and gain consensus and prioritisation on the proposed next steps.

This may develop into different working groups and projects with diverse sources of funding as part of a wider programme of works. Potential sources of funding include Innovate UK, Strategic Innovation Fund (SIF), Pathfinder, and Horizon. More work is needed to identify likely sources of funding that has been outside the scope of this project once the recommendations are agreed.

Post-development of our focus areas and recommendations Ofgem released a call for input for the use of AI within the energy sector. Conveniently, the proposed recommendations align with our recommendations, including the call for an AI Forum (in their case an AI best practice cross industry forum), and for the need for producing guidance. There is also a recommendation for addressing regulatory issues using tools

such as an AI sandbox. The intention will be for ESC to respond to this call for input using the information we have collected as part of this project.

8.1 Focus Areas for Future Work

FOCUS AREAS FOR FUTURE WORK		RECOMMENDATIONS
1. Culture & Skills	A. Create cross-industry consensus on AI risks	Initiate a cross-sector AI special interest group
	B. Leverage the engineering safety culture to embed an AI enabled safety culture	
	C. Skills training and requirements review	
2. Systems Modelling	D. Data sharing and availability	Develop a sandbox testing environment
	E. Developing an outcomes-based AI roadmap and strategy	
	F. AI risk foresight mapping and planning	
3. Regulation & Governance	G. Map the regulatory gaps and priority areas	Create best practice AI risk guidance
	H. Regulatory oversight and enforcement	

Table 2: Summary focus areas for future work and recommendations

8.2 Culture & Skills

8.2.1 Create Cross-Industry Consensus on AI Risks

A clear finding from the stakeholder engagement was that there was not a common discourse on what the key risks that AI might present, how they might manifest themselves and what the impact might be. Perceptions ranged from AI risks being minimal and not something that the industry needed to devote significant time to or prioritise, to those who saw it as an emerging threat to grid stability that needed to be better understood as a priority to enable mitigation planning.

The industry should consider establishing a sector forum that regularly reviews newly identified risks and prioritises key ones to be addressed and mitigated. Increased discussion and collaboration to agree a consensus on the scale of the risks would support the establishment of a programme to address required mitigations. Further, mapping key stakeholders and responsibilities, and how threats might affect them would advance the discourse and break through siloes to encourage different actors to work together on solutions.

8.2.2 Embed an AI Enabled Safety Culture

Building on the above, and as AI becomes more widespread in its adoption across the energy network, there is a need to reduce or remove the culture of secrecy and for increased openness and transparency around security issues, near misses, sharing testing and training data for algorithms.

The energy sector should seek to harness the strengths of the safety cultures embedded in the aviation and financial sectors as well as within its own engineering specialisms, as

examples of how to balance the need for commercial sensitivity with the need for openness and transparency to promote safety and improve learning and feedback loops. This should build on one of the recommendations in the Energy Digitalisation Taskforce[21];

Embed a digitalisation culture - Digitalisation is not valued or understood in all parts of the energy sector, with not enough skills or value given to digital assets and activities. BEIS (now DESNEZ) should employ a Chief Data Officer and importantly investors and the rating agencies need to value digital assets as well as their traditional value assessment for infrastructure.

A digitalised culture with the associated digital infrastructure can improve the sharing of data and resources, encourage the development of skills, and enable the dissemination and uptake of appropriate risk standards and management.

In particular, for the previously mentioned forum to be effective requires a better understanding of what AI is being applied and a greater visibility of AI risks and near misses. Therefore, one consideration for the cross-industry group would be to look at better reporting for AI and their associated risks which would create a tangible basis of evidence for the forum to focus on.

8.2.3 Skills Training and Requirements Review

The increasing penetration of AI across the energy sector necessitates upskilling, across areas such as baseline AI knowledge and understanding, to risk management and system interaction, and understanding and assessing AI tools will all be needed. This is particularly true at a senior decision-making level where a key understanding of AI technologies is needed to develop strategy.

There is evidence[92] that energy stakeholders are struggling to recruit and retain data science and related skillsets, due to competition from both within energy and other sectors, as well as a lack of understanding of the opportunities.

The Department for Energy Security and Net Zero (DESNEZ) should commission an industry skills gap survey and assessment to identify the gaps and the medium to long-term skills requirements for the industry, developing a plan to nurture these internally or be able to draw people in externally. This review could build upon the approach in the Government Digital, Data and Technology Profession Capability Framework[93] and connect with the skills work being conducted by the ESO AI Centre of Excellence[94].

8.3 Sector-Wide Coordination

8.3.1 Data Sharing and Availability

Whilst there has been good progress towards improving data infrastructure, availability and licensing, the data landscape is still patchy and there are limitations with some of the

[92] https://es.catapult.org.uk/report/data-science-skills-in-the-energy-sector-survey-results/
[93] https://ddat-capability-framework.service.gov.uk/
[94] https://www.nationalgrideso.com/news/future-eso-and-artificial-intelligence

types and quality of data which is available and the ability to share this across organisations. Key use cases include internal testing and modelling, the deployment of more advanced algorithm testing tools, and the ability to understand how and where risks might present themselves in the system.

The ESO's Digitalisation plan has set a clear direction of travel to create a digital first culture across the sector but will not be able to accomplish this alone. Cross-sector collaboration is needed to build trust and transparency to encourage all types of organisations to adopt an open data approach. There is much that can be learned from other sectors that have gone through similar transformations such as the local government sector in the UK, and the work of the London Office of Technology and Innovation in standardising data sharing agreements and reducing the time for data to be shared from months to a matter of weeks.

There has also been a move from sharing data through individual hubs to connecting data to those who need it. The recent investigations into the digital spine[95] (also referred to as data sharing infrastructure) as proposed in the Energy Digitalisation Taskforce, has demonstrated one such viable approach to helping improve data sharing within the sector. Further, the extension of the *presumed open principle* to aggregated smart meter data from February 2024 also promises to help support more granular modelling of the electricity networks.

8.3.2 Developing an Outcomes-based AI Roadmap and Strategy

The direction of travel towards a decentralised, more electrified, zero carbon, algorithmically dispatched energy system is clear. To ensure that the opportunities of AI are proactively communicated and harnessed the future National Energy System Operator (NESO) could, capacity depending, lead on co-ordinating the development of an AI specific roadmap and strategy to identify opportunities for future use cases and industry innovation funding opportunities. Whilst there is some emerging work happening in this area already, for example through the ESO's AI Centre for Excellence, there is a need for a cross sector understanding of where the industry is aiming for and what it would like to see developed.

8.3.3 AI Risk Foresight Mapping and Planning

As a sector we need greater definition and granularity on the emerging system level risks and scenarios. Building on the above AI roadmap, industry should review the emerging landscape and collaborate to explore potential use cases and associated risks from balancing services, flexibility, chatbots, etc. The development of potential future risk scenarios would enable a proactive and dynamic approach to risk mitigation.

[95] https://es.catapult.org.uk/project/digital-spine-feasibility-study/

8.4 Regulation & Governance

8.4.1 Map the Regulatory Gaps and Priority Areas

Regulation and governance for AI is a rapidly moving space as organisations get to grips with the technology's potential for good, but also misuse. There are differing schools of thought about how and when to regulate so early in the implementation cycle, and most of the discourse is currently happening at the international or national level with the passing of significant acts or legislation at the EU, US and UK levels.

These efforts will need to filter down to the energy sector level, starting with a sector mapping exercise to develop a clear understanding of where the responsibilities lie with different organisations, both now and longer-term for when AI becomes more prevalent across the system. This should include who is the owner of the risks for the AI algorithms, who is accountable for them, which organisations are responsible for monitoring, developing and testing algorithms, etc.

This could also include the consideration of market-making mechanisms to give organisations the confidence to invest, use and deploy AI for the energy sector. This might also require DNOs to publish their AI risk pathways alongside potential incentives to use AI across the networks.

8.4.2 Regulatory Oversight and Enforcement

A recurring sentiment throughout the stakeholder engagement was to ensure that as regulation is developed, that the regulatory bodies are provided the power, capacity and skills to oversee and enforce these regulations. Appropriate regulation that is well thought through and balances the need for innovation with oversight is needed and there is encouraging progress in this area through the work of the AI Safety Institute. However, without oversight and enforcement, there is a risk that regulations lose their intended impact, and new powers may be required to enforce AI regulations. This could be through the ability to interrogate the models to understand what they are designed to do, to test them in a sandbox prior to deployment or the requirement for certification or equivalent.

There is also a requirement to consider market interventions and their unintended consequences. As costs will provide one of the primary objectives for future AI-informed network devices, market signals could be a direct, or indirect, driver of cascading risks. These risks and their causal outcomes will need to be mapped and understood, and necessarily interventions and protocols will need to be devised.

There needs to be a degree of accountability and transparency in the system so that if things do go wrong and there is a failure in one part of the system that goes on to affect another part of the system, that regulation can be enforced, and people held to account to remedy and mitigate future risks.

8.5 Recommendations

8.5.1 Develop a Cross-sector AI Special Interest Group

Formally engage industry leaders from across the energy sector to create a Special Interest Group. This would allow for debate on the industry's AI roadmap, sharing knowledge and best practice, developing collaborative initiatives, engaging policymakers and regulators, educating wider industry stakeholders.

This could be led by Energy Systems Catapult and NESO and help set the framework for a wider body of work addressing other focus areas mentioned above[96].

8.5.2 Develop a Sandbox Testing Environment

The AI risk profile will continually change as the energy system becomes more decentralised, complex, inter-connected and automated, yet the industry currently lacks the tools to develop detailed modelling insights and understanding of how risks might play out in practice beyond the hypothetical. Building a sandbox testing environment would enable industry to model potential risk scenarios and understand cascade effects from interconnected systems to proactively map risks based on their likelihood and impact and develop mitigations.

It could be used to explore use cases, test algorithm interactions for potential unintended consequences, quantify risks and develop risk mitigation strategies and support best practice guidelines. More specifically, it could look at how AI can support moving from deterministic to probabilistic planning for network operations, or how AI can be embedded in system operation along with how the network would be operated in an emergency if AI fails. It could also act as a training ground for control system engineers to learn how to detect AI malfunctions and how to respond. The Connected Places Catapult Climate Resilience Demonstrator (CReDo)[97] offers a potential model that could be applied to support this.

8.5.3 Develop Best Practice AI Risk Guidance

Regulation and governance for AI in the energy system is a rapidly emerging space as it responds to national level publications, and it is beyond the remit of this project to make specific recommendations for regulation. Building on the previous two recommendations there is an opportunity to set out clear best practice AI risk guidance for the energy sector drawing on risks frameworks, setting the direction of travel, minimum standards expected, particularly for newer entrants into the market who may lack domain knowledge, provide best practice advice and regulation. Some specific ideas on best practice guidance and principles have been collated from our investigations and interviews in Appendix 9.3 but

[96] At the time of publication, the ADViCE programme has set up a working group engaging with AI experts across the sector which will consider some of the challenges posed in this report.
[97] https://digitaltwinhub.co.uk/climate-resilience-demonstrator-credo/

further, research, testing and scrutiny will be required to develop an appropriate, useful and robust guidance for across the sector.

9. APPENDICES

9.1 Literature Review

[1] ADViCE: AI for Decarbonisation – Ecosystem Report, Digital Catapult, Energy Systems Catapult, The Alan Turing Institute, 2023.

[2] ADViCE: AI for Decarbonisation – Challenges Report, Digital Catapult, Energy Systems Catapult, The Alan Turing Institute, 2023.

[3] Resilient Electric Vehicle Charging, Energy Systems Catapult, 2022.

[4] Algorithm Governance: A Briefing, Energy Systems Catapult, 2022.

[5]: Data Ethics and Bias, Energy Systems Catapult, 2022.

[6] AI in Energy White Paper, Energy Systems Catapult, 2021.

[7] Future Energy Scenarios 2023 Main Report, National Grid ESO, 2023.

[8] The ESO: Digitalisation Strategy and Action Plan, National Grid ESO, 2023.

[9] The future of the ESO and Artificial Intelligence, National Grid ESO, 2023.

[10] Quantifying Demand Flexibility: towards a Standardised Approach to Baselining, Centre for Net Zero, 2024.

[11] Energy Data Taskforce: A Strategy for a Modern Digitalised Energy System, Energy Systems Catapult, 2019.

[12] Energy Data Taskforce: Two Years On, Energy Systems Catapult, 2021.

[13] National AI Strategy, HM Government, 2021.

[14] The joy of flex - Embracing household demand-side flexibility as a power system resource for Europe, Regulatory Assistance Project, 2022.

[15] The Impact of Electric Vehicle Charging on Grid Stability, Sygensys, 2022.

[16] Delivering a smart and secure electricity system, Department for Business, Energy and Industrial Strategy, 2022.

[17] Power Grid IoT System Protection and Resilience using Intelligent Edge (Power-SPRINT), PETRAS, 2022.

[18] EV Technical Standards for Grid Operation, enX, 2023.

[19] Resilient Electrical Vehicle Charging: "REV", Sygensys, 2022.

[20] How to Talk about Cybersecurity of Emerging Technologies, Dr Ola Michalec, PETRAS, 2022.

[21] EU AI Act: first regulation on artificial intelligence, 2023.

[22] Active Network Management: Opportunities and risks for Smart Local Energy Systems, Energy Systems Catapult.

[23] Introduction to AI Assurance, HM Government, 2024.

[24] A pro-innovation approach to AI regulation, HM Government, 2023.

[25] Data Best Practice Guidance, Ofgem, 2021.

[26] SR letter 11-7: Guidance on Model Risk Management, Board of Governors of the Federal Reserve System, 2011.

[27] Applying model risk management guidance to artificial intelligence/machine learning-based risk models, Google, 2023.

[28] Transforming the energy industry with AI, Siemens, 2021.

[29] AI Adoption in the UK's Public Sector, PUBLIC, 2023.

[30] AI Readiness for Government, Deloitte.

[31] Government AI Readiness 2022, Oxford Insights, 2022.

9.2 Stakeholder Mapping

A stakeholder map of key organisations and individuals involved in AI related projects and startups across the energy sector was developed, and from this a prioritised list of people was developed to contact for interview. A total of 25 people (see Table 1 below) were interviewed throughout the project using a semi-structured interview process, to gather as wide a range of views and perspectives on the risks of AI deployment. In addition, ESC held a webinar on AI for Decarbonisation: Energy System Flexibility[98] that yielded insights into participants concerns around the risks, was also included. Thematic analysis was then conducted, to cluster emerging perspectives and issues around AI risks, risk management frameworks, and regulation and governance. The thematic analysis and clustering of key risk areas and recommendations forms the basis of the main sections of this report, supplemented through the desk research and opinions of the Energy Systems Catapult team.

	ROLE	**ORGANISATION**
1	CEO	Sygensys
2	Head of Data Science	Flextricity
3	Data Science & Development Manager	UK Power Networks Distribution System Operator
4	Data Science Manager	National Grid Electricity Distribution
5	Professor	The Alan Turing Institute & University of Cambridge
6	Energy and Utilities Consultant	IBM
7	Head of Energy System Digitalisation	Ofgem
8	Principle AI Technologist	Ofgem
9	Contractor	Ofgem
10	Head of AI Policy - AI Taskforce, AI Regulation	UK Government
11	Principle Data Scientist	Arenko
12	Research Fellow	University of Bristol
13	Reader & Associate Professor	The Alan Turing Institute & University of Manchester
14	Head of Infrastructure	TechUK
15	Global Digital Energy Leader	Arup
16	Data Science Strategy Lead	Arup
17	Senior Lecturer in Future Power Systems	University of Manchester
18	Communication and Systems Integration Theme Lead	Power Networks Demonstration Centre
19	Lead Data Scientist (Machine Learning) / Data Science Manager	National Grid ESO

[98] https://www.youtube.com/watch?v=vMc4-JD2-lk

20	Stakeholder Engagement Lead – Strategy, Digital, Data & Technology	National Grid ESO
21	Holder of the ScottishPower Chair in Future Power Systems	University of Strathclyde
22	Innovative Solutions Architect - Flexibility	Energy Systems Catapult
23	Power Systems Engineer	Energy Systems Catapult
24	Practice Manager (Data Science & AI)	Energy Systems Catapult
25	Technical Lead	Energy Systems Catapult

Table 1: List of individuals interviewed for the AI Preparedness project.

9.3 Collected Principles and Best Practice for Managing and Mitigating Risks

In section 5 many challenges of AI adoption are covered, including those on Skills, Data, and Organisational Culture. In addition to these we also collected many other ideas and principles from our interviews, and research, (including the government guidance from their AI Assurance guide[99]) around managing and mitigating risks of AI systems which we present below.

The below principles expand on some of the ways to identify and manage risk, and some other sources of risk such as machine learning operations. The list is by no means exhaustive but aims to highlight the diversity of issues and some of the properties which must be considered to reduce risks in AI systems.

9.3.1 Machine Learning Operations (MLOps)

Although too specific and technical for the core report proper implementation of machine learning workflows will be key to their safe deployment and helping with mitigating and managing risks. Some of the principles we identified in our investigations include:

- **In-house development**: Where possible, AI applications with high risk should be developed in-house, with the necessarily input from interdisciplinary teams, to ensure full interpretability of the models, and full accountability of the algorithms which they deploy. It also guarantees greater control. Enterprising tools may have better performance, but there will likely also be less transparency and ability to explain the models. The utilisation of explainable AI methods could help improve the understanding of black box models.
- **Reproducibility:** Experimental pipelines should be deployed to support reproducibility[100]. This can ensure better testing and development. Ideally AI models/systems should be a compared to standard and state-of-the-art benchmarks to ensure they have sufficient accuracy.
- **Scalability Issues:** Considerations for scalability, especially for complex models, should be planned in advanced. Not only does this create challenges for interpretability and management of the AI systems, but it may also increase computational costs and excessive energy use. To reduce risks all these systems may require extensive auditing which may be prohibitive for large numbers of algorithms.

---

99 https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf

100 https://es.catapult.org.uk/insight/data-science-for-net-zero-the-value-of-reproducibility/

9.3.2 Identifying risks

The main report identified some general frameworks for risk assessment and management in Section 6, but no specific guidance on identifying risks. There are several places that AI systems pose risks to the energy system. Below are some of the places that risks can more likely appear and therefore could help identify risks.

- **Biggest Impacts:** One approach to identifying risks is to work backwards from the biggest potential impacts in the system.
- **Use case risks:** Each individual use case applied within a system should be assessed for the potential risks they pose. Their interactions and emergent behaviour should also be assessed within a whole systems approach.
- **Gaming Opportunities:** Some risks may be the results of how actors operate within the system. Any system, especially those operated using markets, as the energy system does, provides opportunities for gaming. These should be identified, and their implications assessed.
- **Edge Case Assessment:** Often the highest risks come from edge cases which are not assessed within the normal model testing environments. However, they could have the largest implications. Rare or unusual edge case events should therefore be identified and tested, e.g. extreme weather or market conditions.
- **Multiple interdependency applications and events**: Applications with the most interactions, potentially across multiple systems (telecommunications, gas networks, transport etc.) could have the highest impacts, but are also more complicated to model. They may be a useful test case for digital twin models. What AI systems interact, and what are their potential consequences?
- **High-speed automation applications:** As energy networks become more complicated and require rapid control decisions to be made across multiple devices, it becomes more difficult to keep a human-in-the-loop. This makes such applications riskier, and more difficult to manage. Focus on those events rather than the ones where AI is simply used for decision support.
- **Feedback:** Which models are feeding back into each other with their outputs. This could lead to domino effects in the models, or could lead to model collapse, where the performance of an algorithm drops rapidly as it is training on its own outputs (or the effects of its outputs) rather than real world data.
- **Red teaming:** The best way to identify risks is through utilising teams to test a system and how it would respond to attacks or other risks.
- **Interpolating vs Extrapolating:** Identify which algorithms are interpolating within the domain of the historical training data, and which ones are extrapolating. The latter is where there is less available learning data and may produce spurious results.

9.3.3 AI System Evaluation and Verification

- **Explainability:** Since many AI models are black boxes it is not possible to understand the relationships between the inputs and outputs explicitly. This creates risks when applying methods to critical infrastructure. Explainability techniques have been devised to better understand what the models are doing which can help assess a model and understand where potential errors can come from. Some of the most popular are models such as SHAP and LIME which look at individual features or local representations respectively to help improve model interpretation. However, there are many other techniques including permutation-based feature importance, partial dependence plots, model sensitivity, etc.
- **Third party verification:** It can be difficult for every team developing AI to have all the skills to reduce risks arising from their models. Furthermore, they may have some biases which inhibit their ability to independently assess the models. Employment of third-party teams can help provide independent validation and confidence.
- **Formal Verification:** As described in the Introduction to AI Assurance[101], another possibility is to deploy formal verification in which formal maths methods and proofs are used to verify if a system satisfies specific requirements.
- **Open Testing**: The development of robust, trustworthy and accurate models depends on sufficient open testing against common benchmarks using shared datasets. This ensures that the models are more easily comparable, and a better comparison can be made of important features and pitfalls. They also ensure minimum accuracy requirements are fulfilled.

9.3.4 Risk Management and Mitigations

- **Algorithmic Transparency and Governance:** To properly assess and manage the risks posed by algorithms requires documenting particular metadata[102] about the algorithm so that the impact and interactions of the AI systems are fully understood. For particularly high-risk algorithms there may be a need to register the algorithm as has been considered in the Algorithmic Transparency Recording Standard[103].
- **Mapping Interactions**: Algorithm documentation should also include interactions of the algorithms and what other systems they connect to. This makes it easier to trace the effects of the algorithm and source of risk.
- **Risk Registers:** Risky behaviours cannot be learned from if they are not properly reported and refreshed according to new instances of risk, including close call events.

---

101 https://assets.publishing.service.gov.uk/media/65ccf508c96cf3000c6a37a1/Introduction_to_AI_Assurance.pdf

102 https://es.catapult.org.uk/report/algorithm-governance/

103 https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub

- **Retrospective Audit:** There should be a thorough assessment of an algorithm after a failure has been discovered in a system involving AI. Utilising the metadata documentation mentioned above can help to accelerate the process and identify the likely causes of faults. Explainable AI systems can also be used to help understand the causes of unexpected AI model behaviour and their outputs.
- **Feedback mechanisms:** The projects in the Portfolio of AI assurance techniques[104] demonstrate the usefulness of feedback so that improvements can be to algorithms and future risks and biases can be mitigated against.
- **Human-in-the-loop:** Many of the stakeholders we talked to emphasizes the need for human-in-the-loop for most AI systems, especially those where the impact can be highest. It is highly unlikely these systems will be completely automated so the key is to understand which components can be and which must be left in the control of human operators.
- **Retraining:** Energy systems are changing all the time. In particular it is known that the demand and generation in different areas of the network will have a different distribution as new technologies are connected with varying levels of load/generation and efficiencies. This means that the algorithms will also be out of date, increasing the risk to the systems they operate. There is a need to keep track of the algorithms and ensure they are update frequently and retested regularly.
- **Unknown unknowns:** It will be impossible to account for all AI model risks and their causes, but the effects should be limited as much as possible by considering contingencies in the areas with the greatest impacts. AI systems should be continually tracked, reported on and updated based on new observations and/or aberrations. This should also be shared with other relevant parties (see above).

[104] https://www.gov.uk/guidance/cdei-portfolio-of-ai-assurance-techniques

10. Licence/Disclaimer

Energy Systems Catapult (ESC) Limited Licence for AI Risks for Energy Networks: Challenges, Management and Regulation.

ESC is making this report available under the following conditions. This is intended to make the Information contained in this report available on a similar basis as under the Open Government Licence, but it is not Crown Copyright: it is owned by ESC. Under such licence, ESC is able to make the Information available under the terms of this licence. You are encouraged to Use and re-Use the Information that is available under this ESC licence freely and flexibly, with only a few conditions.

Using information under this ESC licence
Use by You of the Information indicates your acceptance of the terms and conditions below. ESC grants You a licence to Use the Information subject to the conditions below.

You are free to:

- copy, publish, distribute and transmit the Information
- adapt the Information
- exploit the Information commercially and non-commercially, for example, by combining it with other information, or by including it in your own product or application.

You must, where You do any of the above:

- acknowledge the source of the Information by including the following acknowledgement:

"Information taken from AI Risks for Energy Networks: Challenges, Management and Regulation, by Energy Systems Catapult"

- provide a copy of or a link to this licence
- state that the Information contains copyright information licensed under this ESC Licence.
- acquire and maintain all necessary licences from any third party needed to Use the Information.

These are important conditions of this licence and if You fail to comply with them the rights granted to You under this licence, or any similar licence granted by ESC, will end automatically.

Exemptions
This licence only covers the Information and does not cover:

- personal data in the Information
- trademarks of ESC; and

- any other intellectual property rights, including patents, trademarks, and design rights.

Non-endorsement
This licence does not grant You any right to Use the Information in a way that suggests any official status or that ESC endorses You or your Use of the Information.

Non-warranty and liability
The Information is made available for Use without charge. In downloading the information, you accept the basis on which ESC makes it available. The information is licensed 'as is' and ESC excludes all representations, warranties, obligations and liabilities in relation to the Information to the maximum extent permitted by law.

ESC is not liable for any errors or omissions in the Information and shall not be liable for any loss, injury or damage of any kind caused by its Use. This exclusion of liability includes, but is not limited to, any direct, indirect, special, incidental, consequential, punitive, or exemplary damages in each case such as loss of revenue, data, anticipated profits, and lost business. ESC does not guarantee the continued supply of the Information.

Governing law
This licence and any dispute or claim arising out of or in connection with it (including any noncontractual claims or disputes) shall be governed by and construed in accordance with the laws of England and Wales and the parties irrevocably submit to the non-exclusive jurisdiction of the English courts.

Definitions
In this licence, the terms below have the following meanings: 'Information' means information protected by copyright or by database right (for example, literary and artistic works, content, data and source code) offered for Use under the terms of this licence. 'ESC' means Energy Systems Catapult Limited, a company incorporated and registered in England and Wales with company number 8705784 whose registered office is at Cannon House, 7th Floor, The Priory Queensway, Birmingham, B4 6BS. 'Use' means doing any act which is restricted by copyright or database right, whether in the original medium or in any other medium, and includes without limitation distributing, copying, adapting, modifying as may be technically necessary to use it in a different mode or format. 'You' means the natural or legal person, or body of persons corporate or incorporate, acquiring rights under this licence.

Energy Systems Catapult is an independent research and technology organisation. Our mission is to accelerate Net Zero energy innovation.

Launched in 2015 by Innovate UK, the Catapult has built a team of more than 250 people, with a range of technical, engineering, consumer, commercial, incubation, digital, and policy expertise. They draw on sector-leading test facilities, modelling tools, and data collected from our back catalogue of more than 500 research projects.

We use that 'whole energy' system capability to support innovative companies -- small and large – to test, trial and scale their new products and services. Our impact comes when those innovators attract new customers, new investment, and new grants so they can thrive in the future energy system.

Based in Birmingham, Energy Systems Catapult is part of a network of nine world-leading technology and innovation centres, established by Innovate UK. The Catapult Network fosters collaboration between industry, government, research organisations, academia, and many others to transform great ideas into valuable products and services.

Energy Systems Catapult
7th Floor, Cannon House
18 Priory Queensway
Birmingham
B4 6BS

es.catapult.org.uk