DATA ETHICS AND BIAS

PRACTICAL STEPS TO AVOID DISCRIMINATION IN FUTURE SMART LOCAL ENERGY SYSTEMS

DIGITAL & DATA

WEDNESDAY 21 SEPTEMBER 2022

CONTENTS

1. EXECUTIVE SUMMARY

In order to promote the sustainability and economic viability of smart local energy systems, energy products and services will need to change and adapt to new technologies, new tools and new techniques in order to be successful. In this report a set of principles are proposed to prevent structural bias from entering the data that underpins these new products and services. These principles are partly drawn from existing work with use cases to support them, but also contextualised by the Energy Systems Catapult authors to focus on energy.

The principles described in this report include:

- **user group principles** or how to identify groups that may be adversely affected by data bias
- **user effect principles** which demonstrate how users can be affected by data bias
- **data mitigation principles** which give a set of ways in which data bias can be mitigated.

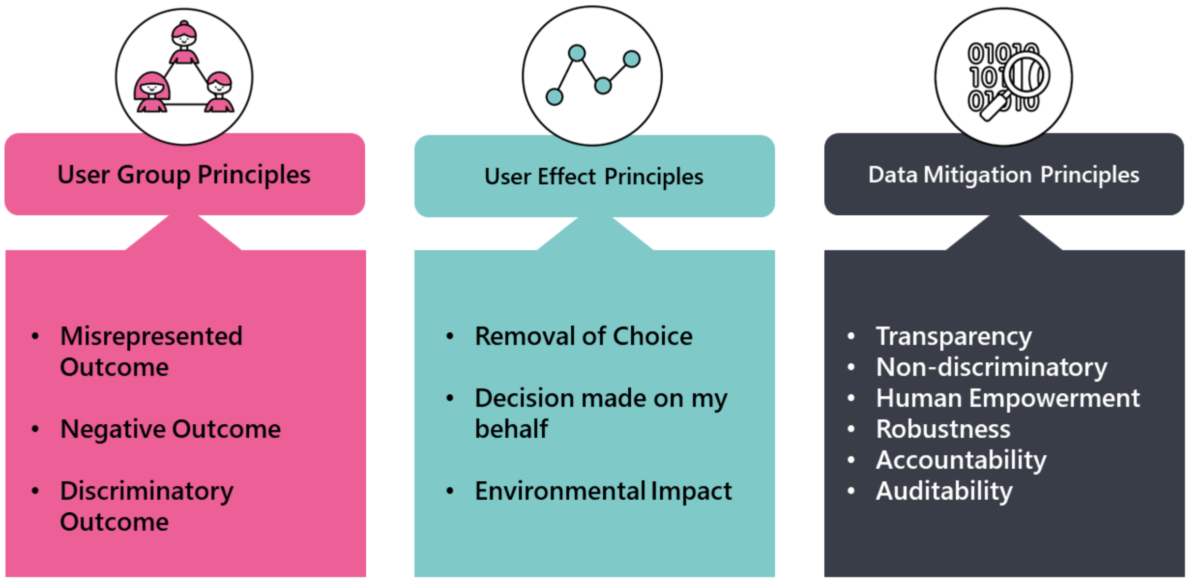


Figure 1. Summary of the principles which will be outlined in this report.

This report demonstrates through a series of relevant use cases that data bias can create negative or discriminatory impacts on specific groups of users, and how the application of the data mitigation principles might have been used to mollify those impacts if pre-emptively considered.

The principles are also contextualised in terms of smart local energy systems since they will be the focus of many of the more sensitive elements of granular and potential personal data. Guidance is given as to how to apply the **data mitigation principles** to promote good outcomes for consumers.

It is expected that further use cases and issues will arise with the development of future smart devices and higher resolution monitoring. Further to this, the government has recently launched initiatives to investigate algorithmic transparency, which overlaps significantly with the messages within this report. In the future we expect to write on algorithmic transparency and it is not considered here. Energy Systems Catapult would welcome the opportunity to work with organisations looking to develop a smart local energy system to help apply these principles and support the development of energy products and services that take account of the individual needs of user groups and provide a decarbonised, decentralised and democratised energy system.

2. INTRODUCTION

In an increasingly digitalised world, data and Artificial Intelligence ethics is rapidly developing as an area of interest. Up until recently the energy sector has not deeply considered this topic due to personal data being siloed within a few organisations, high-resolution data being relatively sparse, and very few technologies/organisations actually utilising advanced algorithms and machine learning. This is expected to change rapidly in the next few years. With the smart meter roll-out there is expected to be granular (half hourly) monitoring of all households and consumers. The increased uptake of controllable and smart technologies mean that responsible and ethical utilisation of personal data and how it interacts with people will quickly become a ubiquitous problem if safeguards and guidelines are not put in place.

This report is an introductory insight paper into a subset of these issues considering data ethics for smart local energy systems (SLES). It is designed to help support those implementing a SLES's tackle the practical ethical challenges surrounding data when:

- **Assessing risk in data collection**
- **Building algorithmic control systems**
- **Applying machine learning techniques.**

Energy needs to work for everybody. This paper aims to provide some initial guidance on how to find and identify places where bias can creep into data which is collected, pre-processed, and then utilised in other algorithms such as those used in control mechanisms and in-home smart systems. The algorithms themselves can also produce biases but we only focus on impacts caused by the input data, and their implications for the subsequent data after modification through the algorithms. Algorithms themselves can themselves introduce biases and discriminatory outcomes and this requires a full treatment beyond the scope of this report and will be considered in a future investigation.

Future papers will build on this work in more detailed ways to describe the specific ethical challenges associated to the algorithms and the new products being developed for smart local energy systems. Other papers, such as [1] have identified the need for ethical practices in AI for energy systems, making recommendations for sector specific guidance enabled by regulation.

2.1. DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

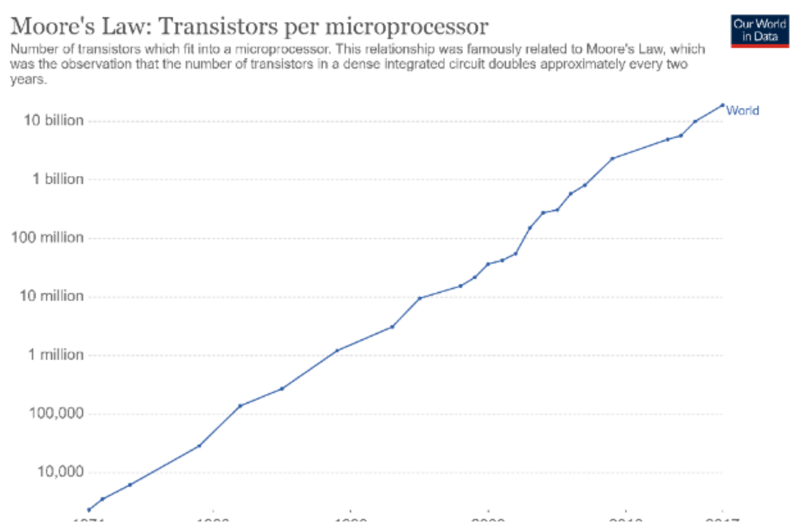


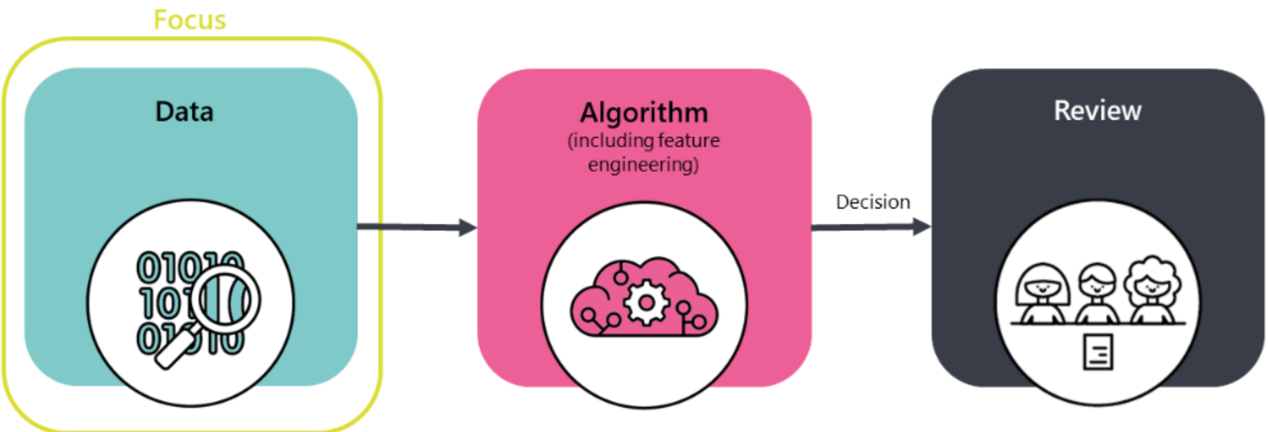
Figure 2. Number of transistors per microprocessor as a function of year. From Our World in Data licensed under CC BY 4.0.

Data storage technologies have increasingly been made cheaper and more accessible. Tools and technologies are now available that make access to data collection, batch processing, and analysis easier and simpler than ever. This has assisted the rise in the utilisation of a range of data science, artificial intelligence (AI) and machine learning (ML) techniques. Many of these are based on methods that have existed and been utilised in a limited form for a long time, but until the last few years, there has been constraints in data, computing resources and storage.

AI and more advanced algorithmic decision-making have the potential to profoundly change our lives for the better. However, there are undoubtably risks involved in such advancement. This paper considers only the risks and challenges associated with the narrow AI (or weak AI[1]) available today, which can outperform humans at specific tasks (e.g. National Grid's use of machine learning to predict solar generation in Britain [2]).

The data science process associated with narrow AI applications consists of three general parts (see figure below):

- The data collected (including any manipulation of data prior to the use of machine learning or other similar techniques).
- The algorithm(s) that makes or informs the decision
- The review of the decisions and outputs which the algorithm helps inform (and the subsequent refinement of the algorithm and or data utilised).



Each part of the process can affect the final decisions made. Therefore, algorithms can result in decisions which are unfair or unintended (or at least perceived as such). While all parts of the process are equally important and can influence the decision, the specific focus of this paper is on the input data and the potential biases that can result. Subsequent work should consider the algorithm and review steps.

The smart local energy systems (SLES) being developed must be sustainable, economically viable, and accepted by the local population. Through considering new and innovative ways of thinking about risk in data collection and algorithm use, we can ensure that individuals and communities can have better trust that their data is being used ethically with companies ensuring that are applying standards and techniques to mitigate risk and address bias.

Data should be collected for a specific purpose or application. For example, to create personalised energy usage tariffs which better support the occupants will require collecting and analysing their smart meter data. However, it should be clear to the household that you are collecting it for this

[1]Although it has many definitions, narrow (or weak) AI typically focuses on solving a particular task, e.g. operating a self-driving car. Such an algorithm couldn't be applied to new tasks, and would not be considered to have "conciousness". This is in contrast to Strong AI, considered by many to be the ultimate goal of AI, a sentient machine that can intelligently solve a wide variety of problems.

purpose, and you should not use it for other applications or experiments outside of the original scope without explicitly informing them. Further to this, if the data is collected for a particular application and the data collected contains only a narrow range of data points or omits diversity in the sample, then it can introduce bias to the results. A major objective of this report is to understand where bias is appropriate to accept and how bias may lead to discriminatory outcomes. Bias can mean several different things in society, algorithms, data etc. and hence it is worth being clear what we mean when we discuss bias in this report. In our report we refer to bias as a disproportionate representation of a particular individual or group either in the original data collected or in the final outputs after pre-processing or algorithms have been applied. Hence, the biases may already be embedded in the data being utilised or can be created/generated by the manipulations of the analysis. The data in the bias in the data may be because it represents the biases in the overall population or may be a result of a non-representative sampling method.

Energy in the UK is a product that is very difficult to opt out of without significant effort and capital, therefore any solution that changes the way in which the general public engages with energy must be suitable for the whole population. For example, although energy use peaks are well understood as a national phenomenon, energy usage is a superposition of many different demand profiles at an individual level that are highly variable across the day and as diverse as the people themselves. Therefore, peaks (and emissions) at national level can be mitigated using individual usage data within say, smart control algorithms, but they must ensure that they do not treat any groups/individuals unfairly or produce harm on those who are most vulnerable (for example, those who depend on dialysis machines and therefore can be harmed by losing their electricity supply).

Furthermore, whilst it can be argued that reducing certain forms of discrimination can create social good, from an economic perspective, bias can create inefficient outcomes. The growth of clean energy and the net-zero target requires the best use of resources to succeed. For the transition to net-zero the best data, skills and innovation will be needed to accommodate the diversity of energy needs. Smart Local Energy Systems provide the opportunity to create energy solutions that cater to all needs.

3. PRINCIPLES SUMMARY

Three sets of principals are proposed for addressing data bias in smart local energy systems:

1. **User group principles**: a simplified method of identifying for whom the solution works and who is likely to have been excluded.
2. **User effect principles:** a guide to negative impacts any bias may create for the users.
3. **Data mitigation principles:** a guide to start de-risking data collection methodology to prevent bias from flowing through the process.

In the next section, examples have been highlighted where data collected or used without a proper understanding of the biases can create negative or discriminatory outcomes. Some examples of discriminatory outcomes covered in this report are isolated cases, others perpetuate structural inequality. **The aim of these principles is to de-risk data collection, not to prevent data from being collected, and to transition towards a more ethical process that embeds an adherence to regulations and builds an ethics of care into the process.**

The principles have been chosen to be aligned with other themes set out for the ethical use of AI elsewhere [3] [4] [5] [6]. They are closely aligned to the European Union's ethics guidelines for trustworthy AI [6]. These principles are explored further in the use cases section within the Appendix for a variety of non-energy focused applications.

3.1. USER GROUP PRINCIPLES

As nearly every member of the UK population uses energy in some way, they are defined as users in this context. Users and user groups participating in smart local energy system projects should be identified with a robust methodology such as the Smarter Consumer Protection Manual [7]. In the context of data ethics, it is important to account for the groups that either are not present in the data collected or have been excluded. This includes protective characteristics under the Equality Act 2010[2] such as sex, gender reassignment, race and sexual orientation, and others such as economic class, education, and geographical location. Ethical approaches to data usage requires framing these characteristics as interconnected to address how complex identities may be more vulnerable to discrimination than others.

As an example within an energy context, those in fuel poverty are of a particular concern due to their vulnerability to increasing energy costs, and the direct impact of sudden weather changes on their health and safety.

The accidental exclusion of data collected on particular groups of users can produce non-desirable and unethical outcomes, which can be categorised into the following based on their use in data and/or algorithmic systems:

- **Misrepresented outcome**: The conclusion is misrepresented, misinformed, perhaps contains conscious (or unconscious) bias or does not reflect reality
- **Negative outcome**: If by utilising data or an algorithmic control scheme, the incorrect outcome makes the user's life more difficult than it was previously
- **Discriminatory outcome**: If by utilising data or an algorithmic control scheme, the solution creates a biased outcome for a certain group of people or excludes a certain group of people from the benefits

To develop algorithms that address and/or identify biases or discrimination in the outputs, requires the collection of sensitive information. The collection of this data will signpost any redundant

[2] https://www.legislation.gov.uk/ukpga/2010/15/contents

encodings (i.e. those variables which are strongly correlated to the sensitive variables) and prevent less easily identifiable discriminatory outcomes. However, without robust transparent data collection and handling protocols, these process could amplify and replicate discrimination. In Appendix 7, several use cases are presented which can illustrate where negative or discriminatory outcomes has resulted from the poor handling and collection of data.

3.2. USER EFFECT PRINCIPLES

Poorly designed data selection and/or inappropriate implementations of decision supporting algorithms can have unethical and undesirable effects on users. Furthermore, the use of data to make decisions does not necessarily only directly affect the consumers of a product or service, it can also affect third parties. Discriminatory outcomes of algorithmic decision making can be categorised in the following way

- **Removal of choice**: A user is restricted to accessing a narrower selection of products or services than they had access to previously. This can create a negative outcome, if the selections remove the opportunity for a better deal, results in an increase in price, or are less suitable for the user's needs. Similarly, a discriminatory outcome occurs if it limits the choice of a particular user or user group, as discussed in the user principles.
- **Decision made on my behalf**: A user's product or service is managed by a system or process outside of their influence or control. It is natural for products and services to change over-time but this creates a negative outcome if the decisions produce a product which is worse than it was previously, or the user has been disadvantaged by the decision. This creates a discriminatory outcome if the algorithm makes negative outcomes only on behalf of a particular group or set of groups, for example, the protected groups identified in the user principles.
- **Environmental impact**: A user who is not directly connected to the product or service but nevertheless may be indirectly affected by decisions made by other users or by the supplier of a product. This creates a negative outcome if for example, it affects the price others pay for their gas. This creates a discriminatory outcome if it raises the prices for some groups and not others.

In all three cases, third parties may be affected in both positive and negative ways. For example, citizens in the local area may benefit from a cleaner environment and lower energy prices as a result of a decision to charge battery storage in their energy system, but other users near the battery storage may suffer an under voltage of their electrical supply. In certain circumstances, this may be an acceptable outcome overall; the point is to pre-empt and understand the effects when developing the service rather than having to address these issues afterwards.

A set of use cases in the Appendix Section 7 seek to explore the implications of decisions, from the point of view of the users for a set of non-energy specific applications.

3.3. DATA MITIGATION PRINCIPLES

There are six key principles that if followed will help guard against undue data bias. This section begins to present guidance on how to manage data in the context of ethics in preparation for the use case discussion in section 4. The key principles for data collection are:

1. **Transparency**: the data being used to inform final decisions should be explained in an accessible and transparent manner to stakeholders, addressing concerns where possible. They should also be made aware of the data's strengths, limitations, and biases, and how it was collected. Good practice around transparency is critically need for traceability and provenance (i.e. good documentation) of the data.
2. **Non-discriminatory**: care must be taken to avoid drawing conclusions when bias exists within data (using principle 1), as it could have multiple negative implications. This could include erasure or marginalisation of vulnerable groups and exhibiting prejudice and discrimination bias. In addition to

the use of diverse data sets, there should be an interrogation around the limitations of existing data sets to ensure outputs are accessible to all, regardless of user group.

3. Gathered and used for **human empowerment** and **societal well-being**: Data should be collected for the benefit of all human beings, including future generations. The environment, other living beings and the societal impact should be considered.

Systems should be in place to ensure:

4. **Robustness:** data should be collected, used and stored in a resilient and secure manner. Care must be taken to ensure that privacy and data protection considerations are fully respected. Appropriate governance mechanisms need to be put in place to: (a) ensure accurate and reliable data collection, (b) maintain the quality and integrity of data and (c) ensure legitimate access.
5. **Accountability**: mechanisms should be put in place to ensure clear responsibility and accountability for the collection, storage use of data, and data breaches. For example, a clear line management to report potential data breaches, including anonymous reporting.
6. **Auditability**: how data is collected, used and stored should enable independent assessement. This should include reviewing how the potential for bias has been addressed and the data protection measures in place. Where necessary, adequate and accessible redress should be provided and updates to the process should be applied.

4. CONTEXTUALISING SMART LOCAL ENERGY SYSTEMS

Examples of ethical challenges in smart local energy systems applications are relatively sparse due to the relative lack of maturity of such systems. For this reason, we have included an appendix (Section 7) that describes use cases from a wide range of sectors and applications such as public spaces, healthcare and technology. What we now discuss are the considerations for the design of future smart local energy systems.

A key part of SLES design is a smooth and efficient path to decarbonised solutions that suits the needs of all current consumers. Understanding the limitations and biases associated with the data collected will go a long way to ensuring future products and services respond properly to market signals and demand rather than catering to small groups, which will promote the long-term viability of any future market solution. The rights of the individuals must be properly balanced with the greater good of society and to ensure success will likely need to be financially viable. It is suggested that a suitable identification of users groups and user needs is conducted before applying the above principles to the data used in a product or service.

4.1. SMART PLUGS

Many businesses are considering so called *smart plugs* as part of their building management systems to reduce demand and carbon emissions. Devices such as laptops, printers etc. can be plugged into these devices, which means the usage of the appliance can be monitored and controlled. Inefficient devices can be identified, as can periods of inactivity. Thus new, more energy efficient devices can be installed and/or devices can be turned off when not in use.

Although such devices could save money, help businesses decarbonise, and play a part in the UK achieving net zero, there are some serious potential implications for the users based on the data that they collect. The clearest potential misuse is the potential for employers to track individual work patterns and habits. As the pandemic has demonstrated, flexible working patterns have been important for the mental health of individuals and strict requirements may unduly effect those who are particularly vulnerable in society.

The principles outlined in Section 3.3 should be applied to ensure that the data is utilised properly. For example, transparency (Principle 1) is key so that employees understand how their data is going to be used, which should focus on reducing carbon emissions and better utilisation of technologies (Principle 3). Additionally, processes should be in place to ensure that the data is used appropriately, and biases are addressed (Principle 6).

4.2. ELECTRIC VEHICLES

As transport is one of the biggest sources of emissions in the UK, it is a core focus for smart local energy systems. The need for electrification of transport is driving an increased uptake of electric vehicles. However, when too many EVs are connected and charging on the same network it can cause grid stability issues and may lead to blackouts. Data collected from EVs can lead to better understanding of usage, driving and charging habits, which can then provide opportunities such as tailored tariffs and coordinated control to ameliorate their impact on the electricity supply.

Much of the data relating to electric vehicles has a lot in common with current transport surveys with the addition of the mechanism used to charge the vehicle. Some data recorded might be journey e.g. start and end point, journey time, purpose of journey, charge location, charge time. In the future product and service satisfaction metrics may be added based upon charging time or charge point performance.

Firstly, the transparency (data principle 1) of the user groups identified in the collection data can be implemented. This will ensure that a basic understanding of user groups is formed for the conclusions of the studies. As an example consider data collected based on gender. This could help delivery of a particular service which are based on gender differences, but unless well-thought out in the original data collection request, could actually have a negative impact if it reinforces gender biases (Principle 2). A final consideration is to limit access to detailed journey information (data principle 4) to ensure the privacy of the participants in any trial.

In addition to journey data, EVs also have charging behavioural data. This can provide insights into usage, and where and when the user will charge. This data may eventually be used to coordinate the charging of several EVs to support the network. It is imperative that there are fair ways to allocate charge to satisfy both the needs of the drivers and the network. It is easy to envision that there could be biases or discrimination (principle 2) in allocating and coordinating preferences for charging especially if an unconstrained or based on economic advantages.

4.3. SPACE HEATING

Due to a lack of information on consumer habits, and poor visibility of EPC ratings and analogue measurement of consumer energy usage, space heating habits are generally poorly understood in the UK. However, space heating is another leading cause of emissions in the UK and so has been targeted by smart local energy systems demonstrators and local authorities alike for decarbonisation.

Data commonly collected in space heating includes: target temperature, time taken to reach temperature, temperature in each room, use of heating boosts, hot water usage, heat pump reservoir temperature, gas consumption, electricity consumption, temporal heating patterns and reported comfort of the house occupants.

A starting point for space heating is to examine comfort metrics, but in the interest of non-discrimination (data principle 2) there needs to be clarity around what user groups are reporting the comfort in the household and disaggregation of the data. For electric heating systems such as heat pumps, methods of identifying vulnerable consumers could be implemented to ensure their needs are catered for (data principle 3). This data must be collected in a robust manner and informed consent given (data principle 4) for its collection, with clear messaging about the benefits of its collection on the product being offered.

4.4. ENERGY STORAGE

Energy storage is emerging as an important way to effectively utilise energy generated from intermittent renewable sources such as wind and solar and distribute renewable energy onto a local grid. Types of data collected can include discharge rate, discharge times, internal resistance, temperature, fast frequency response, degradation, duty cycles and power output.

As an example, grid scale or distribution level battery storage can have knock-on effects on the local energy grid and environment. For batteries providing flexibility services, as well as storage, it is important that data is collected to be accountable (data principle 5) for the outcomes that take place when providing those services. This could be in order to prove the viability of a battery to a local energy system, demonstrate the effective utilisation of local renewable sources, or prove that local decarbonisation has taken place. In a similar way, data should be collected that enables auditability (data principle 6) of the product being offered. Although grid scale batteries are becoming more common, it will be important in the demonstration of sustainability that distribution connected batteries are having a positive impact for the energy system, and consumers with the capital to purchase one.

4.5. LOCAL ENERGY MARKETS

Markets can be effective ways of communicating price signals to interested parties and breaking down the barriers to entry to participants in an industry. Customers benefit from markets that promote effective competition which keeps prices appropriate and encourages high quality products and services.

There are multiple investigations into possible future market structures for several different applications, including the procurement of flexibility, the real-time balancing of distribution networks, and the effective utilisation of renewable sources. Data collected can include flexible capacity available over time, registered market participants, technology types, opening and closing prices, bids for flexibility, promised capacity vs delivered capacity and supply – demand curves.

It is important in market structures to encourage transparency (data principle 1) so that all market participants have access to the same information and that no party is favoured over another. This will promote greater innovation and improve customer and market choice. Much like the current markets, it is important to ensure accountability (data principle 5) from each supplier or procurer for the deliverability of the power agreed and to ensure that each party is adhering to its commitments, which will also demonstrate the benefits of the market for observers. Finally, the auditability (data principle 6) of the market outcomes must be ensured for examination by the participants and by the regulator to ensure adherence.

4.6. ALGORITHMIC CONTROL SYSTEMS

It has been stated multiple times in this document that as smart local energy systems move forward, data flows through to algorithmic control systems will become more common. This could include e.g. residential demand side response (where residential energy usage is altered based on signals) or altering the consumption patterns of large-scale consumers to provide energy flexibility. In any case the idea of using these control systems is that the system can respond in real time with a set of pre-established rules with less human intervention than would have previously been needed.

The full ramifications of algorithmic control systems will be discussed in future papers, but with regard to the data it is important to ensure the rules are designed with non-discrimination (data principle 2) in mind to avoid discriminatory outcomes. This can include processes that promote source data that adequately accounts for user differences and needs, and addresses bias (implicit and explicit) inherent in the data. Finally, it is important to ensure that the products and services are used for human empowerment (data principle 3) to drive sustainability of the business model in a decarbonised energy system.

5. CONCLUSIONS

This paper has been produced to demonstrate how a set of simple principles can be used to mitigate the risk of negative and discriminatory outcomes in future smart local energy systems. This report seeks to highlight that data driven sustainable business models and economically viable energy products and services must be developed, which keep in mind the potential risks and biases when collecting, pre-processing, analysing, and utilising data.

By examining a few use cases and examples, it has been demonstrated that biases can creep into complex systems if the data foundations are poorly conceived, are not sufficiently robust and lack accountability or transparency. In the best cases, this makes it more difficult for people to achieve their full potential; in the worst cases it can be fatal.

The examples presented were all worthy cases for making processes more efficient and helpful in promoting innovative approaches to data ethics and bias considerations. However, it has been demonstrated via the use cases that data bias can create discriminatory outcomes because of insufficient consideration of users, making decisions on behalf of users, limiting the ability of people to reach their full potential, or in multiple cases endangering lives.

The principles applied to the use cases have provided a starter for organisations to consider how best to de-risk data acquisition adopting and applying a set of data mitigation principles. It has been shown how the principles can apply to emerging technologies in smart local energy systems, particularly with the underlying components that will constitute complex products and services. Although a set of principles has been proposed, it is recommended that the Smarter Consumer Protection manual is used with this paper to fully understand users and how they may be affected. Other available resources which readers may find useful include the Tech Transformed tools developed by doteveryone [8] (now run by the ODI) for sustainable innovation design and the data ethics canvas by the ODI [9].

This paper serves as a foundation of data principles for smart local energy systems. Energy Systems Catapult will continue to engage with interested parties to apply these principles to data collection in emerging smart local energy systems. It is important to create a solid and fair foundation of data so that future energy systems work for everyone.

6. AUTHORS AND CONTRIBUTORS

This paper had contributions from many individuals over its development. We thank our colleagues from across ESC who provided insight and expertise that greatly assisted its development, including. Dr Stephen Haben, for taking ownership of this paper and pushing it forward to its conclusion. Authors and contributors of this paper included Jake Verma, Dr Chris Harrison, Dr Anthony Woolcock, Dr Richard Dobson, Naomi De Silva, Dr Andrew Barton, Gordon Graham, Anna Stegman, Jim Lott & Greg Johnston.

We also thank Dr Allison Halford from Coventry University for her time, expertise and insight in the review of this document.

7. APPENDIX USE CASES

This appendix investigates non-energy applications but highlight many of the potential ethical issues and discriminatory outcomes by not considering the risks when collecting and using data and applying advanced machine learning algorithms to make decisions.

7.1. MEDICAL SCHOOL ENTRY

Automated decision making always encodes the past. This in part is because the designer's decisions are always retrospectively using data and decisions from the past. One example of this is the use of an automated system to shortlist applications for interview by the St George's Hospital Medical School [10] [11].

The system was conceived to both reduce the workload on staff and to improve the consistency of the process by removing the subjective nature of the previous process. In many ways it achieved those aims, with the final version showing 90% to 95% correlation with the gradings of the selection panel [11]. However, that did not mean it was providing fair decisions.

The data used to decide who should be interviewed did not include information on the sex or ethnic origins of the applicants, but the Commission for Racial Equality still found the medical school guilty of racial and sexual discrimination. This was not because the new system was introducing new bias to the selection process. It was just reflecting the bias within the old system contained within the data used to develop it that had been encoded in the 1980s.

The system deduced information about the race and sex from the names and place of birth of applicants. This resulted in as many as 60 applicants in 2000 being refused an interview purely because of their sex or racial origins [11]. While the commission did not serve notices against the medical school, there were obvious costs to not having transparent or auditable processes in place. In addition to some undoubtable reputational cost, St George's hospital had to go through the costly process of checking previous decisions and contacting those involved. This resulted in at least three previously unsuccessful applicants being offered places [11]. The reputational damage may also have discouraged female or non-white applicants from applying due to perception of bias against them.

How might this be addressed in the context of the principles in Section 3.3? This example highlights the importance of non-discrimination (data principle 2), that care should be taken that the data does not introduce bias. This may mean collecting data to check for bias. The commission recommended that a question on racial origin be included in future applications. This could have been used to check the original system was not biased before developing a new system to directly mimic it.

The biased nature of data sets is not uncommon and to some extent differences within the data, which may introduce the bias, are required to drive the decision-making process. However, knowing about bias in particular data sets can help avoid these biases influencing the decision. For example, Xerox noticed when they were using similar methods to assist with recruitment decisions that those who lived further from a job were more likely to churn [10]. This, however, also correlated with coming from poorer neighbourhoods and so they choose not to use this to inform the decisions of recruitment algorithms. This transparency and auditability (data principle 1 and 6) allowed Xerox to avoid a discriminatory outcome [12].

This example also highlights the importance of having humans continuously involved in the process of decision making (data principle 3). More human oversight of the system results may have highlighted the issues before the Commission for Racial Equality were involved. The commission highlighted that more members of staff should have been aware of how the automated selections were made and no one person should have sole responsibility for such a process [11]. Clear accountability for the process of filtering candidates and auditability of the data that was feeding the criteria (data principle 5 and 6) could have also avoided these issues.

7.2. FACIAL RECOGNITION

Facial recognition is increasingly being built into the pipeline of many different tasks and the full extent is yet to be realised. Early controversies in facial recognition included privacy concerns when facial recognition was added to applications such as Facebook's image tagging system [13]. Images with multiple people in them were scanned and added to Facebook's database, even if the person was not a Facebook user.

Privacy concerns aside, more pertinent issues relevant to this report is the frequently reported racial and gender bias of facial recognition algorithms. Such issues with facial recognition started to become apparent very early in the roll out of features designed to assist with face tracking, some as early as 2009 [14] where it was reported that a HP webcam tracked a person in frame with light skin but not dark skin. Many of these problems have been caused by the data used to train AI and ML techniques.

Issues with facial recognition utilising machine learning, such as the product released by IBM, have been reported to have its roots in the benchmark datasets used for training which disproportionately contain images of light skinned men. Using the Fitzpatrick Skin Type classifying algorithm on two of the benchmark facial analysis datasets showed maximum error rates of 0.8% when presented with the image of a lighter-skinned male faces, compared with error rates of up to 34.7% for darker-skinned female faces [15].

All three of the above services mentioned offer access to their facial recognition algorithm via API, with none of the biases highlighted in the documentation for 3rd party access. The ramification of this is that a commercial product offered by large technology companies based on this data would be embedding this bias into applications further down the value chain [15]. In a statement by IBM it was shown that they had improved on their error rates, and it was shown that as a result of the study that many commercially available facial recognition services improved their error rates as a result of the bad publicity [16] [17]. It is worth stating that the propagation of biases through multiple algorithmic systems will become increasingly problematic as more complicated systems of connect machine learning are developed. Hence it is particularly important that biases are addressed in all systems and the algorithms that utilise them downstream.

IBM's response highlighted a follow-on issue, the images used by a training dataset published the following year had been uploaded to an image sharing website and used without the consent of the user that uploaded them [18]. Although the images are covered with a creative commons licence and IBM claimed the training dataset was developed to assist academic institutions, whether the training images were used in their commercial offering was not clear [19].

Such unfairness is an all-too-common occurrence and there have been many similar examples since, which are likely to in part be due to the data bias not being understood or controlled for. For example, Amazon's decision algorithms did not provide same day delivery in areas predominantly made up of black people [20][3] or ongoing issues recognising people of colour when using facial recognition

---

[3] Note that since this article was published Amazon have started to extend their same-day delivery service in some areas, see updates in [19].

algorithms. However, it has been shown that auditing algorithms with biases can help push improvements in a companies products and services [17].

How might this be addressed in the context of the principles in Section 3.3? Issues that result from data bias cannot be completely guarded against, but the history shows ignoring the potential issues when developing new products and services is not without risk. The transparency of the source dataset used when training models is a key consideration in all of these factors in this section (data principle 1) and the auditability of those benchmark datasets available for public consideration when deciding on whether to use the commercial product (data principle 6). Furthermore, IBM's decision to utilise images without consent underlines the importance of a robust data collection methodology which is clear on whether it provides legitimate access (data principle 4).

7.3. TEACHER GRADING

Unfair discriminatory decisions can also occur when working with small data sets which can exhibit particular behaviour purely by chance rather than properly represent the underlying population. Value added scores of different kinds have been used to evaluate the performance of schools for many years. These scores attempt to control for factors such as social deprivation and the special educational needs of children. As such they can be useful tools for assessing school performance, where there are a larger number of children. However, in certain parts of the United States they have been used to assess the performance of individual teachers, with poor scores being a major component in teacher's dismissals [10] [21].

This has resulted in teacher's being unfairly dismissed [10] [21], with respected bodies (including the American Statistical Association) cautioning against using value-added scores for such decisions [22] [23]. Results for individual teachers have been shown to vary widely from one year to the next for all teachers [10] [24]. This means either teacher's performance genuinely changes significantly from year to year, or the measure fails to capture teacher's performance. In both cases, making decisions based on this data would be arbitrary and so fail to meet the intended goal. This is not surprising given scores were sometimes based on only a few data points, as little as 25 in some cases [10] [21].

The example also highlights the importance of a robust data analysis processes (as highlighted in principal 4). When using small amounts of data, errors can result in significant changes in the outputs. Here the robustness of student's results has been called into question, with the suggestion that some teachers may have inflated test results [10] [25]. This means the following year pupils will not have shown so much progress. If true, this highlights that robust data collection processes were not put in place to guard against potential fraud.

Finally, the data (and methods) used to calculate the value-added scores was far from transparent (as would be expected by principal 1). When teachers tried to get information about how their scores were calculated this information was not forthcoming [10] [26]. Thus, they could not challenge its validity and highlight why the measure was so poor. Ironically the results of these issues may have produced reduced human empowerment and societal well-being (data principle 3), with poorer teachers remaining in the classroom.

7.4. CARDIOLOGY

If data is collected in a way which introduces biases, these biases become embedded into the final outcomes and decisions. In many cases pre-existing biases can be confirmed if the data collection is influenced by the initial hypothesis being tested. This is particularly apparent when collecting data as part of behavioural studies [27]. One particular way in which biases can creep into data is insufficient disaggregation of observational data. What this means in practice is a lack of understanding as to who the participants in studies or trials are and why they may behave the way that they do.

A key example of this is in medicine and was published in a 2019 editorial in The Lancet. This article highlighted that the symptoms of heart attacks for women were often vastly different from men, but as the data collected had historically had a male focus the advice in medical textbooks advised that women should be assessed in a similar way as men [28]. The Lancet article highlighted in particular that "*Many guidelines for the management of the 50% of heart disease that occurs in women are extrapolated from studies that predominantly enrolled men, such as the Harvard Physician's Health Study done in 22 000 men that formed the basis for aspirin in the prevention of heart attacks*" [28]. In the best-case scenario this leads to misdiagnosis and pain, in the worst cases preventable deaths of women (a discriminatory outcome). It is this stark example that highlights the importance of disaggregating information to understand where bias may be found.

In other studies, this has been quantified in terms of the impact it has had on women. A ten-year study showed that in England and Wales over 8200 preventable deaths were recorded due to the misdiagnosis of symptoms [29]. The study also pointed out this did not cover all hospital admissions, so the actual number was likely to be much higher. Furthermore, a separate study showed that women were 50% more likely to receive a misdiagnosis than men [30] whereas a correct diagnosis increases the survival rate from heart attacks by 70% [30].

How might this be addressed from the perspective of the principles in Section 3? It is clear that a lack of transparency in the data, that studies were primarily focused on men, have led to women's symptoms being misunderstood despite similar historical death rates from heart disease [31]. When collecting data the non-discriminatory principle (data principle 2) must be implemented, and data collection should accommodate the collection of disaggregated data. Clearly understanding where the data has been sourced and who it covers is key to ensuring conclusions are valid for the users.

7.5. TRANSPORT PLANNING

In transport planning and mobility in general, systems are often built with commuter traffic at peak times in mind. In cities where public transport is generally available it is also designed for point-to-point commuter journeys or to/from transport hubs [32]. Caroline Criado-Perez points out that in many societies men dominate access to cars which is what the transport system is designed for. It fails to account for chained trips, for example schools-to-shops, or home-to-school-to-work journeys, of which many are part of unpaid care work [33] [34]. This issue extends to those who do not own cars or are not able to drive.

Criado-Perez highlights that this extends to other ancillary attitudes around the users of the transport system. Snow clearance in a Swedish town was always prioritised by the peak time traffic routes, but a change to the prioritisation in favour of smaller roads and pedestrian routes, it was found that the overall burden on the healthcare system in that area from slips and falls reduced as the walkways were now safer, with no additional impact on the car accident rate [33]. As this mobility was predominantly for unpaid care work, there was a positive outcome for women.

How might these be addressed from the perspectives of the principles in Section 3? By properly understanding the needs of the users of the transport system it is easier to implement a non-discriminatory outcome by ensuring the needs of more of the users are met. Criado-Perez suggests a simple solution, collect sex-disaggregated data to understand where non-discriminatory principles can be applied [35]. This in turn has the potential to increase the human empowerment of the users currently left behind in transport and urban planning for better societal wellbeing, for example by easing the burden of unpaid care work or allowing more mobile (i.e. non-commute) business models to thrive.

In addition to this, the transparency of the users identified (data principle 1) extends further than just the use of the system, it has knock on effects on health benefits from cleaner air, lower congestion and lower transport costs through efficiency. Mobility is a key mechanism in economic output and is already a key focus of the future non-fossil fuel transport system. The auditability (data principle 6) of the data collected about decarbonised transport will be key in assessing the impact of mobility solutions such as electric vehicles or other micromobility services.

8. REFERENCES

[1] E. Morris, K. Stamp, A. Halford and E. Gaura, “The practice of AI and ethics in energy transition futures,” EnergyRev, 2022.

[2] National Grid ESO, “ESO and The Alan Turing Institute use machine learning to help balance the GB electricity grid,” National Grid, 25 July 2019. [Online]. Available: https://www.nationalgrideso.com/news/eso-and-alan-turing-institute-use-machine-learning-help-balance-gb-electricity-grid. [Accessed 07 August 2020].

[3] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo and L. Floridi, “Artificial Intelligence and the 'Good Society': the US, EU, and UK approach,” *Science and Engineering Ethics,* vol. 24, no. 2, pp. 505-528, 2017.

[4] C. Véliz, “Three things digital ethics can learn from medical ethics,” *Nature Electronics,* vol. 08, pp. 1-3, 2019.

[5] S. Pichai, “AI at Google: our principles,” Google, 7 June 2018. [Online]. Available: https://www.blog.google/technology/ai/ai-principles/. [Accessed 27 March 2020].

[6] Independent High-Level Expert Group on Artificial Intelligence, “Ethics Guidelines for Trustwory AI,” European Commission , Brussels , 2019.

[7] R. Chard, “Smarter Consumer Protection Manual,” Energy Systems Catapult, 21 February 2020. [Online]. Available: https://es.catapult.org.uk/brochures/smart-consumer-protection-manual/. [Accessed 06 May 2020].

[8] “Responsible Innovation,” doteveryone, [Online]. Available: https://www.doteveryone.org.uk/work/responsible-innovation/. [Accessed 06 May 2020].

[9] “Data Ethics Canvas,” Open Data Institute, [Online]. Available: https://theodi.org/article/data-ethics-canvas/. [Accessed 06 May 2020].

[10] C. O'Neil, Weapons of Math Destruction, Allen Lane, Penguin Random House, 2016.

[11] S. Lowry and G. Macpherson, “A blot on the profession,” *British Medical Journal,* vol. 296, no. 6623, p. 657, 1988.

[12] J. Leber, “The Machine-Readable Workforce: Companies Are Analyzing More Data to Guide How They Hire, Recruit, and Promote Their Employees,” MIT Technology Review, 27 May 2013. [Online]. Available: https://www.technologyreview.com/2013/05/27/178320/the-machine-readable-workforce/.

[13] A. Glaser, “Facebook’s Face-ID Database Could Be the Biggest in the World.,” Slate, 09 July 2019. [Online]. Available: https://slate.com/technology/2019/07/facebook-facial-recognition-ice-bad.html. [Accessed 07 April 2020].

[14] “HP camera 'can't see' black faces,” BBC news, 24 December 2009. [Online]. Available: http://news.bbc.co.uk/1/hi/technology/8429634.stm. [Accessed 07 April 2020].

[15] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability, and Transparency*, New York, 2018.

[16] R. Puri, "Mitigating Bias in AI Models," IBM, 06 February 2018. [Online]. Available: https://www.ibm.com/blogs/research/2018/02/mitigating-bias-ai-models/. [Accessed 07 April 2020].

[17] I. D. Raji and J. Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," in *Conference on Artificial Intelligence, Ethics, and Society*, Honolulu, 2019.

[18] E. Carter, "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent," NBC news, 12 March 2019. [Online]. Available: https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921. [Accessed 07 April 2020].

[19] J. R. Smith, "IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems," IBM, 15 February 2019. [Online]. Available: https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/. [Accessed 07 April 2020].

[20] D. Ingold and S. Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?," Blooberg, 21 April 2016. [Online]. Available: https://www.bloomberg.com/graphics/2016-amazon-same-day/. [Accessed 03 April 2020].

[21] B. Tuque, "'Creative ... motivating' and fired," The Washington Post, 6 March 2012. [Online]. Available: https://www.washingtonpost.com/local/education/creative--motivating-and-fired/2012/02/04/gIQAwzZpvR_story.html. [Accessed 4 April 2020].

[22] E. Brown, "Education researchers caution against using students' test scores to evaluate teachers," The Washington Post, 12 November 2015. [Online]. Available: https://www.washingtonpost.com/local/education/education-researchers-caution-against-using-value-added-models--ie-test-scores--to-evaluate-teachers/2015/11/12/72b6b45c-8950-11e5-be39-0034bb576eee_story.html. [Accessed 6 April 2020].

[23] American Statistical Association, "ASA Statement on," 8 April 2014. [Online]. Available: https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf. [Accessed 06 April 2020].

[24] G. Rubinstein, "Analyzing Released NYC Value-Added Data Part 1," Teach for Us, 26 February 2012. [Online]. Available: https://garyrubinstein.teachforus.org/2012/02/26/analyzing-released-nyc-value-added-data-part-1/. [Accessed 6 April 2020].

[25] N. Byline, "When standardized test scores soared in D.C., were the gains real?," The Hechinger Report, 28 March 2011. [Online]. Available: https://hechingerreport.org/when-standardized-test-scores-soared-in-d-c-were-the-gains-real/. [Accessed 6 April 2020].

[26] G. F. Brandenburg, "DCPS Administrators Won't or Can't Give a DCPS Teacher the IMPACT Value-Added Algorithm," G F Brandenburg's Blog on Wordpress, [Online]. Available: https://gfbrandenburg.wordpress.com/2011/02/27/dcps-administrators-wont-or-cant-give-a-dcps-teacher-the-impact-value-added-algorithm/. [Accessed 6 April 2020].

[27] N. O. S. R.-T. Heidi Albisser-Schleger, "Avoiding bias in medical ethical decision-making. Lessons to be learnt from psychology research," *Medicine, Health Care and Philosophy,* vol. 14, no. 2, pp. 155-162, 2011.

[28] Editorial, "Cardiology's problem women," *The Lancet,* vol. 393, no. 10175, p. 959, 2019.

[29] C. Wilkinson, O. Bebb, T. B. Dondo, T. Munyombwe and B. Casadei, "Sex differences in quality indicator attainment for myocardial infarction: a nationwide cohort study," *British Medical Journal: Heart,* vol. 105, no. 7, pp. 516-523, 2019.

[30] J. Wu, C. P. Gale, M. Hall, T. B. Dondo, E. Metcalfe and G. Oliver, "Impact of initial hospital diagnosis on mortality for acute myocardial infarction: A national cohort study," *European Heart Journal: Acute Cardiovascular Care,* vol. 7, no. 2, pp. 139-148, 2018.

[31] R. Mehran, B. Vogel, R. Ortega, R. Cooney and R. Horton, "The Lancet Commission on women and cardiovascular disease: time for a shift in women's health," *The Lancet,* vol. 393, no. 10175, pp. 967-968, 2019.

[32] Department for Transport, "Road Use Statistics Great Britain 2016," Depart, 07 April 2016. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/514912/road-use-statistics.pdf. [Accessed 08 April 2020].

[33] C. C. Perez, Invisible Women: Exposing Data Bias in a World Designed for Men, Chatto & Windus, 2019.

[34] G. Ferrant, L. M. Pesando and K. Nowacka, "Unpaid Care Work: The missing link in the analysis of gender gaps in labour outcomes," OECD Development Centre, December 2014. [Online]. Available: https://www.oecd.org/dev/development-gender/Unpaid_care_work.pdf. [Accessed 08 April 2020].

[35] R. Thompson, "We aren't collecting data about women and it's literally putting their lives at risk," Mashable, 05 March 2019. [Online]. Available: https://mashable.com/article/caroline-criado-perez-invisible-women-gender-data-gap/?europe=true. [Accessed 08 April 2020].

[36] S. Bush, "Of course algorithms are racist. They're made by people," New Statesman, 25 January 2019. [Online]. Available: https://www.newstatesman.com/politics/staggers/2019/01/course-algorithms-are-racist-they-re-made-people. [Accessed 03 April 2020].

[37] T. Simonite, "The Best Algorithms Struggle to Recognize Black Faces Equally," Wired, 22 July 2019. [Online]. Available: https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/. [Accessed 03 April 2020].

[38] N. Bostrom, Superintelligence, Oxford: Oxford University Press, 2014.

[39] National Grid, "Jack Kelly: using Artificial Intelligence to address climate change," National Grid, 28 January 2020. [Online]. Available: https://www.nationalgrid.com/stories/engineering-innovation/jack-kelly-using-artificial-intelligence-address-climate-change. [Accessed 07 August 2020].

OUR MISSION

TO UNLEASH INNOVATION AND OPEN NEW MARKETS TO CAPTURE THE CLEAN GROWTH OPPORTUNITY.

ENERGY SYSTEMS CATAPULT
7TH FLOOR, CANNON HOUSE,
18 PRIORY QUEENSWAY,
BIRMINGHAM, B4 6BS.

ES.CATAPULT.ORG.UK
@ENERGYSYSCAT