University of Southampton

Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

Putting AI ethics to work: Are the tools fit for purpose for SMEs?

by

Jacqueline Ayling

ORCID ID https://orcid.org/0000-0002-2450-0881

Thesis for the degree of Doctor of Philosophy

December 2021

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences

School of Electronics and Computer Science

Doctor of Philosophy

Putting AI ethics to work: Are the tools fit for purpose for SMEs?

by

Jacqueline Ayling

Bias, unfairness and lack of transparency and accountability in Artificial Intelligence (AI) systems have raised concerns about the ethical impact, and unintended consequences of new technologies for society across every sector where data-driven innovation is taking place. This thesis first reviews the landscape of proposed ethical frameworks with a focus on those which go beyond high-level statements of principles and offer practical tools. It then provides an assessment of these practical ethics tools through the lens of known best practices for impact assessment and audit of technology. It reviews other historical uses of risk assessments and audits to create a typology that allows us to compare current AI ethics tools to best practice found in previous methodologies from technology, environment, privacy, finance, and engineering. It analyses current AI ethics tools and their support for diverse stakeholders and components of the AI development and deployment lifecycle. Building on this analysis, a series of interviews were conducted with CEO's/founders of smaller tech companies to understand how these tools might be used (or not) in the production of real products and services. This uncovers a narrower conception of ethical concerns and stakeholders in the sector than presented in AI ethics tools and principles. The sector also understands itself as already taking the necessary steps to address ethical issues without the need for specific ethical tools or governance. From this, gaps are identified in current AI ethics tools and their practical application that should be considered going forward.

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Jacqueline Ayling

Title of thesis: **Putting AI ethics to work: Are the tools fit for purpose for SMEs?**

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

Putting AI ethics to work: are the tools fit for purpose? (Ayling And Chapman 2021) Journal of Ethics and AI https://link.springer.com/article/10.1007/s43681-021-00084-x

Signature: Date:24/12/2021

Acknowledgements

To my long suffering and ever patient supervisor, Prof Age Chapman, who I would like to thank for being a faithful guide in what has been a difficult journey. Gratitude is also due to Prof Les Carr and Dr Nick Gibbins for always being kind and supportive, and Prof Kieron O'Hara who offered invaluable guidance during my internal examinations. Thanks to my second supervisors Prof Matt Ryan, Prof Sophie Stalla-Bourdillon and Dr Christopher Hamerton. Thanks also to the admin team, Alison Tebbutt and Megan Chan - always friendly and efficient without whom things would definitely fall apart.

I am also grateful to my colleagues on the journey, Dr Sarah Hewitt, Dr Johanna Walker, Dr Manuel Leon Urrutia, Dr Sami Kanza, Dr Mark Anderson and Dr Richard Gomer for being good friends, good listeners and always willing to offer help and advice. Special thanks to my virtual office team - Gareth Leggett in sales, Alex Templeton in projects, and the writing lab team, Rachel Hayward, Mike Hoffman and Bernard Roper - who have kept me sane and focused through the long haul of writing up during a pandemic.

I must apologise to my friends and family who have had years of listening to me droning on about my PhD. I am eternally grateful for their good humour and support throughout, especially Dr Alison Pearce. Special thanks also go to my son Bernard who had no choice but to come on this journey with me. I am sure he hopes I will stop talking about data and ethics sometime soon.

Dedication

This thesis is dedicated to my parents. To my Mother for always being there to support me. And to my late Father, Clifford Ayling (1933-2020), who died in the first wave of the corona virus pandemic in his care home alone. Miss you Dad. Wish you were here to see me succeed and wish I could have been there at the end.

XLVII

When You and I behind the Veil are past,
Oh, but the long, long while the World shall last,
Which of our Coming and Departure heeds
As the Sea's self should heed a pebble-cast.

Rubáiyát of Omar Khayyám (Trans. Edward FitzGerald 1859)

Definitions and Abbreviations

BEIS	Department for Business, Energy & Industrial Strategy UK
CBA	Cost-Benefit Analysis
CDEI	UK Centre for Date Ethics and Innovation
DPIA	Data Protection Impact Analysis
EIA	Environmental Impact Assessment
ERA	Environmental Risk Assessment
FIA	Financial Impact Assessment
eTA	Ethical Technology Assessment
FIP	Fair Information Practices
FTC	US Federal Trade Commission
GAAP	Generally Accepted Accounting Principles
GDPR	EU General Data Protection Regulation
HCI	Human-computer Interaction
HRIA	Human Rights Impact Assessment
HUDERIA	Human Rights, Democracy, and Rule of Law Impact Assessment
ICT	Information Communication Technology
IFRS	International Financial Reporting Standards
IP	Intellectual Property
PD	Participatory Design
pTA	Participatory Technology Assessment
SIA	Social Impact Analysis
TA	Technology Assessment

Chapter 1 Introduction

1.1 Context

Ethics for artificial intelligence (AI) has been experiencing something of a gold rush in the last few years, with frameworks, guidelines and consultations appearing thick and fast from governments, international bodies, civil society, business and academia. Bias, unfairness and lack of transparency and accountability in AI systems, and the potential for the misuse of predictive models for decision-making have attracted attention across a range of domains from predictive policing to targeted marketing to social welfare (Diakopoulos, 2016; Eubanks, 2018). There is disquiet about the ethical impact and unintended consequences of new technologies for society across every sector where data-driven innovation is taking place, and an increasing recognition that even the latest updates to data protection regulation (e.g. the General Data Protection Regulation GDPR (European Council and Parliament, 2016)) are not addressing all the ethical issues and societal challenges that arise from these new data pipelines and computational techniques.

This thesis sets out to review the landscape of suggested ethical frameworks with a focus on those which go beyond high-level statements of principles (see (Hagendorff, 2019; Jobin, Ienca and Vayena, 2019; Fjeld *et al.*, 2020) for review of principles), and offer practical tools for application of these principles in the production and deployment of systems. 'Efforts to date have been too focused on the 'what' of ethical AI (i.e. debates about principles and codes of conduct) and not enough on the 'how' of applied ethics' (Morley *et al.*, 2019, p. 2143). We can all nod our heads sagely in agreement with principles like fairness and justice, but what does fairness and justice look like in a real-life decision-making context? How are organisations and those within them to reckon with the complex ethical tug-of-war between 'the bottom-line' and upholding ethical principles? To answer these questions (after an initial exploratory case study) the research uses document analysis to uncover the features of proposed tools for operationalising ethical principles for AI (as opposed to statements of ethical principles.) The analysis uses a range of typologies drawn from best practice in well-established impact and risk assessment, and audit domains, that have been used to manage human activities and new technology. The research then gathers the opinion and responses of high-level decision makers in technology companies to gauge if, and how, these tools might work in the real world.

Societies face a series of complex and difficult problems across multiple domains to which the application of data-driven AI technologies is being eagerly pursued. The ability to collect and store vast troves of data, coupled with increases in computational power, provides the substrate for an

explosion of AI applications, particularly machine learning. The kinds of harms that have been of growing concern build on traditional data privacy harms (see for example (Solove, 2006; Citron and Solove, 2021)). Concerns around AI are grouped firstly around epistemic concerns (the probabilistic nature of insights, the inherent inscrutability of 'black box' algorithms, and the fallibility of the data used for training and input). Then there are normative concerns about the fairness of decisional outcomes, erosion of informational privacy, and increasing surveillance and profiling of individuals. Algorithmic systems also create problems of accountability and moral responsibility, where it is unclear which moral agent in the process bears (or shares) responsibility for outcomes from a system (Mittelstadt *et al.*, 2016).

Disastrous outcomes like the loss of human life through machine malfunction (think medical applications or autonomous cars), or the hijacking and manipulation of critical systems by bad actors (think military systems, or smart city technologies controlling essential services). These kinds of outcomes pose significant challenges for both government and business and could result in reputational damage, regulatory backlash, criminal proceedings and a loss of public trust (Hirsch *et al.*, 2020). As Daniel Solove presciently noted we risk creating a Kafkaesque world with 'a more thoughtless process of bureaucratic indifference, arbitrary errors, and dehumanization, a world where people feel powerless and vulnerable, without any meaningful form of participation in the collection and use of their information' (Solove, 2001, p. 1398). It is to meet these challenges that the current interest in ethical frameworks has become so heightened.

In response to increasing public debate and political concern about the negative effects on individuals and wider society of AI, a veritable AI ethics industry has emerged, promoting a variety of different frameworks and tools (Raab, 2020). Several authors (Greene, Hoffmann and Stark, 2019; Morley *et al.*, 2019; Kazim and Koshiyama, 2020; Kind, 2020; Raab, 2020; Ryan and Stahl, 2020) have identified different phases in the response to increasing public debate about the impact of AI technologies. In the first phase from 2016 to 2019 many high level ethical principles for AI were published as evidenced by these catalogues of ethical principles and frameworks for ethical, trustworthy responsible AI (Hagendorff, 2019; Jobin, Ienca and Vayena, 2019; AlgorithmWatch, 2020; Fjeld *et al.*, 2020; Schiff, 2020). This first phase focused on the high-level ethical principles that might best address the impacts of AI and data-driven systems, framed as applied ethics and dominated by a philosophical approach as opposed to a legal or technical approach.

A second phase saw a more technical approach from the computer science community focusing on fairness, accountability and transparency as an engineering 'ethical-by-design' problem-solving exercise (Mitchell *et al.*, 2019; Bird *et al.*, 2020; Gebru *et al.*, 2020). The current phase is seeing a move 'from what to how' (Morley *et al.*, 2019), with proposals for governance mechanisms,

regulation, impact assessment, auditing tools and standards leading to the ability to assure and ultimately, insure AI systems (Kazim and Koshiyama, 2020). There is also latterly a shift towards acknowledgement of political, social and justice issues which move 'beyond the principled and the technical, to practical mechanisms for rectifying power imbalances' (Kind, 2020). As Crawford (2021) argues, AI ethics is not just a 'tech ethics' problem, amenable to 'tech ethics' fixes, but raises deeply political questions about how power is wielded through technology.

Meta-analyses of AI ethics proposals have thus far focused mainly on classifying and comparing ethical principles, where some convergence has been identified for principles like transparency, fairness, privacy and responsibility (Hagendorff, 2019; Jobin, Ienca and Vayena, 2019; Fjeld *et al.*, 2020). What is less clear and needs investigation are other variables for these proposals like scope, applicable context, ownership of or responsibility for the process, method of implementation and representation of stakeholders. There are already established governance methodologies for assessing and mitigating the impact of new technologies, processes, and infrastructure across the domains of environment (Morgan, 2012), information privacy (Clarke, 2009), data protection (ICO UK, 2018) and human rights (The Danish Institute for Human Rights, 2016). These impact assessment and audit methodologies take core societal values and combine them with a process for the public, outside experts, and policymakers to consider complex social and technical questions. This thesis explores how best practice from other domains can give us insight into proposed frameworks for managing risk in AI, and how these processes are viewed and applied in the industry.

1.2 Research questions

I have a background in environmental management techniques, and have previously been trained to conduct ISO:14000 audits (International Organisation for Standardization, 2021). Reflecting on the processes used to provide assurance for the environmental impact of companies, I considered there to be parallels in the need to implement processes to assure the ethical design and deployment of AI systems. This background knowledge informed the decision to reflect on impact assessment and audit processes in other domains, many of which have a long lineage. I wanted to understand better the features of the current proposed tools for implementing AI ethics, to assess if these tools are fit for purpose, and where gaps might be illuminated by previous practice.

I was interested in particularly how these tools might be applied in the context of small and medium sized enterprises (SMEs) as this is where much technology innovation occurs, yet these companies have constrained financial resources and often lack formalised governance procedures and in-house expertise.

This led to an overall research question:

RQ0: Are the AI ethics tools being proposed fit for purpose for use by SMEs?

After an initial case study to explore data flows and potential ethical challenges in geospatial products (Chapter 4) the following research questions were formulated to answer the overall question:

RQ1: What practical tools are being proposed to operationalise AI ethics?

RQ2: What features do these tools have when compared to existing practices in other domains?

RQ3: How are these tools understood and used by senior decision-makers in SMEs?

1.3 Thesis outline

A background literature review is presented in Chapter 2. This is followed by the methodology in Chapter 3 which describes the process for a case study reported in Chapter 4, a document analysis reported in Chapter 5 and interviews reported in Chapter 6.

After an initial case study described in Chapter 4, a methodology was developed (see Section 3.3) to identify the gaps in current mechanisms by analysing the AI ethics tools using a set of typological schemas. These were developed by conducting a review of previous best practice across different domains and a review of current discussions around AI governance (see Chapter 2). An extensive document search process was undertaken for current tools for improving ethical practices in AI technology. This provided the data set for analysis using the typologies drawn from a review of the development of impact assessment and audit across a range of domains, and the key components as they related to understanding impact across participants, technology, and processes. Using the typologies created from the review of previous practice, current AI frameworks are analysed using these criteria to identify the gaps in current approaches.

Having understood in detail the current proposals for AI ethics tools (see Chapter 5), a series of in-depth interviews with senior decision-makers in SME's/start-ups was used to reflect how these proposed tools might be used in the production pipeline of the technology (see Chapter 6). The interviews give voice to the views of senior decision-makers in the commercial production of AI systems, to reveal the kind of processes they currently deploy to address ethical issues they perceive, to elicit commentary on example AI ethics tools, and to reflect on potential regulation in the sector.

The implications of these results are discussed in Chapter 7, with conclusions being drawn in Chapter 8.

1.4 Research contributions

This paper contributes to the literature by mapping the current landscape of suggested tools for ethical assessment of AI systems, placing these tools in a historical tradition of managing the impacts of technology, thereby exposing possible areas for strengthening these tools in practice. It also provides useful feedback from those who might be expected to apply these tools in real world contexts of building, selling and procuring AI systems.

1.5 Publications and related work

Putting AI ethics to work: are the tools fit for purpose? (Ayling And Chapman 2021) Journal of Ethics and AI https://link.springer.com/article/10.1007/s43681-021-00084-x

Sustainability of (open) Data Portal Infrastructures - Developing Microeconomic Indicators through Open Data Reuse (2020) - European Open Data Portal https://www.europeandataportal.eu/sites/default/files/sustainability-data-portal-infrastructure_2_developing-indicators.pdf

How to use AI for Good – the Ethical and Societal Implications for using AI for Scientific Discovery (workshop) (2020) ACM Web Science Conference https://sites.google.com/site/ai3sdusingaiforgood/home

Ethical Data-Driven Technologies (2020) Web and Internet Science Research Group Lecture Series, University of Southampton https://www.wais.ecs.soton.ac.uk/

Algorithmic Accountability and the Role of Provenance (2018) conference Paper, Provenance Week 2018 https://sociam.github.io/saap-workshop/

Chapter 2 Background

2.1 Definition of AI

The term Artificial Intelligence (AI) is slippery and is commonly used as a portmanteau word for a range of computational techniques and domains of application. It has been used of late as a shorthand to point to digital technologies more generally (if not accurately). It also carries social and cultural meanings outside of technical debates through its representation in literature and the arts. Historically, AI has denoted techniques like theorem proving, heuristic search, game playing, expert systems, neural networks, Bayesian networks, data mining, agents, and deep learning. These techniques are applicable to different kinds of problem and have led to the development of a range of subdomains like knowledge representation, reasoning, planning, machine learning, vision, natural language processing and robotics (Wang, 2019). As a report by the Office for Statistics Regulation notes 'terms such as statistical model, statistical algorithm, data-driven algorithms, machine learning, predictive analytics, automated decision making and artificial intelligence (AI), are frequently used interchangeably, often with different terms being used to describe the same process. The findings of this review apply to all these data-driven approaches to supporting decisions in the public sector whatever the context' (Office for Statistics Regulation, 2021, p. 7). The term AI will be used in this thesis to denote computational systems that use data to predict, categorise or model the world, but with the understanding that this term does not just capture the mathematical model or the data, but understands AI as an embedded sociotechnical system (Baxter and Sommerville, 2011).

2.2 Impact Assessment Practices

Ethical tools and frameworks for AI do not spring like Dionysus fully formed from Zeus' thigh, they are part of a development of governance tools to tackle health, environmental and privacy impacts of technology that began in the 1960's. Impact and risk assessment is 'a type of fact-finding and evaluation that precedes or accompanies research, or the production of artefacts and systems, according to specified criteria. Assessing the impact of some X upon some Y has been practiced for generations, and has engendered debates over methods, purpose, focus, policy relevance, terminology, and efficacy' (Raab, 2020, pp. 6–7). These assessments are shaped by notions of relevance (what is important to society and which phenomena are worthy of attention), evidence

(identification of causes and effects), and normative claims (what is good, acceptable or tolerable) (Renn, 2008, p. 4).

2.2.1 Technology Assessment

Technology assessment (TA) is a practice that began with the US Office of Technology Assessment (OTA) 1972-1995 (Coates, 1974; IAIA, 2009). TA was 'foremost an attempt to gain political control over the potential negative effects of technological development by means of early warnings. TA was supposed to predict unintended negative consequences of technical innovations in order to facilitate more adequate policy-making' (Palm and Hansson, 2006, p. 544). In the 1990's Europe also developed its own TA institutions like the Scientific Technological Options Assessment (STOA) and recent activities include setting up the STOA Centre for AI (STOA, 2021). Several different varieties of TA have been developed, for example in the Netherlands and Denmark TA was extended to address issues of participation. Instead of the traditional TA model with panels of experts producing reports for policy-makers, participatory TA (pTA) includes contributions from a much wider group of stakeholders like lay people, journalists, trade unions and civil society groups (Hennen, 2012). pTA uses various forms of public deliberation including focus groups, citizens' assemblies and consensus conferences to gather data for reporting (CSPO, 2021).The lack of an ethical dimension to TA has also led to suggestions for an ethical TA (eTA) (Palm and Hansson, 2006; Kiran, Oudshoorn and Verbeek, 2015), which mirror many of the concerns found in AI ethics frameworks (Jobin, Ienca and Vayena, 2019).

2.2.2 Environmental Impact Assessment

Environmental Impact Statements (EIS) were pioneered in the US by the 1969/70 National Environmental Policy Act (NEPA), leading to many other jurisdictions enacting environmental legislation Environmental Impact Assessments (EIA) broadened the process to include the identification of future consequences and included in the process public consultation and review mechanisms (Clarke, 2009). '[T]he role of the stakeholders or parties at interest plays such a critical role in technology assessment, and involvement of citizens in environmental impact statements is mandated by law' (Coates, 1974, p. 374). These assessments are part of many jurisdictions planning and/or environmental legislation, intended to allow stakeholders, including the public in its widest definition, to contribute to decision-making on infrastructure development like dams and roads (Suter, Barnthouse and O'Neill, 1987; UN Environment, 2018). It should be noted though that there is a lack of clear definition in EIA literature and practice as to what 'participation' actually means (Glucker *et al.*, 2013). There are also specific assessment techniques for products and materials to assess environmental impact which map life cycles (IMA Europe, 2020).

Figure 1 below illustrates the typical process model for conducting an EIA.

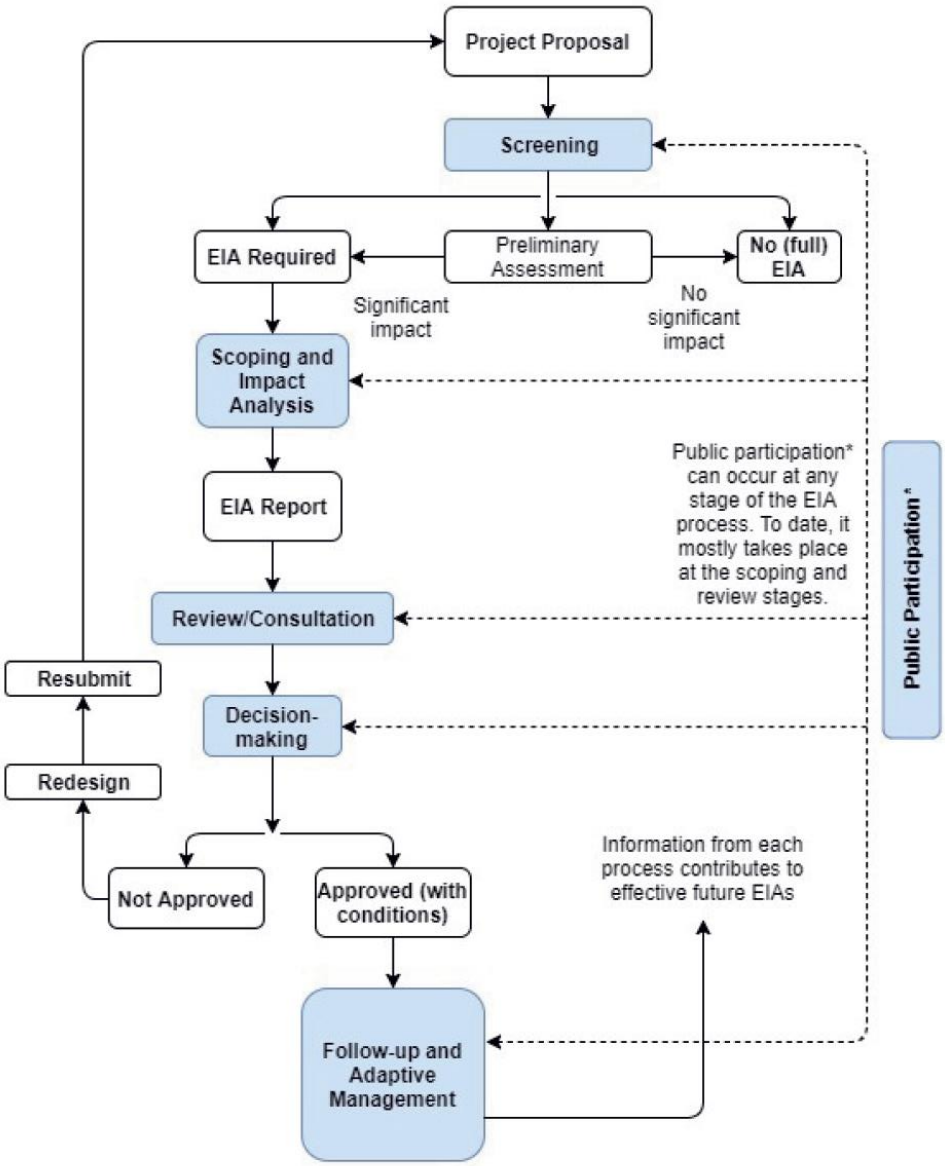


Figure 1 UN Environment Agency process model for EIA
(UN Environment, 2018, p. 32)

Environmental Risk Assessment (ERA) developed as a separate practice from the broader scope of EIAs, using formal quantitative analysis of probabilities for undesirable outcomes of a process or substance. Environmental risk assessment focuses on specific regulatory problems like air or water quality, or the impact of specific pollutants on human health (Suter, Barnthouse and O'Neill, 1987; Aven, 2016). Risk assessments can be included in the contents of an EIA as part of the evidence gathering for specific impacts. Fiscal Impact Analysis (FIA) is an economic impact tool commonly used in land use planning decisions (Edwards and Huddleston, 2009), and EIAs often include forms of cost-benefit analysis (CBA) (Pearce, 2016). CBA has a long history and increasingly complex methodologies, but in essence is used a tool to determine if the benefits of a proposed project

outweigh the costs, and if benefit can be established, what is the magnitude of this benefit to society (Mishan and Quah, 2020). This approach has caused dispute since the 1970's to the present about how economic values can or should be attached to bio-physical resources, see for example the recent Dasgupta Review 'The Economics of Biodiversity' (Dasgupta, 2021). There are parallels in the economic imperatives driving the AI industry (and the arguments against regulation), where social values (like privacy or autonomy) and social structures (like democracy) that do not lend themselves to a rational assignment of economic value (Ackerman and Heinzerling, 2001).

2.2.3 Social and Human Rights Impact Assessment

EIAs and ERAs were criticized for focusing on only bio-physical and economic impacts and not including the social and cultural impacts of proposed developments or technologies, leading to the development in the 1990's of Social Impact Assessments (SIA). SIA is not a widely applied form of assessment 'largely because of the challenge of defining, predicting and measuring social change and impact, in addition to legal and regulatory frameworks that are persistently weak or ineffectual in terms of social impact' (Kemp and Vanclay, 2013, p. 91). They remain fairly uncommon, but have been used in policy impact assessments, for example, by the IMF to try and understand the impact of macro-economic policy changes (Kende-Robbe, 2003).

Human Rights Impact Assessments (HRIA's) have been suggested as impact assessment tools to build on data protection to assess the impact of algorithmic systems on human rights (Mantelero, 2018). The Council of Europe's Ad Hoc Committee on AI (CAHAI) (CAHAI, 2020) suggests a Human Rights, Democracy, and the Rule of Law Impact Assessment (HUDERIA) which should be used by AI developers to meet their human rights due diligence obligations. HRIA's have not been mandated in, for example, the draft EU AI Act (European Commission, 2021) despite the Act being based on fundamental rights (ECNL and Data & Society, 2021).

2.2.4 Privacy and Data Protection Impact Assessment

The concept of privacy, which underpins modern data protection legislation, is essentially normative and represents the cultural and historical values of societies. In the Western tradition there are two core assumptions, the first appealing to a 'natural' divide of the public (the state and politics, work and business) and the private realm (the realm of the home, family, body and personal property, where the individual is considered the best judge of their privacy interests. The second assumption posits privacy as a prerequisite for the liberal democratic state. There are shifting social norms around the value and definition of privacy, with debates revealing tensions,

for example, between the goals of privacy vs security and privacy vs economic growth (Roessler, 2008).

The 'fair information practices' (FIP) movement emerged in the US in the work of Westin (Westin, 1967, 1971), in response to growing societal concerns over the collection and processing of personal data in both the public and private sector. It was not until the mid-1990's that Privacy Impact Assessments (drawing on the model of EIAs) emerged in various forms across different jurisdictions (Stewart, 1996). By 2007 the UK Information Commissioners Office published a handbook describing a methodology for conducting a PIA (Clarke, 2011), which was further developed in Europe into a Data Protection Impact Assessment (DPIA), a key tool in the latest iteration of data protection regulation, the GDPR (European Council and Parliament, 2016)

Privacy impact assessments were developed to meet the need for public trust in information processing by identifying and managing risks. This is part of a wider move in industrialised societies to manage potential risks of new technologies, processes or products that can also be seen in TA and EIA (Raab, 2020). DPIA's use checklists and risk assessments to document the data processing and any necessary mitigations if risks are identified in an iterative review process (ICO UK, 2018).

2.3 Audit Practices

There are long established techniques for auditing processes and systems, for example in the financial sector where there are globally agreed standards like the Generally Accepted Accounting Principles (GAAP) and International Financial Reporting Standards (IFRS) (Financial Reporting Council, 2020). These rules lay down the process for transparent 3rd party auditing which have been adopted into law by the majority of jurisdictions around the world. There are also audit and assurance standards in safety critical engineering for industries like aviation, nuclear power, or more recently, autonomous vehicles (Rusby, 2015; Bloomfield *et al.*, 2019).

Audit techniques are also used for third-party verification for accreditation to international industry standards e.g. International Organization for Standardization (ISO) (International Organization for Standardization, 2021b). An audit consists of the examination of evidence of a process or activity, like financial transactions or an engineering process, and then evaluation of the evidence against some standards or metrics, which could be a regulation or standards regime (International Organization for Standardization, 2021a), or internal management metrics (PwC UK, 2013; Financial Reporting Council, 2020), as illustrated in Figure 2.

Figure 2 Standard stages in an audit

In order to conduct an audit, there first needs to be a set of auditable artifacts that record decisions, systems and processes. Brundage et al. (2020, p. 24) define as this as problem space for current AI production in that they 'lack traceable logs of steps taken in problem-definition, design, development, and operation, leading to a lack of accountability for subsequent claims about those systems' properties and impacts'. This is where some of the technical tools addressing AI ethics can become part of an audit process by providing evidence for evaluation by auditors. Audit also requires non-technical governance processes (Kazim and Koshiyama, 2020) to ensure consistency with relevant principles or norms (Mökander and Floridi, 2021).

Impact assessments like EIA, and audits such as those conducted in the finance sector have well established protocols regulated by legal requirements. Independence of assessment and audit is used to ensure transparency and places liabilities on both the parties assessing and the assessed parties. 'Whether the auditor is a government body, a third-party contractor, or a specially designated function within larger organisations, the point is to ensure that the auditing runs independently of the day-to-day management of the auditee' (Mökander and Floridi, 2021, p. 2). External assessment provides publicly available documents which can also serve a broader range of stakeholders beyond the entity or process in question to include users, customers and wider society.

2.4 Risk assessment and techniques

While a myriad of processes, tools and applications of these tools at various parts of the production cycle exist across the historical impact assessment and audit activities above, one of the key elements is risk assessment.

Modern conceptions of risk (risk = accident x probability) became a fully-fledged part of modern societies with the risk assessment practices developed in response to concerns over the impact on the environment and human health from human activity in the form of development, technologies and industrial processes and materials. In 1969, in an article entitled 'What is our society willing to pay for safety?' (Starr, 1969), articulated a systematic and quantitative approach to risk, and

introduced the concept of trade-offs between risks and benefits (Thompson, Deisler and Schwing, 2005). Debates within the environmental movement, and the associated legal and organisational structures that grew out of this period, came to be famously characterised by Beck in the 1980's as the 'Risk Society' (Beck, 1992). Beck posited that the project of modernity had become not how to distribute wealth or goods, but how to distribute the risks, or 'bads', of modern industrial society, where technical experts are given pole position to define agendas and impose bounding premises a priori on risk discourses' (Beck, 1992, p. 4).

A risk-based approach has developed throughout the latter part of the 20th and into the 21st century, taking the methodology and approaches from environmental management and risk assessment and applying them to areas like occupational health and safety, business risk (financial, operational, reputational), quality, and information security. Risk assessment techniques vary from quantitative to qualitative approaches depending on the sector and application (Aven, 2016). Risk assessments often rely on scoring or 'traffic light' systems for ranking risks (Moses and Malone, 2004), and highlighting those areas that need treatment, either in the form of mitigation (changing the risk score) or in taking measures (like insurance) or documenting decisions to 'trade off' the risk against the potential benefits. Risk assessments are also used for achieving compliance with the existing regulatory frameworks. The latest European iteration of data protection (GDPR) also takes a risk-based approach to privacy protections for data subjects. Many of the ethical frameworks proposed for AI build on these models and approaches to risk assessment.

For the business sector managing reputational risk is an important consideration and providing evidence of responsible behaviour has direct links to both users/customers and also to investors and boards. Many investors use Environmental Social and Governance (ESG) assessments, where they look for evidence of compliance with international standards and norms, where the risks (especially reputational) could impact across all three areas of ESG assessments for investors. Business-focused AI ethics tools fall into the suite of tools organisations deploy to protect their core value. Managing risk allows institutions to 'adopt procedures and self- presentation in order to secure or repair credibility' (Beck, 1992, p. 4), a core purpose of contemporary risk management strategies (Hayne and Free, 2014).

Risks in AI can manifest as either underusing the technology and missing out on value creation and innovation, or overusing/misusing the technology. Floridi et. al. (2018) draw attention to risk that results from not using the technology, and how these risks need careful trade-offs to ensure the greatest benefit. As Jobin et. al. (2019) note in their systematic review of global AI guidelines, conflicts can be identified in the different proposals 'between avoiding harm at all costs and the perspective of accepting some degree of harm as long as risks and benefits are

weighed against each other. Moreover, risk–benefit evaluations are likely to lead to different results depending on whose well-being will be optimized for and by which actors. Such divergences and tensions illustrate a gap at the cross-section of principle formulation and their implementation into practice' (Jobin, Ienca and Vayena, 2019, p. 396).

2.5 Stakeholder Theory and Participation

The influential European Commission's report on 'Trustworthy AI' proposes that 'management attention at the highest level is essential to achieve change. It also demonstrates that involving all stakeholders in a company, organisation or institution fosters the acceptance and the relevance of the introduction of any new process (whether technological or not). Therefore, we recommend implementing a process that embraces both the involvement of operational level as well as top management level' (High Level Expert Group on AI, 2019, p. 25). A wide-ranging network of stakeholders can be plotted in the production and deployment of new technologies that extend far beyond the domain of engineers and developers (see Table 3 Typology of stakeholders).

Since the development in the 1980's of corporate stakeholder theory (Freeman, 2010) it has become common parlance to refer to 'stakeholders' across a range of organizational domains. Stakeholder theory provides a well-established framework that allows us to:

1. Identify and describe all interested and affected parties in the deployment of a technology
2. Acknowledge stakeholders have legitimate interests in technology
3. Affirm that all stakeholders have intrinsic value, even if their concerns do not align with the concerns of the technology producers
4. Identify the responsibilities of parties with relation to a given process (Donaldson and Preston, 1995).

Table 3 identifies the broad categories of public and private sector stakeholders who either have direct roles in the production and deployment of AI technologies, or who have legitimate interests in the usage and impact of such technologies. Stakeholder theory has long challenged the assumption that a company's exclusive obligation is to their shareholders or investors, with business leaders increasingly recognizing the need for a wider set of obligations beyond the narrow vision of 'shareholder primacy' (Business Roundtable, 2020).

2.6 Technical and design tools

Another active space in the AI ethics debate is within the machine learning (ML) community itself where much attention and research has been focused on metrics like fairness, accountability,

explainability and transparency[1]. A range of computational approaches have been suggested, offering quantitative metrics for fairness, methods to 'debias' training data sets, test models against protected characteristics and provide explanations of 'black box' algorithms, packaged up into AI fairness toolkits (Bantilan, 2017; Bellamy *et al.*, 2018; Bird *et al.*, 2020; TensorFlow, 2020). These toolkits have been criticised for offering a 'reductionist understanding of fairness as mathematical conditions' (Lee, Floridi and Singh, 2021, p. 1), and reflect a longer history of attempts to reduce (un)fairness to a metric (Hutchinson and Mitchell, 2019). Studies with ML developers highlight that considerations of a model's context, and the specificity of the domain in which it is used, are vital in order to improve features like fairness (Veale, Van Kleek and Binns, 2018). Many would argue that in fact, developing ethical AI requires not only technical 'fixes' but the deployment of social science disciplines is vital to address negative outcomes (Veale, Van Kleek and Binns, 2018; Hoffmann, 2019; Radford and Joseph, 2020).

Other suggestions focus on design processes, for example awareness raising for design teams in workshop style events (Institute for the Future and Omidyar Network, 2018; Doteveryone, 2019), or participatory design processes (Madaio *et al.*, 2020). The human-computer interaction (HCI) community is also concerned to translate previous work in, for example, Value Centred Design, to address the issues in human-AI interactions (Stephanidis *et al.*, 2019).

2.7 'Ethicswashing'

Business orientated risk-management is premised on value creation and protection of an organisation from penalty, or reputational damage (Arena, Arnaboldi and Azzone, 2010). The adoption of an EDIA could be viewed as an attempt at 'ethics-washing' (akin to 'green-washing' for environmental concerns).

> 'With ethics-washing, a performative ethics is being practised designed to give the impression that an issue is being taken seriously and meaningful action is occurring, when the real ambition is to avoid formal regulation and legal mechanisms. It is, in effect, virtue signalling, providing empty or superficial support for a position and prioritising appearance over action. The hope is to reassure the public, policy-makers and government with respect to any concerns they might have, and in so doing, promote products and initiatives, enhance reputation and attract investment' (Kitchin, 2019).

[1] E.g. new conferences have been created like ACM FAccT https://facctconference.org/index.html and high profile conferences in the AI/ML space increasingly including work on ethical problems like NeurIPS https://neurips.cc/.

Using policies and procedures like EDIAs or DPIAs can result in a check-list mentality, that can not only fail to protect against the very risks it purports to protect against, but results in an avoidance of more fundamental normative questions about the kind of society we want to create. For example we can look to the failure of internal financial controls and regulation in the financial sector that led to the 2008 economic crash (Power, 2009). As Arena, Arnaboldi and Azzone (2010) note:

> 'The danger is that these systems become box-ticking exercises that have little effect on decision-making and outcomes.' (Arena, Arnaboldi and Azzone, 2010, p. 673)

Narrow technical and compliance concerns avoid wider questions like social justice and the public good, and locate ethical problems in individuals and systems rather than in the structural power relations where the real ethical challenges may lie (D'Ignazio and Klein, 2018).

2.8 Summary

This chapter has served as survey of existing practices across a variety of domains to manage and regulate the impacts of different technological and economic activities. It is against this landscape that the current proposed tools for implementing AI ethics will be assessed.

Chapter 3 Methodology

3.1 Overview

The research process is presented in Figure 3. The research questions in Section 1.2 were developed from an interest in the practical challenges for SME's in the tech industry in applying ethical principles to their products and services and drew from my previous experience as an auditor for environmental management standards. An initial pilot case study (see Section 3.2) was focused on mapping the data flows in a GIS insights company. The process of conducting the interview and writing up the results clarified the next stage of research which required a granular examination of proposed ethics tools (see Section 3.3). This analysis then informed the selection of key representative ethics tools for response from senior decision makers in industry (see Section 3.4).

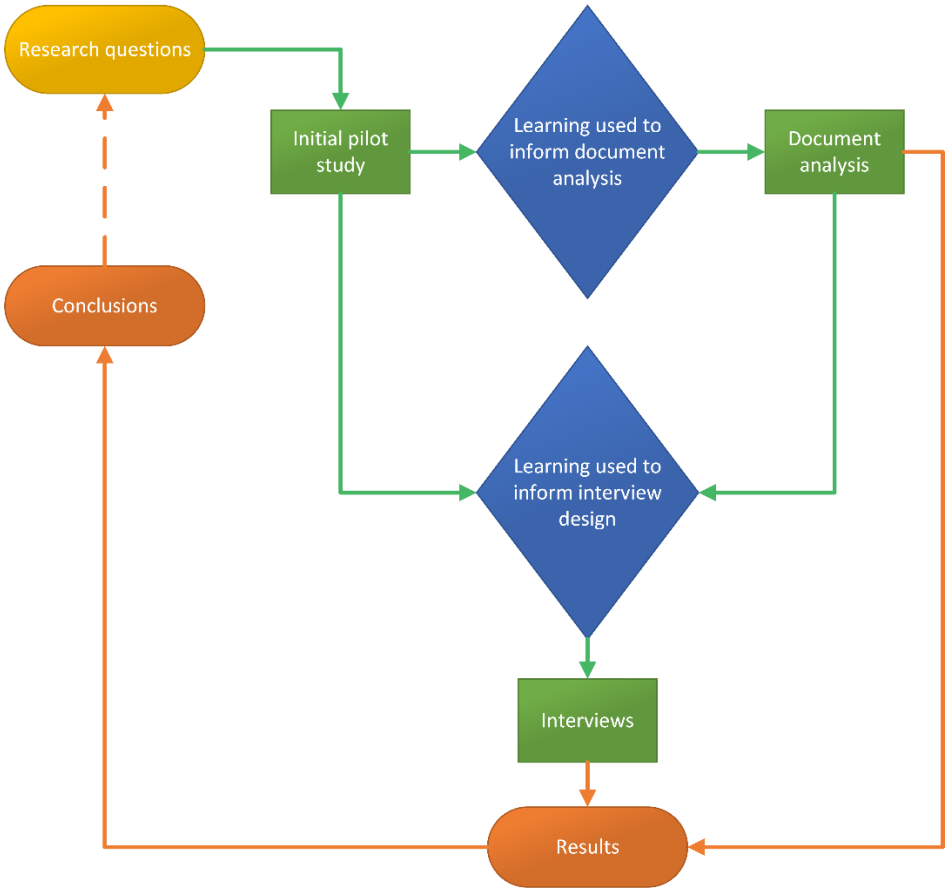


Figure 3 Overall research process diagram

3.2 Pilot case study

3.2.1 Rationale

In the first phase of the research, a pilot study was conducted to explore the ethical issues in one company. This was to test out interview formats, and potential data collection tools for the research project. Initially it was thought that perhaps mapping data flows and examining the issues through this lens might be fruitful. To this end the interview was designed around a spreadsheet to collect the information. However, this was found to be impractical and the questions from the tool were used during the interview and recorded and transcribed later.

3.2.2 Interview

A semi-structured interview (University of Southampton Ergo II application 46086) was conducted with the CEO of a small geospatial insights company based in the UK, referred to as Company X. Their company uses geospatial Big Data, Machine Learning and AI to create products and services for applications in the transition to a low carbon economy. Their business provides a range of geospatial insights for a range of public and private customers using GIS datasets (maps, satellite and aerial imagery) for renewable and sustainable energy, EVs, smart grids and energy management and sustainable mobility.

3.2.3 Limitations

The case study was an exploratory project which enabled the research to develop with a sharper focus in highlighting the need to understand the tools and principles being proposed, and for more wide-ranging interviews to be conducted at a later stage. This case study is therefore an interesting deep dive into one company but can only be used as an illustration of a set of use cases in a particular context.

3.3 Document analysis

There are a number of different ways of identifying the methods and tools available to help all stakeholders reflect on and apply ‘ethics’ when creating AI systems. For example, Vakkuri et al. (2019) sought to answer the question ‘what practices, tools or methods, if any, do industry professionals utilise to implement ethics into AI design and development?’ by conducting interviews at five companies that develop AI systems in different fields. However, whilst analysis of the interviews revealed that the developers were aware of the potential importance of ethics in AI,

the companies seemed to provide them with no tools or methods for implementing ethics. These findings did not imply the non-existence of applied ethics tools and methods, but rather a lack of progress in the translation of available tools and methods from academic literature or early-stage development and research, to real-life use.

In order to gain a richer understanding of the translation problem identified in Vakkuri et al. (2019) this thesis draws from the rich impact assessment and audit literature from other domains to develop a typology for comparative document analysis of proposed AI ethics tools. The AI ethics documents themselves provide the primary data for study, with codes being produced by a mixed methods approach, where thematic codes were developed in response to research questions, a review of related literature, and iteratively refined through examination of the documents under examination themselves. (Fereday and Muir-Cochrane, 2006; Bernard and Gravlee, 2014). Using the results from the document analysis, interviews were then conducted with senior decision-makers to investigate how these tools are understood and applied in real world situations.

The features exhibited in impact assessment and audit literature from other domains were used to develop typologies for comparative document analysis of proposed AI ethics tools. In order to understand how proposed AI ethics tools might be applied, it is first necessary to understand what they are offering, how they differ, and to identify any gaps. This understanding can be used to refine and develop these tools for future use. The AI ethics documents themselves provide the data for study, which have been analysed using qualitative content analysis, ‘a research technique for making replicable and valid inferences from data to their contexts’ (Krippendorff, 2013, p. 403). Typologies of salient features were developed in response to research questions, using a review of related literature and AI ethics documents, and iteratively refined. Typologies are useful heuristics to enable systematic comparisons (Smith, 2002), and extensive related literature was reviewed to build representative typologies for the tool types under examination which would yield useful comparisons across a diverse range of documents.

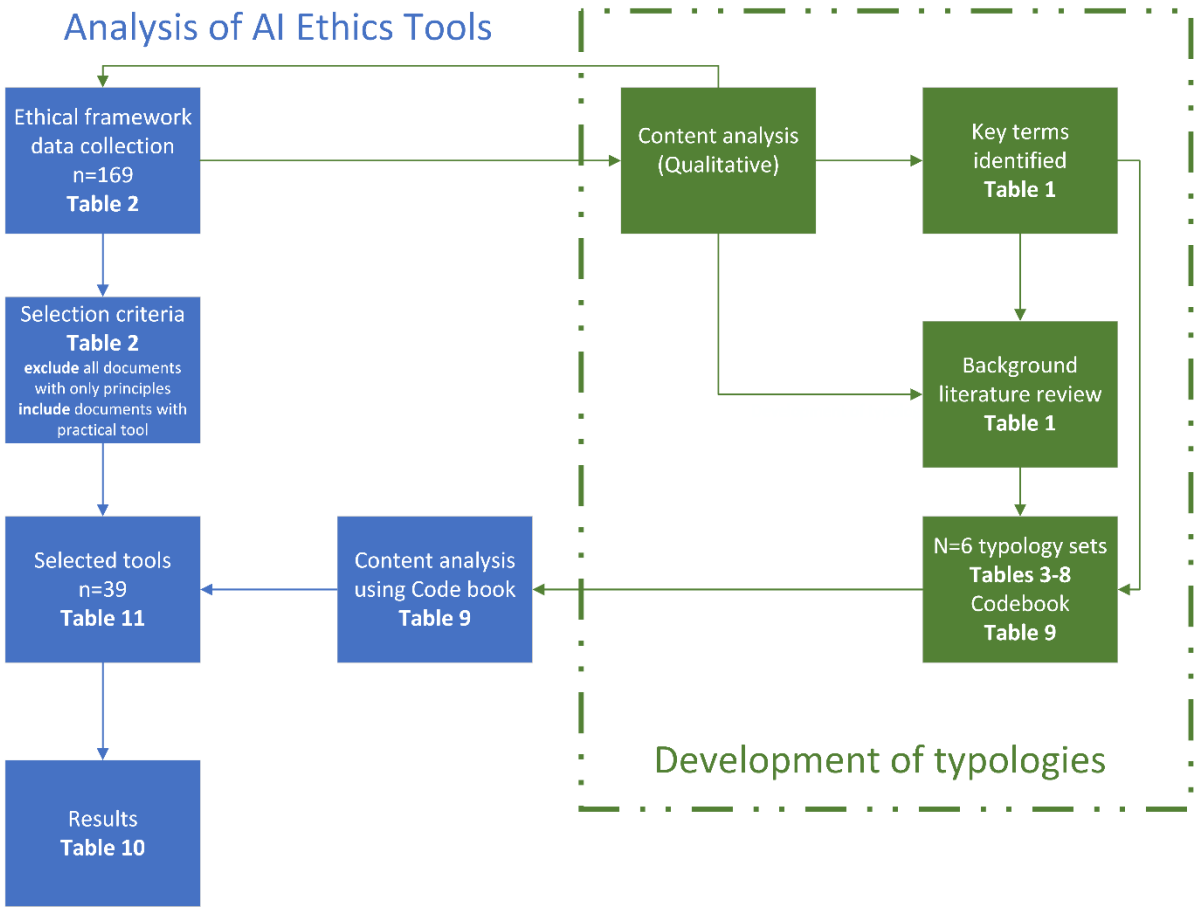


Figure 4 Flow diagram of methodology for document analysis

The research process is set out in Figure 4. The process began with a systematic collection of AI ethics documents using the document types and keywords detailed in Table 2. This comprised a combination of web searches, citation scanning and monitoring of relevant social media and news items to identify suitable candidates between May 2019 and December 2020. Other collections of AI ethics documents were also used both as sources of relevant documents, and for validation (Singh *et al.*, 2018; Hagendorff, 2019; Jobin, Ienca and Vayena, 2019; AlgorithmWatch, 2020). The initial search yielded n=169 documents. Many of these documents are drafted by public, private or not-for-profit organisations and constitute 'grey literature' not typically found in academic databases (Schopfel, 2010). Academic sources were also included, particularly as the private sector is active in producing and publishing academic papers on this topic (Birhane *et al.*, 2021).

This initial data set was analysed using a qualitative content analysis methodology (Bengtsson, 2016) to elicit frequently applied terminology and approaches. The documents were stored in Zotero reference manager and coded in an MS Excel spreadsheet iteratively to identify recurrent key words and concepts that were used to describe their main purpose, type of document, author, and audience. The key terms derived from this process are shown in Table 1.

From this a set of sub-questions were devised to query the data which shaped the categories and codes which were developed (see Table 9). These questions were considered the salient features that would allow detailed comparison of the set of AI ethics tools. Key terms were then used to search for literature that mirrored these terms across different domains as shown in Table 1. The deep background literature review of previous practices was used to identify categories which became the codebook (see Table 9). This was a reflective process where I identified principles and categories across domains and used the salient features to create typological sets as follows:

Table 3 Typology of stakeholders

Table 4 Typology of impact assessment methods

Table 5 Typology of audit methods

Table 6 Typology of internal vs external process

Table 7 Typology of technical and design tools

Table 8 Typology of when tool used and if applied to data and/or model.

Table 1 Key terms and background literature

Key terms from initial content analysis of ethics frameworks	Background literature review to build content analysis
Impact assessment	Technology Assessment
Audit	Environmental Impact Assessment
Technical tool	Social and Human Rights Impact Assessment
Design tool	Privacy and Data Protection Impact Assessment
Application stage	Risk Assessment
Stakeholder	Audit
Risk assessment	Technical and Design Tools
Procurement	Stakeholder theory
Type of author	

The next step was to narrow down the initial large data set of n=169 documents, which contained many documents that were statements of principles or discussions of AI ethics. The focus of interest was only in those documents that would give an organisation or practitioner a concrete tool to apply to AI production or deployment. See (Whittlestone *et al.*, 2019) for a discussion of why principles are not enough on their own, and how we need to bridge to gap between principles and practice. Documents were excluded that did not contain practical tools to apply ethical principles (see Table 2), leaving a data set n=39 documents that offered practical tools to operationalise ethical principles in the production and deployment of AI systems.

Table 2 Criteria for sample identification

CRITERIA	INCLUSION	EXCLUSION
DOCUMENT TYPE	Codes, principles, checklists, risk assessments, reports, white papers, academic research, technical tools, documentation, impact assessments, audits, guidelines, standards, registers, contracts, policy documents, recommendations, webpages, institutional reports, declarations, professional ethics	Opinion articles, speeches, audio/visual materials, images, legislation
KEYWORDS	AI, artificial intelligence data - ethics, stewardship, big data machine learning, deep learning algorithms predictive analytics automated decision making advanced analytics automated scoring, profiling, aggregating, sorting data science digital technology	Traditional data protection, privacy
TYPE OF CONTENT	Practical proposals for implementing ethics for AI, including both model and data	Ethical principles and frameworks without proposals for how to apply these principles
AUTHOR	Public, private and not-for profit sector (including NGO's), academic research, standards bodies	Authors not representing an organization, or not peer-reviewed publication
LANGUAGE	English	
AVAILABILITY	Public, online	
DATA COLLECTION TIME PERIOD	May 2019 to December 2020	
DOCUMENT PUBLICATION DATE	2016-2020	Pre-2016 and post 2020

3.3.1 Typology of stakeholder types

After review of stakeholder theory (Donaldson and Preston, 1995; Clarke, 2005; Freeman, 2010; Plessis, Hargovan and Harris, 2018), a categorisation of key stakeholder groups relevant across both public and private sector was developed, adapting a typology from (Foden, 2019; High Level Expert Group on AI, 2019; Stanley, 2020; National Crime Agency, 2021). Table 3 presents a typology of stakeholders that has been adapted and extended from the identification of possible stakeholders

described in (High Level Expert Group on AI, 2019, p. 25), where it is interesting to note the table did not include users or customers, or shareholders. The categories have therefore been extended to mirror the roles in the public sector, and also widened the stakeholders beyond the confines of the production or deployment of the technologies to include all stakeholders who are affected or have in interest in the process.

Table 3 Typology of stakeholders

STAKEHOLDER	PUBLIC SECTOR	PRIVATE SECTOR	
VOICELESS	**Environment** **Marginalised or excluded groups**	**Environment** **Marginalised or excluded groups**	Impacts on physical environment, ecosystems and its members, energy and raw material extraction and use. Workers in extractive or digital industries (e.g. mining, content moderation, data annotation). Traditionally marginalised groups with limited voice in society (e.g. the poor, minority ethnic groups, refugees and immigrants, disabled, incarcerated, women, children).
VESTED INTEREST	**Citizen**	**Shareholders** **Investors**	The electorate have a right to transparent processes and should have the ability to contribute to decision-making (participation). Shareholders and investors also have fiduciary duty to consider the ethical behaviour of their investment vehicles.
DECISION MAKERS	**Elected Official** **Chief Executive** **Director**	**Senior Management (C-suite)** **Board**	Senior management discusses and evaluates the AI systems' development, deployment or procurement and serves as an escalation board for evaluating all AI innovations and uses, when critical concerns are detected. It involves those impacted by the possible introduction of AI systems and their representatives throughout the process via information, consultation, and participation procedures.
LEGAL	**Compliance/Privacy** **Legal Department** **Policy**	**Compliance/Privacy** **Legal department** **Corporate responsibility department**	The responsibility department monitors the use of an ethical assessment and its necessary evolution to meet the technological or regulatory changes. It updates the standards or internal policies on AI systems and ensures that the use of such systems complies with the current legal, regulatory and policy frameworks and to the values of the organisation.
DELIVERY	**Delivery Managers** **Service Managers** **Domain Experts**	**Product Managers** **Service Development or equivalent**	The Product and Service Development department uses an ethical assessment to evaluate AI-based products and services and logs all the results. These results are discussed at management level, which ultimately approves the new or

			revised AI-based applications.
QUALITY ASSURANCE	**Policy** **Service delivery staff** **Quality assurance**	**Quality Assurance**	The Quality Assurance department (or equivalent) ensures and checks the results of an ethical assessment and takes action to escalate an issue higher up if the result is not satisfactory or if unforeseen results are detected.
HR	**HR**	**HR**	The HR department ensures the right mix of competences and diversity of profiles for developers of AI systems. It ensures that the appropriate level of training is delivered inside the organisation.
PROCUREMENT	**Procurement**	**Procurement**	The procurement department ensures that the process to procure AI-based products or services includes an assessment of ethics.
DEVELOPER	**Data Scientists/Engineers** **Developers** **Project Managers**	**Developers** **Project managers**	Developers and project managers include an ethical assessment in their daily work and document the results and outcomes of the assessment.
USERS	**Service users**	**Users** **Customers**	Participation of users in development, and/or publication of assessments for public interrogation. (NB: this layer is missing from the EU categories)
OVERSIGHT	**Independent Oversight Bodies** **Expert Committees** **Freedom of Information Requests** **Regulators** **Courts**	**Independent Review/Oversight Bodies** **Expert Committees** **Regulators** **Courts**	Public Sector governance has a variety of structures aimed at accountability and transparency and compliance with the law,

Table adapted from (Foden, 2019; High Level Expert Group on AI, 2019, p. 25; Stanley, 2020; National Crime Agency, 2021).

3.3.2 Typology of tool types for Impact Assessment

Table 4 shows the key features of impact assessments derived from the literature review.

Table 4 Typology of impact assessment methods

IMPACT ASSESSMENT	
CHECKLIST; QUESTIONNAIRE	Widely deployed tool across impact assessments and audits to describe activity and interrogate aspects of project or process. Can be used for both potential projects and to documentation for audit.
BASELINE STUDY	Commonly used in EIA and policy assessments to ascertain baseline conditions against which proposed projects or policy can be measured.
PARTICIPATION PROCESS	Mandated part of EIA process, public stages of EIA involve scoping and review, and publicly available documentation.
COST-BENEFIT ANALYSIS	Assessment tool to compare economic costs with potential benefits.
RISK ASSESSMENT	Can be qualitative or quantitative, frequently translated to a scoring or traffic light output.
LIFE-CYCLE ASSESSMENT	Assessment technique for products or materials to calculate environmental or health impacts.
CHANGE MEASUREMENT	Commonly used in policy or human rights impact assessment to determine impacts.
EXPERT COMMITTEE	Used in assessment process to provide expert evidence or domain knowledge.
GOVERNANCE PROCESS	Business and administrative processes to document activity and provide verifiable documentation.
PROCUREMENT PROCESS	Structured process to assess the impact of a purchasing decision.

3.3.3 Typology of tool types for audits

Table 5 shows key processes mapped from the review of audit techniques.

Table 5 Typology of audit methods

AUDIT	
CHECKLIST; QUESTIONNAIRE	Widely deployed tool across impact assessments and audits to describe activity and interrogate aspects of project or process. Can be used for both potential projects and to documentation for audit.
DOCUMENTATION	Audits require artifacts for inspection and assessment such records of processes, materials, outcomes and decisions.
REPORTING	Output from audits is commonly in the form of auditors' reports.
GOVERNANCE PROCESS	Business and administrative processes to document activity and provide verifiable documentation.

3.3.4 Internal vs external process

Table 6 shows the codes created to identify if the tool was designed for internal organisational use or provided for third party inspection.

Table 6 Typology of internal vs external process

INTERNAL VS EXTERNAL ASSESSMENT/AUDIT	
INTERNAL/SELF-ASSESSMENT	Designed to be used only as internal organisational tool. Outcomes assessed only by internal parties. No process for wider transparency or participation.
EXTERNAL/3RD PARTY	Designed to be used by external auditors, standards body. May include provision for publication of results/outcomes for wider transparency.

3.3.5 Technical and design tools

A sub-set of tools being suggested for operationalising ethical AI comprise design and engineering tools for use in specific stages of the production pipeline (see Table 7.) These are either materials for use in design teams in workshop style events (Institute for the Future and Omidyar Network, 2018; Doteveryone, 2019), tools for producing documentation of the design, build and test process (Mitchell *et al.*, 2019; TensorFlow, 2020), or technical tools for testing models, protecting privacy and security, testing for bias (Bantilan, 2017; Badr, 2019; Kaissis *et al.*, 2020), or tracking provenance of data (Chapman *et al.*, 2020).

Table 7 Typology of technical and design tools

TECHNICAL AND DESIGN TOOLS	
WORKSHOP MATERIALS	Materials produced for use by design teams as workshop or discursive events e.g. scenarios, design cards, agile design events.
DOCUMENTATION	Technical documentation like logs and incident reports, technical descriptions.
TECHNICAL TOOLS	Specific technical applications for addressing issues like privacy, security, bias, transparency, provenance in models and data.

3.3.6 Production and deployment process for AI Systems

AI systems go through stages of production, from initial definition of a use case, development of a business case, through the design, build, test and deploy process (ICO, 2020). Assessment and audit tools can be applied at different stages of the process (or attempt to capture cover the whole pipeline) and can be focused on the data flowing through the pipeline, or the attributes of the

model, or both. Table 8 defines codes for these stages. The pipeline for deployment of systems often includes selling the AI system to a customer, who will deploy the system, at which point ethical considerations can be included in the procurement process.

Table 8 Typology of when tool used and if applied to data and/or model

STAGE IN PROCESS TOOL USED	
BUSINESS/USE CASE	A problem space, or area for improvement is identified, and the use case and business case are developed.
DESIGN	Business case is translated into design requirements for engineers.
TRAINING DATA COLLECTION	Training and test data is identified, collated, cleaned, and prepared for training the model.
BUILDING	AI application is built.
TESTING	The system is tested.
DEPLOYMENT	The system goes live.
MONITORING	System performance is monitored as it performs in the wild.
PROCUREMENT OF SYSTEM	Third party buys system for their own use.
DATA	Depending on the focus of the tool, either the data pipeline is the main object of assessment, or the model itself.
MODEL	

3.3.7 Document collection process

A total of n=169 items were identified under the broad category of AI-related ethics frameworks, which after application of the exclusion criteria resulted in a final list of n=39 ethics tools (see Appendix 1). The documents were analysed using qualitative content analysis (Mayring, 2019), through the development of a codebook of variables to identify key features (see Table 9). This was an iterative process where the codes were refined during the process of reading and coding the material.

The terms impact assessment and audit are used in differing ways in the domain of AI ethics tools made coding of these documents complex. As Carrier and Brown (2021) note, there is much ambiguity over the use of the term 'audit' in relation to AI ethical assessment being used by what they term as the 'AI ethics industry'. Across the landscape of AI ethics audit and impact assessment tools, terms are often used loosely, or are used interchangeably. An Ada Lovelace Institute report, 'Examining the black box'**,** categorised algorithmic *audit* into two types, a narrow 'bias audit' or a broader 'regulatory inspection' which addresses 'compliance with regulation or norms, necessitating a number of different tools and methods; typically performed by regulators or auditing professionals' (Ada Lovelace Institute and DataKindUK, 2020, p. 3). Algorithmic *impact assessment* is divided into an ex ante risk assessment, and what the report terms an 'algorithmic

impact evaluation' which assesses the effects of an application after use (Ada Lovelace Institute and DataKindUK, 2020, p. 4). The codes reflect a decision by the researchers to define 'impact assessment' as an ex ante process which was predicting possible impacts, with audit being an ex post process for examining ongoing activities. This is not necessarily reflected in the language of the documents themselves, depending on the author and field or discipline from which they originated.

Table 9 Sub-questions and derived codes for analysing documents

QUESTION POSED TO DOCUMENTS	CODES
WHICH SECTOR WERE THE AUTHORS/USERS FROM?	Public Sector Private Sector Not-for-Profit Academic Research
WHICH STAKEHOLDER WOULD EITHER USE THE TOOL, OR ENGAGE WITH THE RESULTS? [SEE TABLE 3 FOR DETAILED CATEGORY BREAKDOWN]	Voiceless Vested Interest Decision Makers Legal Delivery Quality Assurance Procurement HR Developer Users Oversight
WHAT TYPE OF TOOL WAS IT? WHICH STRATEGIES DID IT EMPLOY?	**Impact Assessment** Checklist questionnaire Baseline study Participation process Cost-benefit analysis Risk assessment Life-cycle assessment Change measurement Expert committee Business process Procurement process **Audit** Checklist questionnaire Documentation Reporting Business process **Technical Tools** Workshop materials Documentation Technical tests
WERE THESE TOOLS FOR USE INTERNALLY, OR HAVE EXTERNAL ELEMENTS?	Internal/self-assessment External/3rd party
WHICH STAGE IN AI PRODUCTION AND USE WAS THE TOOL USED?	Business/use case Design Training data collection Building

	Testing Deployment Monitoring Procurement of system
WAS THE TOOL APPROPRIATE FOR ADDRESSING THE MODEL, DATA, OR BOTH?	Model Data

3.3.8 Limitations

The coding process consisted of reading and re-reading the documents and coding them against the typologies to create the results (see Table 13.) The research methodology used is reflexive and adaptive (Bengtsson, 2016), creating a robust process for relating the document data to its context as shown in the diagram in Figure 2. Despite this, a single researcher analysing and coding the documents presents a limitation in that often validity of qualitative analysis is considered to be justified by the process of recurrent iterations with different coders (Patton, 2014). Despite this limitation, I believe every effort has been made, from the conception and planning of the project, through to development of typologies and coding of results, to consider where bias and omission could occur in the process (Krippendorff, 2013). The categories enable assessment of AI ethics tools that reliably surface salient features which can be used to compare across disparate types of tool or procedure.

Some limitation of the analysis derives from the categories chosen in the typologies, in particular the somewhat blunt categorisation of tools as focusing data/model/both. Although these categories have utility in determining how the tool could be applied, this part of the typology could be nuanced if it included categories that captured broader applications of tools. This part of the typology could have benefited from expanded categories to denote aspects such as use case, context and downstream risk. This would have enabled a richer description of the tools that address broader ethical considerations and contexts beyond a focus on the model and/or data.

This research does not set out to provide an exhaustive review of the computational techniques in the AI/ML research to address ethical issues like fairness and explainability, for this see for example (Lee and Singh, 2021; Mehrabi *et al.*, 2021).

3.4 Industry interviews

The next section of the research moves on from analysing the suggested frameworks for addressing ethical issues in AI to canvassing responses from actors in industry to the 'ethical turn' (Raab, 2020) in digital technology production. This was to understand how the types of tools analysed in Chapter 5 are being deployed, if at all.

3.4.1 Target group

The target group for the interviews was defined as:

1. A founder/CEO – high level decision-maker with responsibility for overall company
2. A technology company involved in development and commercialisation of emerging data - driven technologies
3. Start-ups (micro-enterprises of less than 10 employees) and small enterprises (with less than 50 employees) (European Commission, 2016)

3.4.2 Summary table of interview participants

Table 10 Summary table of interview participants job role, company size and sector

RESPONDENT #	JOB ROLE	COUNTRY REGISTERED	MARKET	AGE OF CO.	WORKFORCE	PRODUCT/ SERVICE AREA	TARGET CUSTOMER/ SECTOR
#1	President Europe	US	B2B	<10	~40	Biometric authentication	Financial services Banking Aid agencies
#2	Head of Data Science	UK	B2G B2B	<20	~60	Research Public policy technology	Government NGO Private sector
	Founder/ CTO	UK	B2G	<5	~2	Data science	Government NGO
#3	CEO	US	B2B	<15	~50	Biometric authentication	Financial services Banking
#4	Director	UK	B2B B2G B2C	<10	~4	Community energy	Citizen Commercial Social housing Local government
	Information & Governance Lead	UK	B2G	<5	NA	Smart City projects	Local business Citizen
	CIO	UK	B2B B2C	NA	NA	Telecoms	Telecoms
#5	Founder	UK	B2B B2G	<10	~6	GIS Mapping Energy infrastructure	Energy Transport Local & Regional government

RESPONDENT #	JOB ROLE	COUNTRY REGISTERED	MARKET	AGE OF CO.	WORKFORCE	PRODUCT/ SERVICE AREA	TARGET CUSTOMER/ SECTOR
#6	Founder	UK	B2B B2G	<15	~6	Consent/ privacy Smart city/ energy	Public Local government
#7	Founder	UK	B2B B2G	<5	~2	Patient management	Healthcare
#8	Founder	US	B2C	<5	~2	Music	Advertising Film/TV
#9	Founder CTO	UK	B2B	<5	~2	Retail platform application	Online retailers
#10	Founder CCO	UK Swiss/ German	B2B B2B B2C B2G	<10 <10	1 ~10	Product commercialisation Compliance	Privacy Security Privacy Security

3.4.3 Rationale for start-ups/small enterprise as participants

Large organisations generally find it easier to implement governance and compliance processes as they possess the necessary human and financial resource. Findings from the implementation of GDPR shows that it is the smaller organisations who find it the most challenging to implement (Ayling, 2017; Sirur, Nurse and Webb, 2018). Start-ups generally lack skills in for example, data protection, and lack the finance to hire external consultants, unless they are working in a highly-security conscious market (Sirur, Nurse and Webb, 2018; Norval *et al.*, 2021). It might be expected that if small companies struggle with data protection requirements which are matters of legal compliance, then they might also be similarly challenged in implementing wider (non-regulated) ethical governance processes.

Tech start-ups are also key drivers of innovation in technology development and application and as such 'influence how new technologies come to be designed, deployed, perceived, and used – and can shape standard industry practices in the process' (Norval *et al.*, 2021, p. 279). Despite operating at the 'cutting edge' of technology, these small companies often have financial and human resource constraints with limited expertise beyond those needed for production. They are driven by a desire

to disrupt and the need to establish their place in the market. These factors make this group of particular interest in examining their relationship to ethical impacts and governance.

3.4.4 Rationale for participant role

It has long been recognised that executives are key actors in companies (Penrose, 1995). In order to understand organisations much of the information needs to come from those who lead them. It is they who understand how decisions are made and are responsible for making and enacting policy. In start-ups there may only be one level of staff hierarchy (the founders) anyway. 'For researchers seeking process information, the CEO or other members of the organization's upper echelons are not just the best but might be the only sources for some variables. However, the willingness or ability of these executives to share such evidence with researchers is another matter.' (Cycyota and Harrison, 2006, p. 133) Despite the difficulty of recruiting 'elite' respondents (Richards, 1996), this group provide the best insight into how ethical challenges might be met (or not) within their organisations as they are responsible for driving governance and business processes.

3.4.5 Interview questions

The interview questions were informed by an in-depth review of literature, and particularly a close reading of a collection of AI ethics framework documents (see Chapter 5). From this collection of frameworks, which included many documents explaining and describing principles, a sub-set of documents was extracted that presented practical tools for application by an organisation or developer. The question still remained after analysis of these tools – how are/could these tools be applied in the real world?

3.4.6 Designing the interviews

The interview questions were kept open-ended and intended to be flexible for a range of respondents, while covering the key questions raised by the previous analysis of tools and case studies.

See Appendix B for the interview script. The respondent was invited to describe their job role, company and product or service. This description was then used by the interviewer to make relevant follow up questions around ethical issues related to their own domain. Respondents then gave their reactions to 4 example tools (see Appendix C).

Respondents were then asked to comment on:

1. The suitability/useability of the tool in general
2. The suitability/useability of the tool for their own companies/projects

3. Any tools or processes they use (similarity/difference)
4. What costs these tools might make (financial/resource/time/staff)

Follow up questions were then asked about impacts on innovation and their view of existing/potential regulation in this space.

3.5 Example ethics tools

3.5.1 Example A – IBM AI ethics statement

The first example document[2] (see Appendix C) chosen to show respondents was a web page from the IBM website, setting out three high-level principles. These form part of IBM's ESG reporting, under 'Principles for Trust and Transparency' (IBM, 2018). This item was chosen purposely to represent a set of ethical principles, in contrast Examples B, C and D which were applied tools. This was to elicit participants views on statements of ethical principles and how useful they might be, and how these were viewed in comparison to an applied ethical tool.

Respondents were asked what they thought of the principles themselves, what their purpose might be, how effective they might be, and if they used anything similar themselves.

3.5.2 Example B – EU Trustworthy AI – Statement of Principles and checklist

The second example[3] discussed with respondents was the checklist at the end of the 2019 report from the High Level Expert Group on Trustworthy AI. This document has been influential in the AI ethics debate (Brundage *et al.*, 2020; van Wynsberghe, 2021), and informed the latest iteration of EU responses to AI and digital technologies in their draft AI regulation proposals (European Commission, 2021). At the time of preparation, this was a pilot tool, later re-published as the Assessment List (ALTAI) for self-assessment tool following feedback from the pilot (European Commission HLEG AI, 2020). The original pilot questionnaire was used for the interviews as it was reasonably concise compared to the later version to allow respondents to give a view of it within the interview time frame and covered the same topics.

[2] IBM's Principles for Trust and Transparency 2018 https://www.ibm.com/blogs/policy/wp-content/uploads/2018/06/IBM_Principles_SHORT.V4.3.pdf

[3] High Level Expert Group on Artificial Intelligence (2019) Ethics Guidelines for Trustworthy AI, pp 25-31 https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

This was selected as the next step on from high level principles, as a document to be used internally, and including comprehensive coverage of issues. The format is presented as a set of questions to be answered under the areas listed in Table 11:

Table 11 Summary of EU Trustworthy AI checklist

Summary of ethical issues identified for 'Trustworthy AI' EU Commission Futurium (High Level Expert Group on AI, 2019)		
Ethical Issues	**Ethical principles**	
Human agency and oversight	Fundamental rights	Respect for human dignity Freedom of the individual Respect for the rule of law, justice and democracy Equality, non-discrimination and solidarity Citizens' rights
	Human agency	Able to enact informed decision-making and protection from forms of unfair manipulation, deception, herding and conditioning
	Human oversight	Ensure that public enforcers have the ability to exercise oversight in line with their mandate Include in design governance mechanisms to monitor and control systems
Technical robustness and safety	Resilience to attack/security	Protection against vulnerabilities that can allow exploitation by adversaries
	Fall back plan, general safety	Safeguards that enable a fall back plan in case of problems
	Accuracy	Ability to make correct predictions, recommendations, or decisions based on data or models
	Reliability	System that works properly with a range of inputs and in a range of situations
	Reproducibility	Results can be replicated under the same conditions
	Respect for privacy	Systems must guarantee privacy and data protection throughout a system's entire lifecycle

Privacy and data governance	Quality and integrity of data	Data may contain socially constructed biases, inaccuracies, errors and mistakes Malicious data may change the behaviour of the system
	Access to data	Data protocols governing data access
Transparency	Traceability	Documentation to record data processes and system production
	Explainability	Explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system)
	Communication	The system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand
Diversity, non-discrimination and fairness	Avoidance of unfair bias	Oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner Data sets used by systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models
	Accessibility and universal design	Inclusion of a wide range of users needs in design Adherence to accessibility standards
	Stakeholder participation	Consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle
Societal and environmental wellbeing	Sustainability environmental friendliness	Consideration of the environmental friendliness of systems' entire supply chain
	Social impact	Deployment of systems that negatively impact social relations, mental or physical well-being
	Society and democracy	Take into account its effect on institutions, democracy and society at large

Accountability	Auditability	Evaluation by internal and external auditors In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited
	Minimising and reporting of negative impact	Conducting risk assessments Training and awareness Whistle-blower and reporting mechanisms
	Trade-offs and redress	Relevant interests and values implicated by the system should be identified and if conflict arises trade-offs should be explicitly acknowledged and evaluated

3.6 Example C - The Information Accountability Foundation (IAF) 'Ethical Data Impact Assessments and Oversight Models'

Data protection practice, like GDPR, has developed risk-based Data Protection Impact Assessments (DPIAs) (CNIL FR, 2018; ICO UK, 2018). There is now a move to extend these to Ethical Data Impact Assessments (EDIAs), that incorporate privacy concerns, and the wider technical and ethical concerns of data-intensive technologies to produce an 'ethical by design' tool and process for use by stakeholders.

The model EDIA[4] was developed from the output of a project by IAF commissioned by the Office of the Hong Kong Privacy Commissioner for Personal Data. The goals as stated by IAF were:

> 'to encourage ICT innovation and competition by demonstrating that an organization has considered the interests of all parties before deciding to pursue an advanced data-processing activity' (IAF, 2019, p. 3), and
>
> 'to help translate sound, implementable business processes into a framework that would enable the economic benefits of technology driven data innovation within a foundation of digital accountability. In other words, ethical data stewardship' (Cullen, 2019).

[4] The Information Accountability Foundation (2019), Ethical Data Impact Assessments and Oversight Models https://secureservercdn.net/192.169.221.188/b1f.827.myftpupload.com/wp-content/uploads/2020/04/Model-Ethical-Data-Impact-Assessment-January-2019-002.pdf

The model EDIA proposed by IAF is an attempt to incorporate data stewardship (see their earlier work on information accountability, big data and stewardship (Cullen, 2015, 2018)) with technology, design and engineering practices (Shannon, Green and Raicu, 2018), that would enable organisations and public authorities to effectively manage the risks and opportunities of the deployment of data-driven technology. IAF considers it part of the 'fourth wave' of privacy regulation (Abrams, no date) which addresses issues arising from a 'cybernetic' industrial revolution that move beyond traditional concerns of privacy protection and regulation.

3.6.1 Composition of IAF Ethical Data Impact Assessment (EDIA)

The impact assessment is recommended to be used when planning 'advanced analytics such as: evaluation or scoring (including profiling and predicting), automated individual decision-making, systemic observation or monitoring, data processed on a large scale, matching or combing data sets, innovative use or applying new technological or organizational solutions (such as AI and ML)' (IAF, 2019). This framework does not focus on the techniques (e.g. ML models) but on the data – data stewardship. This approach avoids labelling everything as 'AI' and assuming that ethics frameworks and tools need to be targeted not primarily on the AI techniques themselves, but on the data that flows through these systems. Statistical analysis, for example, is not necessarily 'AI', but can still derive powerful insights, the application of which can raise ethical dilemmas.

The IAF model is intended as an extension of (not a replacement for) a Privacy Impact Assessment (the IAF point to the French data protection regulator CNIL for a suitable template for conducting a DPIA (CNIL FR, 2018)). An EDIA is to be used where impacts on individuals, groups and wider society encompass a wider set of risks and benefits beyond those considerations that focus on personally identifiable data. The model uses qualitative questions, with the addition of a risk modelling process (significance, likelihood and effectiveness of controls). Risk scores are aggregated into an overall risk/benefit heat map. This model follows well-established Enterprise Risk Management (ERM) approaches that are drawn for the wider principles of risk assessment that developed out of environmental concerns of the 1950s and 60s (Aven, 2016).

The IAF model (2019) uses sets of questions (see Appendix C) to yield descriptions of data, processing and impacts

3.7 Example D – Consequence Scanning toolkit – Doteveryone

The final example[5] was chosen as representative of tools that provide developer teams with materials for considering ethical problems. The materials are designed for a workshop, as part of an agile production process (Agile Manifesto, 2001). The tool comes from (a now defunct) responsible technology think tank in the UK. Doteveryone describes the tool as 'a way for organisations to consider the potential consequences of their product or service on people, communities and the planet. This practice is an innovation tool that also provides an opportunity to mitigate or address potential harms or disasters before they happen' (Doteveryone, 2019).

Consequence Scanning was designed to be lightweight and adaptable, and fit in with established agile development processes. It is suggested for use at initial conception of a product, during roadmap planning, and during feature creation. Three questions form the core of tool:

1. What are the intended and unintended consequences of this product or feature?
2. What are the positive consequences we want to focus on?
3. What are the consequences we want to mitigate? (Doteveryone, 2019)

It is then suggested that the answers be logged, and responsibility assigned for actions. This is the main output from the event, which can be plugged into other existing processes.

This example was included to give participants an opportunity to comment on a developer level tool, rather than the organisational level tools in examples A, B and C.

3.8 Interview process

10 semi-structured interviews (Silverman, 2016) were conducted with CEO's/founders of AI/data driven technology companies (University of Southampton Ergo II Application 55529). The participants were recruited from business contacts of the researcher (n=5), with additional participants recruited using the original participants and the researcher's academic network for introductions (n=5) using snowball sampling (Biernacki and Waldorf, 1981).

Senior staff such as these are difficult to access and can be reluctant to take part in such research due to pressure of time, or potentially not being comfortable talking about their companies and products to a researcher. Personal recommendation is generally needed to engage such participants, who comprise a limited demographic pool. Drawing on a small, hard to recruit set of

[5] Doteveryone (2019) Consequence Scanning – an agile practice for responsible innovators https://doteveryone.org.uk/project/consequence-scanning/

participants meant that a semi-structured interview was selected as the best way to elicit as much breadth and depth of responses as possible from a constrained set of respondents (Richards, 1996).

The interviews were conducted over an extended period (from early March 2020 to July 2021) due to interruption from the pandemic. The first interview was recorded in person before the first national lockdown, the rest of the interviews have been conducted via MS Teams. The interviews were scheduled to take 45-60 minutes, with some participants extending the duration of the interview by up to 30 minutes as they were engaged in the conversation. Recordings were made of all the interviews, which were then transcribed. Quotes from respondents have been edited for clarity, and all references to individuals and companies have been redacted to protect privacy and professional/industry reputation.

3.9 Limitations of interview methodology

Limitation 1 – Potential for selection bias – respondents who may not wish to reveal the ethical challenges in their products would not agree to an interview, therefore the respondents are a group who believe their companies behave in an ethical way ('bad actors' self-exclude).

Limitation 2 – Potential for response bias – respondents will give answers that they think the researcher wants to hear/will put them in the best light or may conceal information that might damage their own or their company's reputation or need to protect IP or are under NDAs'.

Limitation 3 – A small number of interviews using existing networks and snowball sampling which cannot purport to represent the views or behaviours of tech companies as a whole – this limits the generalisability of the results. Therefore, the interview results are presented as modest conclusions and as an indication that further research is needed.

Limitation 4 – Most respondents did not find the time to view the examples before the interview (they were supplied with pdf's on accepting an invitation). Therefore, the time spent examining Example A, B, C and D by respondents was limited. The documents were shared with the respondent during the interview and reviewed and discussed in real time. On reflection, this resulted in participants giving opinions based more on their pre-exiting knowledge and ideas than more robustly engaging with the example tools and responding to them in a more detailed way.

Overcoming the time limitations that respondents have for engaging with materials prior to interviews was a difficult dilemma given the nature of the participants. Perhaps it might have been useful to provide them with a short summary of each tool to read through beforehand on one short document, with the option to refer to the full example if they felt motivated. At least then they would have an overview of the tools to discuss and have some background to the examples. Ideally

it would have also been good to engage participants in a workshop or focus group activity where they would have been tasked to engage more closely with the example tools and discuss them with peers. The results would then have shown a closer relationship to the example tools.

Limitation 5 - Not possible to do inter-rater reliability checks for coding interview data in a PhD project with one researcher. Possibility of skewed interpretation of results.

Limitation 6 – Diversity of participants. The demographic of the individuals recruited to the interviews lacked diversity – all were male, 9 of which were white British or Irish, and 1 South Asian. While it would have been desirable to have wider demographic represented, the tech industry itself has diversity challenges[6] which are reflected in the participant group. It is acknowledged that the research has limitations stemming from the limited demographic, and the recruitment process for participants should have included a more determined strategy to recruit participants from a wider demographic that represented a more diverse range of lived experience. This could have been done by, for example, contacting organisations representing specific groups in the tech industry (for example women, those with disabilities) and organisations focused on diversity and representation. While it should be noted that AI ethics teams in large companies and corporations have been conspicuously led by pioneering women[7], SMEs do not have dedicated teams, and in the cases examined, not even a dedicated staff role for AI ethics. It is acknowledged that a more diverse participant group would have resulted in a wider range of experience and opinion about ethical tools and approaches from participants with differing status and power. This would have yielded a richer set of responses.

Limitation 7 – Lack of interview data from other actors. The original proposal for this research was curtailed by the impact of the Covid 19 pandemic. Despite seven of the participants (#2, 3, 5, 6, 7, 9) being developer founders or having a developer background, it would have been insightful to gather data from developers whose roles do not include running companies or making senior management decisions. The original plan for workshops with groups of developers would have given a more rounded view of AI ethics tools from both a senior management and practitioner perspective and enabled a triangulation of evidence from documents and two different participant groups.

[6] For example, women only make up 19% of UK IT professionals, Tech Talent Charter https://www.techtalentcharter.co.uk/about-the-tech-talent-charter

[7] See for example Margaret Mitchell and Timnit Gebru who previously led the ill-fated Google Ethics Lab, Kathy Baxter at Saleforce, Alice Xiang at Sony, Rumman Chowdhury at Twitter and Maria Luciana Axente at PwC)

Chapter 4 Results - Pilot case study

The Pilot Case Study was conducted at the beginning of the research process, which used the lens of 'smart cities' to consider ethical problems in data flows in this context. The process of undertaking the case study, especially the background literature review needed to complete it, led to a later development of the research project that is reflected in the current set of research questions, which looks at ethical tools and frameworks for AI/ML more generally, rather than situating the study in the particular context of smart cities and the ethical challenges embedded in the data flows.

4.1 Use Case 1 – National Grid Affordable Warmth Solutions Project

Use Case 1 describes a project for Affordable Warmth Solutions CIC (AWC) in the UK. Established in 2008 by National Grid Plc as part of government policy obligations under the Energy Company Obligation scheme (Ofgem, 2016a). AWC is a Community Interest Company that assists qualifying homes in the 25% most deprived areas in England by offering:

- New gas connections to consumers not currently connected to the Cadent gas distribution network.
- Free or discounted gas central heating systems to qualifying households.
- Income maximisation, energy efficiency and tariff advice. (Affordable Warmth Solutions CIC, 2019)

AWC commissioned a geospatial insight company (X) who specialise in the energy sector, to deliver leads for properties which were likely to qualify for free (to the occupier) gas connection under the Fuel Poor Network Extension Scheme (Ofgem, 2016b). X delivers energy system insights for clients using geospatial Big Data and Machine Learning to identify, measure and monitor assets and infrastructure using a variety of data sources.

The AWC project was for an area across North London, to identify addresses that might qualify for connection to the gas network to reduce reliance on expensive electric heating. At the time of the project the qualifying criteria were for households located within the 25% most deprived areas, as measured by the government's Index of Multiple Deprivation (IMD) (Ofgem, 2017).

X delivered a final GIS dataset that linked address, property type and age, tenancy details (owner occupier, private rental, social housing, and council property) and distance from the gas network.

Exclusions were applied in the data for properties which were either in very close proximity to existing gas pipelines (which were assumed to be already connected), and those properties which were too far away from the network to be viably connected. AWC intended to use this data to make doorstep or postal contact with the occupier to inform them of the potential for subsidised gas connection.

AWC received a final dataset identifying properties by address that would be likely targets for an approach by its team for receiving a free-to-customer gas connection. This approach would be by post or by doorstep visits. The names of occupants of the properties were not included in the supplied dataset, and AWC made approaches to 'the occupier'. X felt confident that they had complied with all relevant regulations regarding the data as they had fulfilled the licence obligations of the datasets they deployed and did not consider that the data was in any sense personal data, as they were not sourcing, analysing or supplying names, telephone numbers or email addresses, only property addresses. See Table 12 for a full list of the data sets used.

'Other than the address, there is probably nothing in there that you would consider as personal data…' stated the CEO of X. Property addresses have been traditionally considered public, non-personal data which does not fall under the remit of data protection regulations. The marketing data bought in from Intermedia Global was sold under licence conditions that assured X that only data with consent attached for sharing and reuse was included in the dataset.

Table 12 Data sets used in Affordable Warmth Project

Data Owner	Feature/Field	Format	Source	Licence
Ordnance Survey MasterMap	OS MasterMap Topography Layer (Ordnance Survey, 2019b)	GML 2.1.2 (OGC, 2019)	OS Mapping	1yr
Ordnance Survey AddressBase	Postal address (Ordnance Survey, 2019a) Domestic	CSV	Post Office	1yr
Google Places	Address	CSV	Google maps	Single use
Geoinformation Group	Property type Property age	GIS	Arial photography - identify roof materials, delineate boundaries, then interpret if is e.g Victorian terrace, apartment block etc.	1yr
Cadent Gas	Gas network map	GIS SHP - GEMINI 2.2 standard (AGI Standards Committee, 2018)	Client mapping	Supplied by customer
Intermedia Global Marketing	Tenure type Address	CSV	traditional marketing company – collect data via customers (e.g. insurance, telephone surveys) – sold as DPA compliant - only secured tenure data for approx. 60% of property list, 40% of	1yr single use

			addresses had opted out of data being shared	
Final dataset delivered to client as a GIS dataset with pop-ups for building footprint with further info (address, tenure, building type, distance from gas network). 1yr licence incorporating licence conditions from contributing data set owners.				

X feels confident that their data pipeline is secure and well managed, a data flow diagram illustrates the stages of data processing in Figure 5. X use Dropbox as cloud storage, with a copy in OneDrive as a backup, and local copies of data are deleted as soon as they have been copied to secure cloud environments. Access controls are set for staff working on the data, with permission only being granted to access the datasets necessary for that person to complete their specific tasks in the project. X explained that this siloed pipeline is not designed particularly for privacy reasons, but to create the most efficient workflow for staff. This also has a secondary benefit of exposing data to as few people as possible. X felt that the client (AWC) was a reliable partner, and that the purpose of the data use did not present any business risks to X, or to individuals. They felt confident they had completed all necessary due diligence and that AWC would use the data appropriately and comply with the terms of the licence agreement.

As the quote below from the CEO of X illustrates, he has thought through the implications for the project he was undertaking and saw no particular ethical challenges and felt confident that the data processing he was engaged in was legal and properly managed. Interestingly, he quite clearly saw the limits of his control and responsibility over the insights from the project. After the dataset is handed over to the customer there ceases to be any responsibility or effective means of control, unless it came to his attention that a customer had broken the conditions of the data licence, which would enable legal action under contract law.

> ‘I think, with us, and I think, for most geospatial data companies, it’s what our clients do with the data, rather than what we do with the data. We process data and we provide information, we don’t do anything with it. So, umm you tell us all the solar suitable roofs in Bournemouth, OK, we’ve done it, and then, what gets done with that down the line, it’s almost, it’s not that it’s not our responsibility, when we work with a client we have some insight as to what they want the end data for, but it’s also not necessarily our responsibility to police our clients. We have to have some trust in what they’re going to do with the data. If we’ve been through all our due diligence processes, you know, getting data from GDPR compliant sources, keeping it siloed in the office, storing it in the cloud, and so forth, once we pass that to an organisation, we have to have some trust that they’re going to do the same. It doesn’t always happen. Someone like National Grid would

– you know they're going to store data in the right way, you hope they're going to store data in the right way, use it in the right way, they've got their own customer database, and all they're going to do is join our data with theirs and say, 'Alright, this customer of ours needs to be connected to gas, we can go and do for free', you know, they've probably got their phone number anyway, 'Let's ring them up and organise an appointment.'

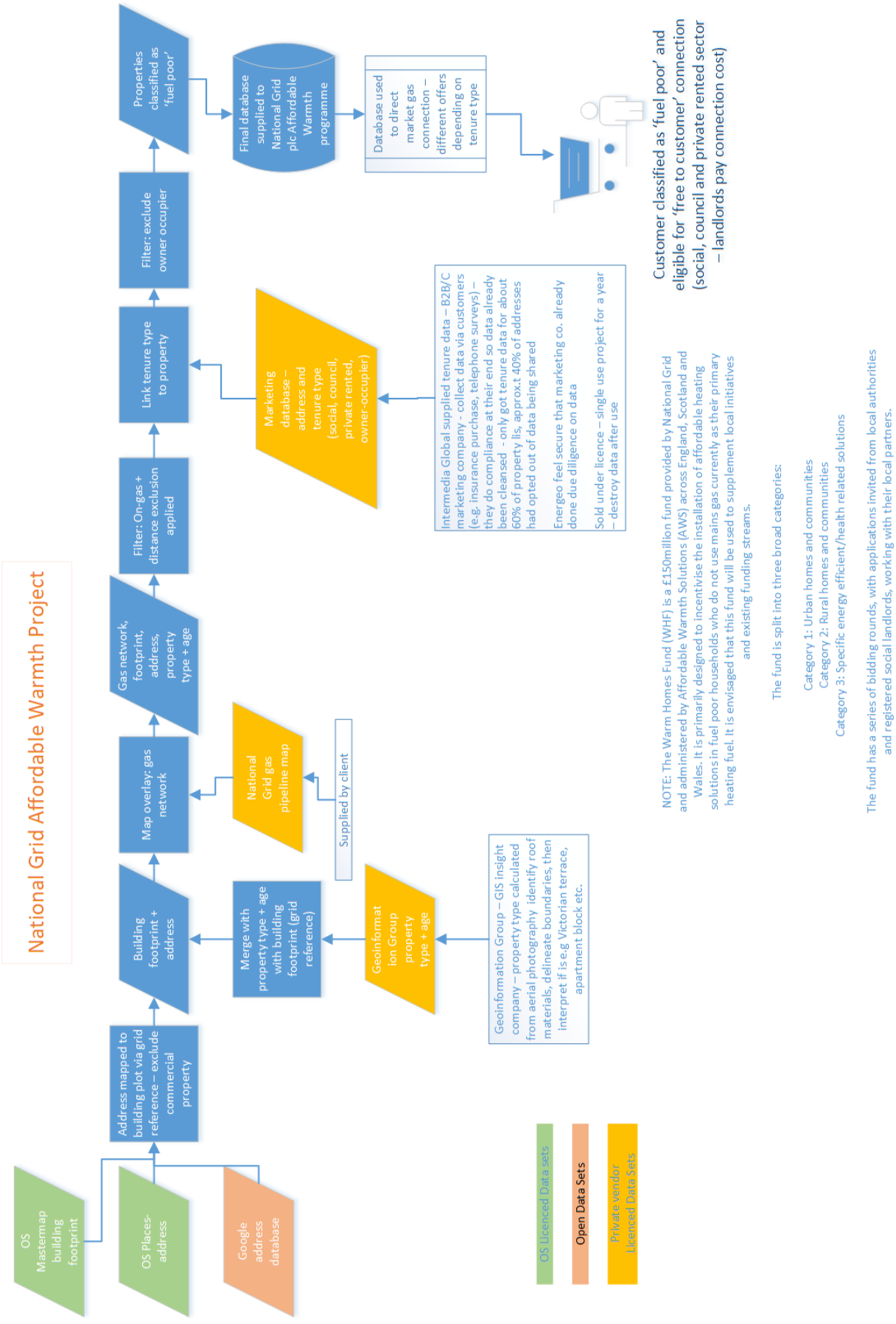


Figure 5 National Grid Affordable Warmth Project Data Flows

4.1.1 Potential for social sorting

The IAF Ethical Data Impact Assessment (IAF, 2019) has questions regarding the potential for negative impacts from social sorting e.g. in Section 3: Impact to Parties and in Particular to Individuals.

'1. During the activity, how will data be used and are there identifiable expectations of individuals, groups of individuals, and society for each use of the data?

For example, could there be an impact (real or perceived) to social or reputation status?

2. Could the data be used in a way that may result in a group of individuals being treated differently from other groups of individuals?' (IAF, 2019)

The purpose of the Affordable Warmth project does identify households by tenure type, property type and address, all of which could be used as proxies to infer characteristics like income bracket and social class.

4.1.2 Regulatory framework for Affordable Warmth

Under the ECO obligations set by Ofgem, (2016a), energy companies have targets to meet for connecting fuel poor households to the gas network. AWC commissioned X to produce a data set to identify those homes with a greater likelihood of being eligible for the scheme. AWC needed to identify those who only have access to a electricity (not connected to the gas network) and who also fall into a profile where they might also be assumed to be in, or in danger of, 'fuel poverty'. This is defined by the UK Government using the Low Income High Costs (LIHC) indicator. A household is considered to be fuel poor if:

- they have required fuel costs that are above average (the national median level)
- they were to spend that amount, they would be left with a residual income below the official poverty line

There are 3 important elements in determining whether a household is fuel poor:

- household Income
- household energy requirements
- fuel prices

(BEIS, 2018)

The GIS project supplied by X did not ascertain if households were in fuel poverty or not, but it did use proxies for likelihood of being so, or in danger of becoming so. The Affordable Warmth Scheme enables households who meet the eligibility criteria (receipt of certain benefits/overall household income) to receive a free connection to the gas network. See Figure 6 which shows the relationship between lack of access to gas and fuel poverty. There is also a separate scheme under the ECO scheme to provide subsidised gas boilers. The insights delivered by X did not ascertain tenants' ability to pay for energy bills in the GIS data set supplied, but instead used proxies for the likelihood of being eligible for the scheme (tenure type, property age/type.) As the Governments fuel poverty statistics reveal, it is older dwellings, and those in the private rented sector, who are more likely to be fuel poor.

- 'Older dwellings tend to have a higher proportion of households in fuel poverty compared to newer dwellings. Households in dwellings built between 1900-1918 were most likely to be fuel poor (18.6 per cent) with an average gap of £379. This is compared to just 4.2 per cent of fuel poor households in dwellings built post 1990 with an average fuel poverty gap of £226.

- The level of fuel poverty is highest in the private rented sector (19.4per cent) compared to those in owner occupied properties (7.7 per cent). Those in the private rented sector also tend to be deeper in fuel poverty, with an average fuel poverty gap of £383, compared to just over £200 for those in local authority and housing association properties.

- When considering household composition, those living in 'multi-person (adult) households' are deepest in fuel poverty with an average fuel poverty gap of £413 compared to a single person under 60 (£208). However, the highest prevalence of fuel poverty is seen for lone parents with dependent child(ren) (26.4 per cent).' (BEIS, 2018, p. 4)

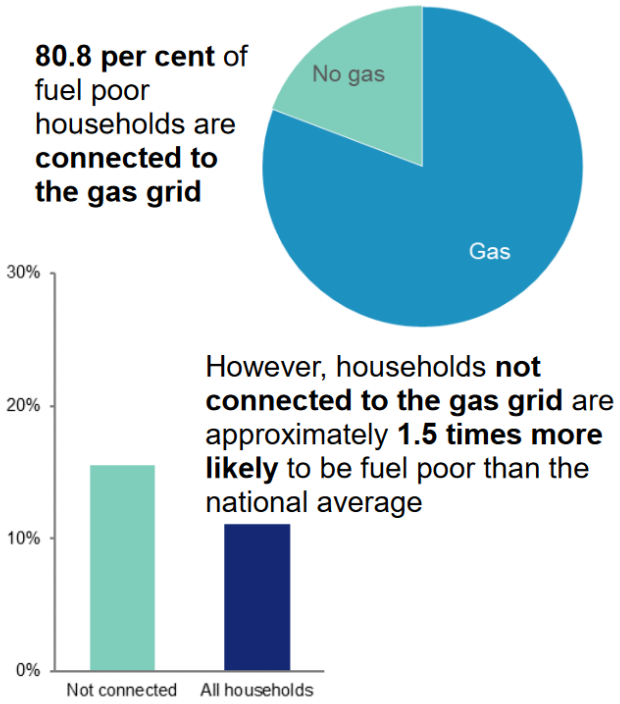


Figure 6 Gas connections - households not connected and fuel poor
(BEIS, 2018, p. 27)

AWC, (as the CEO of X understood the purpose of the project), was to use the data to provide targeted marketing data for approaches by their connection teams. Households would then need to fulfil the necessary means tested criteria to qualify for a connection (AWC CIC, 2019). Private rental tenants would also need to permission of their landlord for connection to proceed, although the scheme would be awarded on the basis of the tenants' circumstances, not the landlords.

For the purposes of the IAF framework, despite individual households (not people) being identified as being fuel poor, the data insights were just to be used for the purpose of offering affordable energy access, which is likely to meet public expectations of what would constitute reasonable data processing that would not cause harm to individuals or groups by the process of identification. If, of course, the dataset was to be repurposed for other purposes (like offering or excluding individual households from other goods or services) then ethical challenges may arise. This is where the aspect of trust in the customer arises again for X – once the dataset is passed on, they cease to have any idea how it might be used, apart from the reputational trust they assume in quasi-public entities like National Grid and AWC.

4.2 Use Case 2 – Bournemouth City Council Energy Transformation Project

This project was to deliver a GIS dataset to Bournemouth Borough Council (BBC) that identified various aspects relevant to the Local Authority energy transformation plans.

The dataset comprised of 120,000 properties across the city, with data 10 fields for each address which identified the solar potential of each roof, included selected fields from Energy Performance Certificates (EPC) (Ministry of Housing, Communities and Local Government, 2019), and identified locations for potential electric vehicle charging points (EVCP). The EVCP data included data derived from ML that identified driveways and pavement width from aerial photography, and additionally a one- off satellite imagery count of on-street parking throughout the city.

4.2.1 Solar potential

BCC were interested in promoting the uptake of solar in the Borough, which had tailed off due to the changes in government funding and subsidies for solar installations. BBC was also under pressure to build new homes and found it could not meet its government target of 1,500 new homes per year due to grid constraints in the electricity infrastructure. They hoped that by encouraging properties with high solar potential to generate power and use battery storage, this would relieve pressure on the grid to free up the potential for connecting new homes. X was unsure exactly how BCC intended to use the solar insights but thought that they would be used as part of a public information campaign, not for direct marketing. The ability for BCC departments to quickly answer residents' queries regarding the solar potential for their property rather than direct approaches to properties. The data was mainly intended for strategic planning for BCC as part of their wider strategy and planning for the City.

The solar potential was calculated using the OSMasterMap building footprint as a cookie-cutter for aerial LiDAR images of the City. X has developed Machine Learning to automatically calculate the roof pitch and aspect for property which is then geocoded (a postal address from Google Places and given a co-ordinate). The EPC database join was achieved by geocoding the EPCs.

4.2.2 Electric Vehicle Charging Point Project

As a lead project of the UK Space Agency's Space for Smarter Government Programme, to demonstrate the potential of using satellite technology to solve challenges faced by the public sector. The project aimed to identify charge point requirements via visualisation of different features and influences on EV roll out, such as existing charge points, residential driveway

availability and size, and footpath width and potential obstructions. X delivered an interactive, web-based tool, designed to support roll out of EVCP infrastructure in urban environments via intuitive visualisation, contextualisation and analysis of data in a map-based user interface.

'Public sector end users will be able to efficiently identify charge point requirements via visualisation of different features and influences on EV roll out, such as existing charge points, residential driveway availability and size, and footpath width and potential obstructions, in order to provide an accurate overview of current and potential EVCP preparedness. In addition, users will be able to identify specific features manually through the map interface, or run queries in order to model pre-determined scenarios that will support the strategic decision making process.' (UK Space Agency, 2018)

Driveways were identified using ML to pull out hard surfaces from a data that combined satellite imagery with OSMasterMap cadastral (property) boundaries. The same process was used to calculate pavement width, combined with the street light maps, to give an indication of the possible location for on-street EVCPs. The driveway data, combined with the solar potential for a property, identifies properties which could potentially charge their electric vehicle from their own solar installation (with the addition of battery storage to meet charging demands for car batteries that happen more frequently at night).

X also supplied a snapshot of on-street parking volumes to further assist with planning where on-street EV points could be sited. This data has also fed into thinking around dynamic street parking systems served by an app, where on-street and private parking potential can be viewed and rented, and available EV charging points are identified. The pavement width data is also used to identify areas where pavements are wide enough to allow for the addition of bus and cycle lanes. The data set is hosted on a portal with a dashboard allowing access to interrogate and report from the data set, with logins supplied to various departments across the Council.

BCC asked X if the dataset could be released as Open Data, which would enable companies to come in and use it as a marketing tool to sell households products. X have said no as they want to protect what they have built (IP), and also to retain the ability to re-sell the dataset multiple times, as this is the business model for generating returns on geospatial insight data. There are also licensing issues, for example OS may well agree to an open licence for a Local Authority, but Google certainly would not. It also leads to questions about how this data could be used if it was open data - Is it ethical to make it open data knowing that solar companies (for example) will use it to knock on doors? How could interested parties use this data for unforeseen applications which may not meet the expectations of either the Local Government body, or the households identified in the data set.

4.3 Use Case 3 – Newcastle Solar Energy Usage Project

This project was commissioned by Northern Power grid to estimate where they might need to reinforce their system, and to investigate how data insights could be used for grid management – more specifically how solar energy flows in a grid. This required the identification of solar panels on roofs, and while imagery cannot detect if the array is passive or photovoltaic, it is possible to use LiDAR to calculate the pitch and aspect of the solar array combined with real-time weather data from the Met Office (2019). These insights can enable an estimate of the energy flows, and then link to the grid map and electricity substations to estimate how much electricity the solar is likely to be pushing back to the grid. While this project did not have access to smart meter data or any data regarding energy usage, it does include grid data from Northern Power identifying substations and capacity. From this it proved possible to make predictions about usage and export of energy to the grid.

The future development trajectory for this project is to combine the GIS insights with electricity grid company data in real-time about the power being drawn from each substation. This will enable large scale identification where solar generation currently occurs and identify where new solar installations can be installed relative to the constraints of the grid – energy project identification. It also entails being able to make models of individual households' energy usage, and therefore to gain insight into household activities and composition.

4.4 Discussion

> 'I think one of the ways geospatial data has skirted privacy is that we can't see inside someone's house, it's an empirical observation by the nature of it, but we are starting to get to the point of being able to look at behaviour.' (comment by CEO of X)

As this case study indicates, the increasing ability to gain insights and infer knowledge from what have traditionally been considered fairly uncontroversial data sets (e.g., OS maps) pushes at the boundaries of what individuals might consider 'fair use' and justified intrusion, particularly as there is no form of consent in these models. The examples given above all seem to be dealing with data in a way that is fair given the purpose of the processing, but as the CEO of X points out, industries based in GIS have traditionally not considered the ethical boundaries of the data processing itself if it does not contain personally identifiable data (like names and contact details). This does leave open the question of data processing based on maps and addresses – it is only one short, and often publicly available hop, from an address to a person. Social sorting through mapping and interpreting

spaces and features and the ability to gain insights into behaviour, and then to use that information for marketing, behavioural nudging, or possibly social exclusion.

If we consider how applicable the IAF EDIA is for the case study in question, it is irrelevant. The kind of processing in these examples, although they could meet the requirement to conduct an EDIA as they do deploy 'advanced analytics such as: evaluation or scoring (including profiling and predicting), automated individual decision-making, systemic observation or monitoring, data processed on a large scale, matching or combing data sets, innovative use or applying new technological or organizational solutions (such as AI and ML).' (IAF, 2019) In reality, the CEO of X applies his own personal ethical code of conduct (i.e., has thought through the purpose and the implications for the processing), and has borne in mind due diligence aspects like adherence to regulations (those for personal data) and to the contracts and licenses attached to the data he is using. He has processes in place to make sure that the data is treated securely and that legal aspects are respected. This has much more to do with his own personal value system than an adherence to a checklist of externally provided ethical principles.

There are of course good business reasons for behaving in an ethical manner. Many of X's customers are public or quasi-public bodies, and the business has an interest in behaving ethically in order that they maintain the trust of both customers and data suppliers. (X is, for example, a 'trusted partner' with the OS, giving free access to OS products for design and development, with licenses only being issued for data that is used in the final product X delivers to a customer.

Intuitively using an ethical framework because you are a 'good actor' still leaves the field open for those 'bad actors'. We cannot rely on the goodwill/ethics of each CEO, Board or manager – so how do we ensure suitable oversight? Exactly same problem as is faced in domains like environmental risk and finance and accounting. The extraction of insights from ever larger data sets creates value, but can also result in concentrations of power, the new face of the complexities of managing the risk technologies in modern societies. How will societies develop effective responses, and whose voices will be taken into account? As noted in the discussion of the IAF EDIA, or in the EU's Trustworthy AI framework, these tools are being produced by either industry interests or high-level academic and governmental actors. Hardly representative of the range of stakeholders who are involved, especially missing, for the most part, are the voices of those on the receiving end of much of this innovation.

Chapter 5 Results - Document analysis

Chapter 5 provides an assessment of the myriad of frameworks, principles, templates, guidelines, and protocols that have arisen around AI through the lens of known best practices for impact assessment and audit of technology. Categorising these tools and considering how these relate to existing approaches serves to move the field forwards in operationalising ethical principles. Looking at the environmental movement of the mid-20th century, in which ethical considerations for many diverse parties, application of technology and societal concerns all converged, there are parallels for best practice in the current AI ethics, impact assessment and audit conversations. There are also robust, long-established audit and assurance practices in other sectors like financial services.

5.1 Results

The data set of 39 proposed tools for ethical AI were coded using the typologies described in Section 3.3, the results of this analysis are shown in Table 13. The documents are arranged in ascending year of publication, with the majority of documents being produced in in 2019/2020, 2020 comprising half the total. Some judgement was required in coding these documents as to whether they were an impact assessment or audit, as the terms are used with varying meanings across the AI ethics documents.

Table 13 Overall results for coded document set (n=39) see Appendix A for document details

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Title	Risks, Harms and Benefits Assessment	AI and Big Data A blueprint for a human rights, social and ethical impact assessment	ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY	An Ethical Toolkit for Engineering/Design Practice	Ethical Data and Information Management: Concepts, Tools and Methods	Ethical OS	Ethics & Algorithms Toolkit (beta)	AI Fairness 360	AI Procurement in a Box	AI-RFX Procurement Framework	Algorithmic Impact Assessment (AIA)	Codex for Data- Based Value Creation	Consequence Scanning – doteveryone	IBM Watson OpenScale	IEEE SA - The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical	Model Cards for Model Reporting	Model Ethical Data Impact Assessment	ODI Data Ethics Canvas	Understanding artificial intelligence ethics and safety: A guide for the public sector	A Proposed Model AI Governance Framework - Second Edition	AI Blindspot: A Discovery Process for preventing, detecting, and mitigating bias in AI systems	Algorithm Register	Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment	Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic	Co-Designing Checklists to Understand Organizational Challenges and Opportunities	Corporate Digital Responsibility	Data Ethics Framework	Datasheets for Datasets	Empowering AI Leadership	Fairlearn: A toolkit for assessing and improving fairness in AI	IEEE Draft Model Process for Addressing Ethical Concerns During System Design P7000™	IEEE 7010 Recommended Practice for Assessing the Impact of Autonomous and Intelligent	Responsible AI	Standard Clauses for Municipalities for Fair Use of Algorithmic Systems	Toward situated interventions for algorithmic equity: lessons from the field	Value-based Engineering for Ethics by Design	Welcome to the Artificial Intelligence Incident Database	White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions
Year	2017	2018	2018	2018	2018	2018	2018	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020
Sector produced by																																							
Public Sector							•				•										•		•	•				•				•			•				
Private Sector					•			•						•	•		•							•	•	•	•		•		•	•	•	•			•		
Not-for-Profit	•					•	•		•	•		•	•					•	•										•								•	•
Academic Research		•	•	•			•			•						•	•			•		•			•	•	•		•									
Sector used by																																							
Public Sector	•	•	•				•		•		•		•		•			•	•	•	•		•	•	•	•		•	•		•	•	•		•	•	•	•
Private Sector	•	•		•	•	•		•		•	•	•	•	•	•	•	•	•	•			•		•	•	•	•	•	•	•	•	•	•	•		•	•	•	
Not-for-Profit	•																									•											•		
Academic Research				•												•									•	•				•				•				•		
Stakeholder Type: ● Applying Tool ○ User of tool output																																							
Voiceless															○								○			•	○												
Vested Interest							○								○					○	○		○		○	•	○						○			○			
Decision Makers	○	○	○	○	○		○		•	○	○	○			○		○	○	○	○	○			○	○	○	•	○		•	○	○	○		○	○	○	○	○
Legal	•	•	•				○				•							•					•		•		•		○					•				
Delivery			•	•	•	•	•	•	•	•	•	•	•		•		•		•	•	•	•		•	•	•	•		•	○	•	•	•	•		•	•		
Quality Assurance			•	•	•	•		•		•					•	•		•		•	•			•	•	•		•		○	•	•	•	•		•	•		
Procurement							•		•	•	•								•	•	•			•				•		○					•	•			•
HR																																							
Developer			•	•	•	•	•	•		•	•	•	•	•	•	•	•	•			•	•		•	•	•		•	•	○	•	•	•	•		•	•	•	
Users																			•	•																			
Oversight							○		○						○		○	○	○		○		○	○	○	○	○	○	○		○	○	○		○	○		○	○
Impact Assessment	•	•	•	•		•	•				•	•						•	•	•	•			•		•		•							•			
Checklist questionnaire	•	•	•	•		•	•				•	•						•	•	•	•			•		•		•							•			
Baseline study																																•							
Participation process			•				•																			•										•	•		
Cost-benefit analysis	•																																						
Risk assessment	•				•		•				•			•				•			•			•				•								•		
Life-cycle assessment																																							
Change measurement																																							
Expert committee		•																																				
Governance process	•			•								•						•		•	•																	
Procurement process							•		•	•										•																			•
Audit					•				•	•				•	•		•						•		•		•		•		•	•	•	•			•	•
Checklist questionnaire									•	•																				•									•
Documentation					•									•	•		•			•				•				•		•		•	•	•			•		
Reporting					•									•	•		•							•							•	•					•		
Governance process					•									•	•										•				•			•	•	•					
Technical & Design Tools						•							•			•						•					•		•							•		
Workshop materials						•							•			•						•			•											•			
Documentation														•															•					•			•		
Technical tools								•						•																•				•					
Internal or external assessment/audit																																							
Internal/ self-assessment	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•			•	•	•	•		•
External/ 3rd party															•								•								•	•					•		
Stage in process tool applied to: ● data and/or ◊ model																																							
Business/use case	•	◊	◊						◊	◊	• ◊	•	• ◊					•	•	• ◊		◊		◊	◊	◊	◊	• ◊	•		◊		◊			◊	◊		◊
Design	•	◊		◊		• ◊	◊				• ◊	•	• ◊		◊	◊			•	• ◊		◊		◊	◊	◊	◊	• ◊	•		◊		◊	◊	◊	◊		◊	◊	
Training data assembly					•			• ◊				•			◊	◊			•	• ◊									•					◊	◊			◊		
Building							◊	• ◊				• ◊														◊					◊									
Testing								• ◊				• ◊					•			◊						◊			•		◊			◊	◊				
Deployment											• ◊	•		◊	◊		•	•		◊	◊					◊	• ◊				◊			◊			◊		
Monitoring												• ◊		◊	◊				•	◊																◊			
Procurement	•		◊				◊		◊	◊		• ◊						•			◊			◊	◊					◊				◊	◊				◊

5.2 Analysis by sector

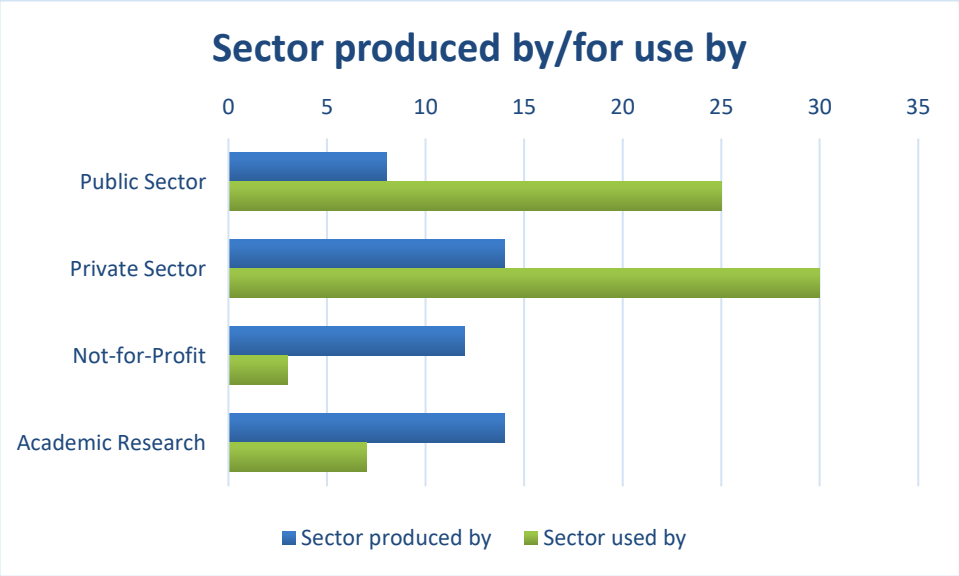


Figure 7 AI ethics tool by sector produced by/for use by

Figure 7 illustrates the main sectors who are either producing ethical AI tools, and compares this to those sectors for whom the tool is intended for use. It shows the main sectors targeted by ethics tools are the public and private sector, which reflects the main sectors where AI systems are being designed and produced (private sector), and the concerns around deployment of AI in public sector institutions. There is also interest from the academic community in AI ethics tools and how to address these issues, with the not-for-profit sector (civil society, NGO's and think tanks) also looking to provide solutions to ethical issues in AI production and deployment, although not-for-profit are not producers of AI systems, some sectors of not-for-profit (like development agencies) do deploy these systems. It is interesting to note that it can be difficult to separate academic research from private, corporate research in AI as there is strong cross-fertilization between these, with scientists moving between sectors, and technology companies funding their own research outputs, and funding university research.

5.3 Analysis by stakeholder

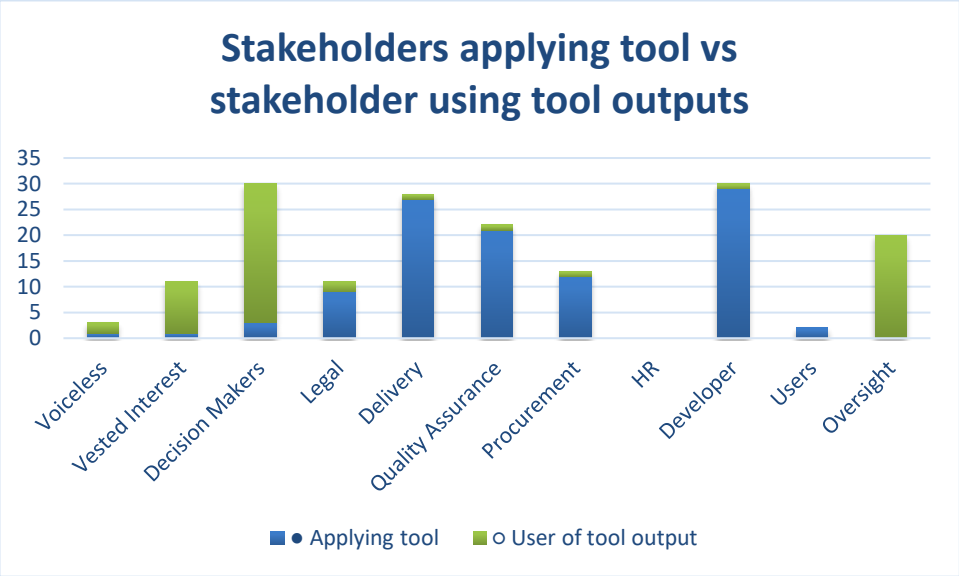


Figure 8 Stakeholder type using tool vs stakeholder engaging with output from tool

Figure 8 shows the number of tools that include which type of stakeholder in their terms of reference either as producers of artifacts, or consumers of the product. For example, a developer team uses an ethics tool to assess a system which produces an output (e.g. report). This output can then be released to other stakeholders who can act on or respond to the findings. As might be expected, the stakeholders who are likely to be applying the tool are mainly in the production side of AI systems (developer, quality assurance and delivery roles), with the results of the tool being used by decision makers and senior staff. The tools can also comprise evidence for shareholders and citizens, and oversight bodies. Despite participation processes being recommended in some impact assessments (see Figure 9 Types of tool suggested to produce impact assessments), we can see that the range of stakeholders involved in the proposed tools only really captures those involved in producing AI systems, or procuring them, and wider stakeholders (to whom negative impacts of deployment of an AI system actually accrue i.e. users and wider stakeholders in society) are not included in these processes.

5.4 Elements of impact assessment tools

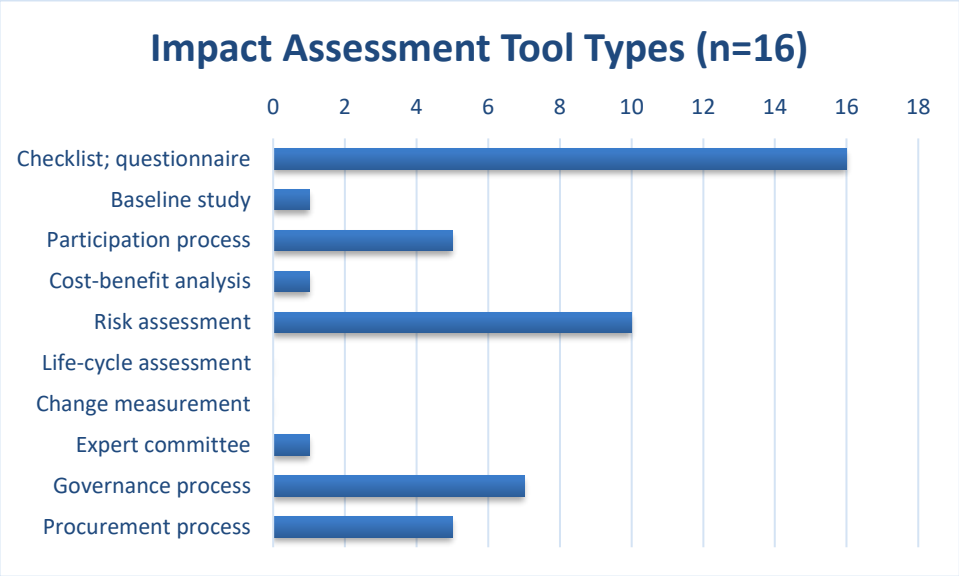


Figure 9 Types of tool suggested to produce impact assessments

Figure 9 represents the number of component tools used within an Impact Assessment. A checklist or questionnaire is used in all 16 Impact Assessments coded in the study (as compared to only 4/16 audit tools Figure 10 Types of tool suggested to produce audits). It is a structured way to record proposals, decisions and actions, and can also be used to embed a governance process for the process of applying an ethical tool. Risk assessments were also commonly included, often embedded as part of the checklist process. Impact assessments were also used as part of a procurement process to assess ethical impacts and risks of purchasing an AI system. Unlike other types of impact assessment like EIA, little attention was paid to measurement of baseline conditions or predicting change. There were also omissions in these proposed tools for AI which did not include the types of impacts that would be measured in a life-cycle assessment for a product or process, leaving out key considerations like resource or energy use and sustainability.

5.5 Elements of audit tools

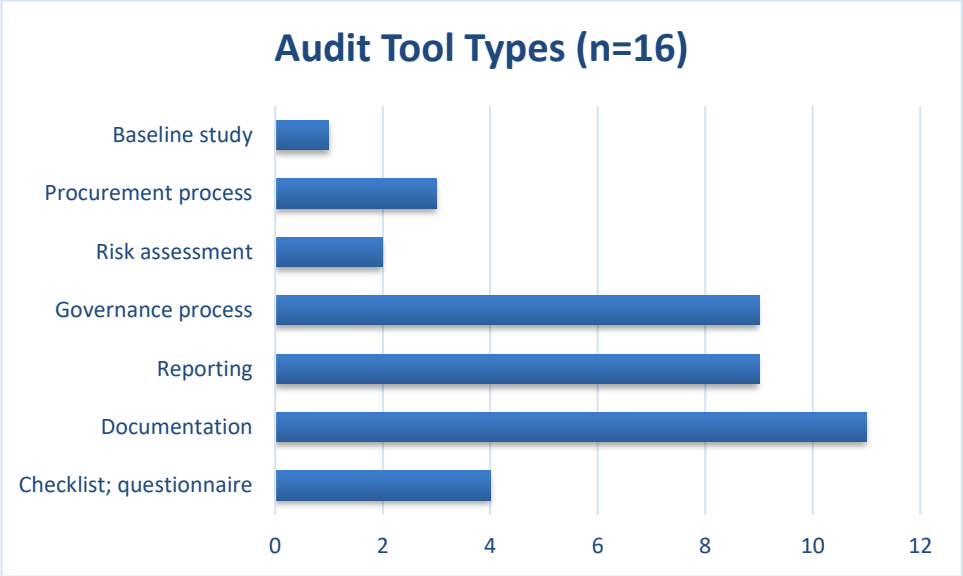


Figure 10 Types of tool suggested to produce audits

Figure 10 shows the tools types identified in ethics tools that are categorised as audits. The focus of these is on appropriate governance processes, reporting and documentation for verification and assurance in the audit process, unlike the impact assessment tools in Figure 9 which rely on checklists.

5.6 Internal vs external assessment

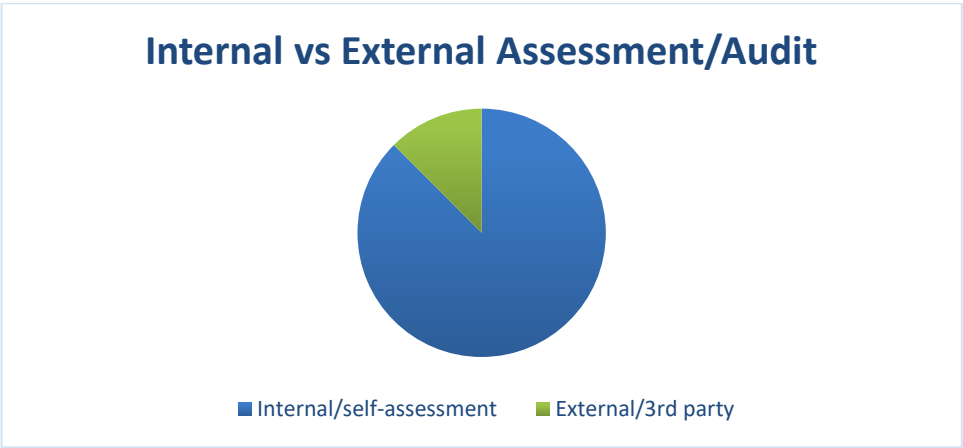


Figure 11 Internal/self-assessment vs external/3rd party assessment/audit

Figure 11 illustrates whether the ethical tool is an internal assessment or audit, as opposed to a verification process from a 3rd party. External verification only occurs in 5 of the 35 tools analysed, surfacing in either the certified standards from IEEE or in tools like incident databases which are designed for transparency.

5.7 Technical tool types

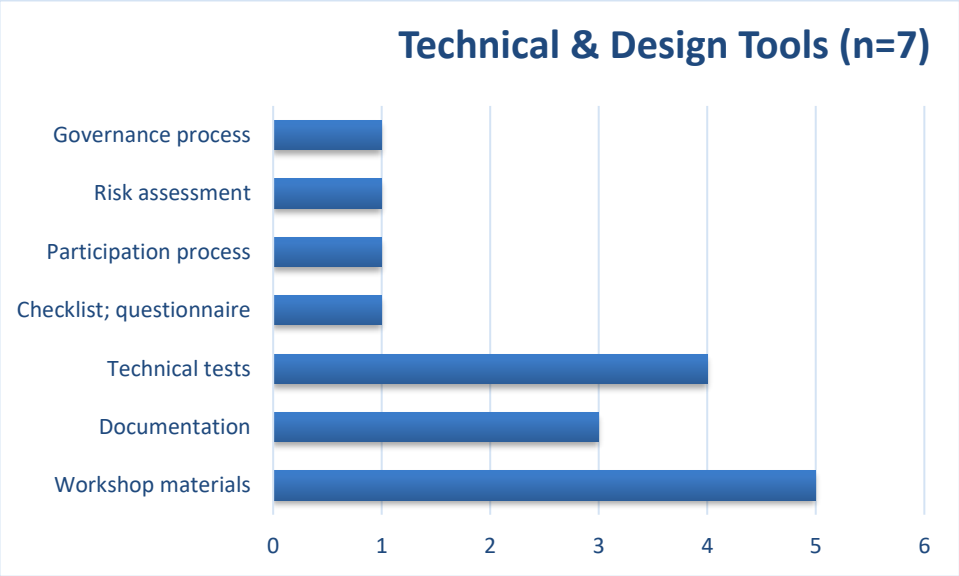


Figure 12 Technical and Design Tool Types

Figure 12 breaks down these tools into workshop and design tools, forms of technical documentation, and tools for testing or monitoring data and models. The workshop materials do not fit into an impact assessment or audit framework and are not designed to provide verifiable evidence of process, but more to elicit 'ethical thinking' from design teams, unlike the documentation tools which can provide evidence for audits. The technical tests are part of creating robust systems and can also provide an audit trail.

5.8 Stage in production pipeline

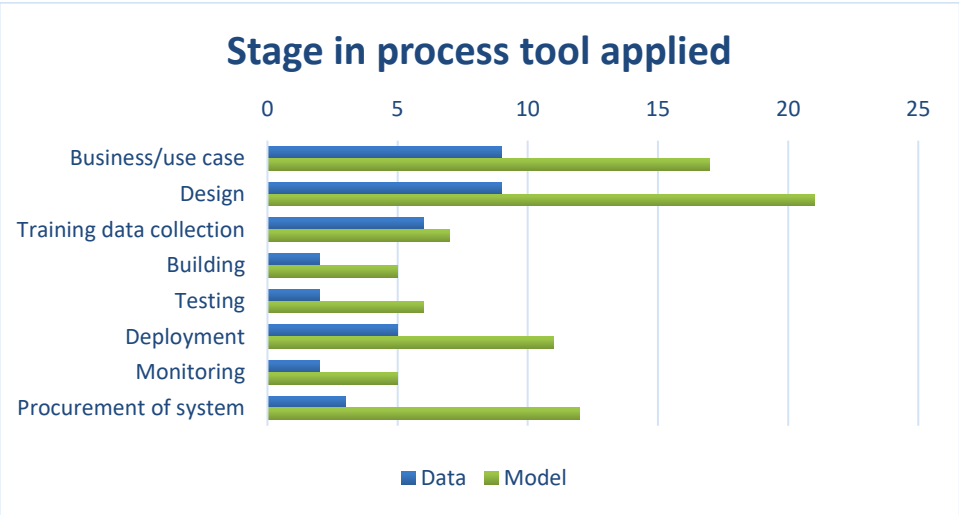


Figure 13 Stage in process tool applied in production process

Figure 13 illustrates the stage in the production process pipeline that the proposed tools apply to, and also categorises these tools depending on whether they are focused on the data or the model. Many of the tools are designed for use early in the process – at the use case and design phase, where the main focus on the model is found. The attention to the model is also more marked in the deployment and procurement process, with data also being an important object for assessment early in the process.

5.9 Key findings

The available guidelines cluster around the product development phase of AI and are focused on being used by and documenting the concerns from developers, delivery, and quality assurance roles. The reporting output from these tools is then used to inform management decisions, as well as shaping developers' workflows resulting in better practice. There is little participation in the assessment or audit process by certain stakeholder groups, particularly the voiceless, vested interests and users, who are not included in the process of applying the tools or interacting with the outputs as tools for transparency or decision-making. Nearly all of the tools available are for internal self-assessment, with only the IEEE standards requiring any kind of external verification, and the two examples of public registers providing explicit transparency. In addition to missing large stakeholder groups, the current set of AI Guidelines and tools do not fully utilize the full range of techniques available, including; participation process, baseline study, life-cycle assessment, change measurement or expert committees. Finally, it is noted that there is no regulatory requirement for any utilization of impact assessments or audits within this field at the moment, minimizing likely adoption and true application of them.

a. The focus has moved from data to models from 2017 to 2020. Earlier documents were often concerned with issues around 'big data', with concerns shifting to models and algorithms. This does not mean that data is not considered in these later iterations (particularly training and test data), but the focus shifts from a more traditional data protection approach.

b. Stakeholder types directly using the tools are clustered around the product development phase of AI (developers, delivery, quality assurance), with the output from the tools (reporting) being used by management decision makers.

c. There is little participation in the assessment or audit process by certain stakeholder groups (voiceless, vested interests and users) who are not included in the process of applying the tools or interacting with the outputs as tools for transparency or decision-making. Perhaps most surprising is how little inclusion there is of users/customers in these tools.

d. Nearly all of the tools are for internal self-assessment on a voluntary basis, with only the IEEE standards requiring any kind of external verification, and the two examples of public registers providing explicit transparency.

e. Techniques and practices deployed by other forms of Impact Assessment (like EIAs) are not present or rarely suggested in ethical AI impact assessments (participation process, baseline study, life-cycle assessment, change measurement or expert committees.)

f. Checklists/questionnaires are ubiquitous across impact assessment tools. Audit tools less frequently use checklists but do require documentation of processes.

g. The output from the tools can provide documentation for oversight from external actors, but as the majority are internal activities there is generally no process or requirement for the wider publication of the results of these tools.

h. A third of the impact assessment tools focus on procurement processes for AI systems from 3rd party vendors, indicating the need for not only producers of AI products to engage with ethical assessment, but also the customers for these products, who will be the ones deploying the products.

Chapter 6 Results – Industry interviews

10 semi-structured interviews were conducted with senior decision-makers (c-suite/founders) in small tech companies (see Section 3.8 for a description of this process). Information about the participants and their companies can be found in Table 10. The interviews were divided into three parts, first to elicit details about the participants and their companies and products, and any reflections on ethical issues they perceived. The next part of the interview used four examples of ethics tools for participants to comment on. Finally, they were asked to reflect on regulation, innovation and provide any further thoughts on the topic.

6.1 Overview of responses to 'Describe your company and products and any ethical challenges?'

Participants overall discussed data protection, privacy, and security issues as main ethical concerns. Consideration given to issues directly related to their product and market, and any potential reputational risks.

In the first part of the interview respondents were invited to describe their own products and what their perception (if any) of any ethical issues. Their responses showed that they had were very aware of the specific ethical challenges of their own products which may affect their ability to successfully sell their products. Data protection, privacy and security emerged as a key theme across the respondents regardless of the marketplace. Linked to this was protection of their reputation, and de-risking themselves from possible scandal. This included screening of customers for potential blow back risks to themselves. It was clear from this section of the interview that respondents were thoughtful and reflective across issues which were directly related to their product R&D and sales pipeline and felt that their processes reflected appropriate ethical considerations in their own domain.

6.1.1 Privacy and data protection as a central concern

All of the respondents, regardless of the nature and application of their product had thought long and hard about risks around personal data, and specifically designed their systems to make their liability as minimal as possible. Either protecting data, or minimising liability for data drove important design and sales decisions. For example, in the health application a small range of lower risk conditions was chosen to minimise the data needed for analysis, the retail shopping application

was designed to avoid collecting any identifiable data about end users. In the biometrics space their products revolved around creating secure, encrypted representations of biometric patterns, and for the secure messaging company their whole USP was around not holding any customer data beyond basic download and sales transactions.

6.1.1.1 Privacy and security – biometric data

Biometric companies keenly aware of the privacy and security issues. Avoid responsibility for personal data where possible. Customer takes responsibility for the data.

Respondents in the identity space work hard to develop systems that do not fall under the category of Personally Identifiable Information (PII), and to have the processes they have developed not be considered under specific parts of biometric data protection regulation. This is done by, for example, creating irreversible tokens from the biometric record which removes the need for storing reference templates, therefore improving security. Other approaches use behavioural signals to monitor the user continuously when on the client's website or app, where ML is used to monitor for any anomalies in behavioural patterns and flag these as potential issues. As the data breach of security platform Biostar 2 in 2019 illustrates (Taylor, 2019), poorly managed authentication systems pose high threat levels to both organisations and individuals.

> 'Our team was able to access over 1 million fingerprint records, as well as facial recognition information. Combined with the personal details, usernames, and passwords, the potential for criminal activity and fraud is massive. Once stolen, fingerprint and facial recognition information cannot be retrieved. An individual will potentially be affected for the rest of their lives.' (Rotem and Locar, 2019)

Privacy and security issues are key drivers for innovation in this sector. Research and development is focused on how to create systems that are highly secure and meet the requirements of legislation (data protection and biometric) in the jurisdictions they sell into. The systems they have designed do not collect or store any personal data, that is done by their customers to avoid any of the problems that come with privacy issues, data breaches and complying with data protection legislation in differing jurisdictions.

> *'We're simply selling tools to people to be able to protect the identities when they're in a system.' (Respondent #1, 2020)*

The responsibility for data lies firmly with the customer deploying their system. Another respondent described how privacy concerns meant they do not wish to enter the market in

managing customer data, choosing only to sell the tools for customers to protect data, but leaving the ultimate responsibility for the personal data to lie elsewhere.

> *'They [the customer] run and operate it under their own, you know, SLAs [Service Level Agreements] and terms, and the collection and the storage and all that is done by our customers, there are peers of ours who offer that as a service. But to be quite frank, again, the concerns around privacy to me just mean it's not worth it. You know what, while there may be another dollar to be made to run and operate the service that we're the experts in, the can of worms that you open around national legislation and where is data stored, and all that kind of stuff to me, so far anyway, it just looked like yeah, I'm not going to do this.' (Respondent #3, 2020)*

6.1.1.2 Privacy as a driver for design – retail recommender

Privacy as a driver for product design for retail personalisation system to avoid the type of data collection used by the traditional adtech ecosystem.

Privacy preservation was a key design consideration for the respondent in the retail recommender market. Their product was designed not to keep any identifying information about users, or to collect data for targeted marketing.

> *'As a privacy focused start-up, we're very keen to keep people as anonymous as possible. So whenever people answer our questions, we basically just assign them a unique identifier. And that's all we ever know about someone. We don't store any kind of emails or ask for any contact information. Because we don't think that you need to have all this kind of tracking and give all of this information. It's like, what do we actually need to make a good recommendation to this person?' (Respondent #9, 2021)*

They do keep analytics on customer engagement with their product which they also supply to the retailer, but described this as a 'completely anonymous aggregated, analytic' which they felt did not compromise the privacy of users but was a useful data source for their customers (retailers). The product was explicitly designed to move away from tracking and collecting data from users to make purchasing recommendations, and to ask users explicitly for information on their preferences to make recommendations.

6.1.1.3 Privacy concerns in lead generation – GIS insights

Manage potential misuse of data insights for lead generation and marketing in GIS data.

Geospatial insights for projects like renewable energy projects or EV charging points create commercial potential for lead generation and marketing. The respondent in this sector regarded this aspect to be the main risks in the data they worked with.

> *'We have to be careful that what we're providing is not being gone a contravention of any of the privacy rules that have come into play over the last few years. And balance what gives our client what they want, but not so much that it becomes kind of intrusive.' (Respondent #5, 2021)*

They have been thinking about the projects they deliver, and proactively cleaning up the data to clients, by, for example, omitting address data.

> *'So one of the things we've been trying to do is actually reverse that and not include address information in what we're doing. To try and kind of just clean up our data a little bit in that in that sense, and make it a bit harder for people to use it and just kind of blanket bomb people with mailings or canvassing and that kind of stuff…. I guess it's sort of trying to draw a line on our data and say that our data is the structural or the physical. Yeah, look, it's not the rest of it. So that's kind of how we've been trying to tackle it a little bit.' (Respondent #5, 2021)*

6.1.1.4 Privacy versus security – secure messaging

Ethical and moral dilemmas in secure messaging market. Right to privacy of communications versus national security and law enforcement.

The respondent in the secure messaging marketplace described ethical debates across his sector about the balance between privacy and security. Providing a completely secure platform where there is no access to the data flowing the product raises important questions about the freedom to have completely private conversations versus the need to monitor and extract evidence to prosecute criminal activities or for counterterrorism.

> *'But of course, that does bring massive ethical and moral challenges. Because if you're imagining the system, and you've seen it in the news that is so secure, that law enforcement can't monitor can't break into it can't do that, there's obviously a chance that that will be used for bad purposes.' (Respondent #10, 2021)*

Their target market is high net worth individuals, and organisations like hedge funds but they are very aware that the reasons users may want complete privacy is because their activities are criminal. There are of course other users like journalists or political activists who may need complete anonymity for their own protection. This has caused much debate within their company about whether to screen potential customers.

'How do you maintain the privacy of the individual in the organisation whilst still not wanting bad things to happen on your system and being open to helping authorities? And so the decisions that have to be made there about who has access to it, who makes the decision about what's right and wrong, what pre-screening, if any, you do? What do you look for? Because, as you know, one person's terrorist is another person's freedom fighter.' (Respondent #10, 2021)

6.1.2 Business and reputational risk

Business and reputational risk a key consideration across respondents.

Business and reputational risk was a concern for vendors when considering how their customers might use their system. This took different forms. Firstly, to ensure their product was not responsible for privacy and security breaches (privacy and security by design). Secondly, to consider the nature of the customer and the uses to which their product, insights and data might be put. For some use cases this led the vendors to be very circumspect about who they would sell to, but in some cases to acknowledge they could not control this. Another approach was to limit the inclusion of personal data in a product, leaving the responsibility for data with the customer.

None of the interviewees was operating in a direct relationship to end users of their product or service as they were all either selling B2B or B2G or procuring systems. This raises the question of accountability across vendor chains. For some respondents serious moral and political issues arose, especially in the biometrics and secure messaging space. Issues of reputational risk arose for respondents when they talked about who they sell to, this may be in terms of damaging the reputation of their product or business if they are involved in a scandal. For the public sector actors political risk was also an important consideration.

6.1.2.1 Business and reputational risk - vendors

Questions of business and reputational risk which is something senior decision makers/CEO's/founders consider carefully in business model and use case development. These are questions that a continuously reviewed, but often without being able to completely close the window of opportunity for negative outcomes. Questions arose in this space about the behaviour of their customers and how their use of a product or data could affect the vendor, where the customer was the one responsible for storing and using the data in a system.

'Question is, to what extent do we get involved in this [customer holding data]? And understanding what that customer will do with our technology? And what knock-on effect is there to us? Because it's, you know, its mission ending for us.

> *If we get associated with scandal, we are too small to survive a big scandal.' (Respondent #1, 2020)*

Respondents generally felt they did have an ethical responsibility to ensure their products and services were used in ways that fitted their own personal ethical position, as well as treating it as a key part of ensuring they protected the reputations of their companies.

> *'But that still doesn't mean we shouldn't take an ethical responsibility to make sure it's not going into something which is going to be used in the wrong way. And to damage, you know, to damage people in any way. That is a very important principle.' (Respondent #1, 2020)*

The company selling geospatial insights had also thought carefully about how customers would use their data. Currently much of their work serves public sector clients, which the respondent felt were reliable and reputable clients, as they are perceived as trustworthy, and they were using the data and insights for strategic not commercial purposes. As the company moved to selling insights to large commercial interests, they felt they would have to be more careful about what they supplied, and to monitor, where possible, how their data and insights were being used.

All the respondents were aware of the dangers of data breaches and lax security for their business and products and depending on the nature of their product expended considerable time and resource on these issues. They all understood what the commercial impacts of not addressing privacy and security issues presented for them, in terms of damage to their business, but risk perception obviously differed across companies depending on how exposed they might be to negative events.

6.1.2.2 Business and reputational risk - public sector

Public sector purchasers of AI systems put the privacy and security of citizen data at the forefront. Value for money and possible political damage was also a key consideration.

The respondent working for a LA 'smart places' programme described the concerns of purchasers buying in AI systems. Avoiding negative political repercussions for being perceived as profligate by wasting tax payers money, or compromising citizen data and privacy were high priorities.

> *'What the council basically understood, quite rightly, is trust is everything. And the ability to tell a reasonable story and to demonstrate that information that is collected, stored and utilised in a smart place system is used in a responsible way, its everything right?... If they get this wrong, and there's a data breach they'll lose trust. And the big concern I have about the whole rush to open data is people are going to look at it and go, what an absolute waste of council*

> *taxpayers money. Why on earth are you paying for this sort of rubbish?' (Respondent #4, 2021)*

The risks of collecting new data and bringing disparate data sets together for new insights was a major concern as was maintaining oversight of data and its processing in order. The local authority was keen to avoid any reputational harms and loss of trust from citizens in new technology projects. They work with UK government bodies like the Centre for the Protection of National Infrastructure (CPNI) and the National Cyber Security Centre (NCSC) to evaluate risks.

6.1.3 Procurement – public sector

Procurement of AI moving beyond standard local government procedures to work closely on understanding a potential private sector partner.

The respondent described the how the focus on trust in the programme meant that they had moved beyond the standard government procurement processes.

> *'So standard procurement within local authorities still follows the tick box mentality. And there are criteria around that. As a supplier, you get sent a formula, are you? You know, do you not have any slaves and, you know, blah, blah, blah, right.' (Respondent #4, 2021)*

The respondent thought that standard procurement just flagged up issues which may be a red light (financial issues, country it was based in, basic compliance e.g. having relevant ISO standards in place), but the LA wanted to move beyond the standard checks and work closely with potential technology partners to really understand them and their business before engaging in any public-private partnership projects.

6.1.4 Vetting customers to mitigate risk

Several respondents described their awareness of the risk to their companies from misuse of their products by customers. Companies are very aware of reputational risk resulting from the use of their products and if the risk is perceived as substantial will take measures to vet customers where possible.

6.1.4.1 Biometric systems

The potential for misuse of biometric systems lead to vendor closely vetting customers.

Trusted customer issues arise for systems that store and process biometric data as the technology can be used against, rather than in the best interest of, customers and citizens. The potential for the use by states against their populations was a concern for one of the respondents.

> *'If this got into the hands of a very evil actor, who wanted to use this to store the identities on the grey list of people they didn't like politically in the wrong country, be a problem. Definitely. And we will never sell to anyone that you think might be like that. So be very, very, very, very careful. Where we sell this to will only be people that we trust, and very, very large places.' (Respondent #1, 2020)*

6.1.4.2 Secure messaging platform

For some vendors despite acknowledging the risks of misuse of their product, the inability to complete effective due diligence on customers resulted in choosing not to conduct any form of vetting.

For the secure messaging company there had been a long process of developing their current position on customers for their products. Previously they tried using a due diligence process for customers to try and establish that their product was going to be used for legitimate purposes. Other businesses in the same market were just using a basic set of terms and conditions, which customers signed up to and then gained access to the product. The respondent described an evolving process where eventually the decision was made not to try and vet users, but to adopt the same approach as other competitors.

> *'We tried to put together a list of kind of pre-screening questions, and I read it afterwards. And it doesn't make sense, because I'm gathering this information about people for me to make a decision on it. I'm not doing a full background in criminal records check as a small commercial organisation. I think that would still have been easy to get round, where it's relying on me making a decision or the company making a decision. So what we've decided now is just to make it available on the website, so anybody can go to the website, anybody can download it, and pay for it on there. So with that, yes, you have still you've got the tiniest bit of a chain, of who's purchased it, because there's obviously the credit card purchase that that goes through, but still, we're not able to get access through to the information. So we've decided basically to go along the same route as the rest of the industry, which is, we were fighting a losing battle.' (Respondent #10, 2021)*

6.1.5 Geospatial redlining and potential bias

Potential for bias and discrimination from geospatial insights.

A concern for the respondent creating geospatial insights was the potential for a form of geospatial redlining. Assumptions can be made about individuals, groups and geographic areas from this kind of data, influencing, for example, who may be offered renewable energy installation, retrofit measures, or the placing of EV charging points.

> *'We can start to get bias towards certain things. And that is one thing I am slightly concerned about when doing machine learning, even of the built environment. I think you can start to typify places. And then by its very nature, you'd start to typify the people that are in those places. Again, it might not be us doing that. But our customers might do that.'(Respondent #5, 2021)*

The respondent noted that there did not seem to be any way to prevent this on the part of the vendor.

6.1.6 Bias in medical records, medical research and treatment

Positive impacts on bias and transparency in patient records and treatment through sharing records and enabling patient to log their own data.

The respondent in the health sector also raised issues about bias in medical data both from medical records and from medical research. Their product uses GP records to generate insights for the delivery of hyper-personalised care in chronic conditions. They raised the issue of bias in medical records, and that they had found some anomalies in the historical patient data they were working with which was possibly related to bias in record keeping. The respondent was not able to explain further due to confidentiality. They felt that their system could be a useful tool in reducing bias in record keeping and treatment, as the AI could flag up anomalies, but also because the system they were developing gave the patient access to their records, allowed the patient to add their own reporting, and the ability to challenge medical records.

> *'So that's why we strongly believe in having patient control over their condition. So that they can help monitor and track their own health… if we start small, on the conditions that we're focusing on, which is diabetes, high blood pressure, asthma, and COPD. I think that empowers the patient a lot. And helps make the data set a lot less biased. I mean, it's a bit hard to argue if it's less biased or not, but you've got data from both sources, at least.' (Respondent #7, 2021)*

The respondent was very keen on empowering patients to take ownership of their own data and treatment and felt that using a health platform where the patient could contribute their own data, and see their records contributed by health professionals would make the process more transparent and give patients the ability to challenge the judgements made by those treating them.

6.2 Responses to example ethics tools

The second part of the interview presented four different examples of ethics tools for discussion. Respondents were asked for their opinion of the tools in turn, and to comment on how they might, or might not, be applicable to their own context (see Section 3.5 for a discussion of the example tools).

6.2.1 Responses to Example A – IBM Principles

6.2.1.1 Ethics statements as a branding exercise 'ethicswashing'

Public-facing ethics statements generally understood to for marketing and branding purposes.

The majority of respondents described Example A as a marketing or branding exercise which was either just *'management spiel'* or a corporate tick box exercise, but some thought it was good to publicly commit to principles. One respondent felt they already stated their ethical stance on their website in a similar fashion, and another was prompted to consider producing a similar document for their own marketing materials.

> *'I think we have probably now just grown to the point where we need to do something like this.' (Respondent #1, 2020)*

Some respondents were very cynical about such a document which presented *'fairly anodyne'* principles that anyone could agree with. One respondent, a former IBM employee, said they were very brand aware, but that profit was the most important thing to them.

> *'IBM can be a right bastard to its customers. It can be bloody evil. Because it's all about driving the bottom line meeting that quarterly profit that keeps the shareholders happy that drives the organisation forward.' (Respondent #6, 2021)*

6.2.1.2 Public statements vs internal behaviour

Clear identification of the potential gap between public statements and internal company behaviour.

Respondents from the larger more established companies reflected on how these statements affect the behaviour of staff. One commented

> *'It feels like you're checking a box? You know, you've got one. So we have to have one, I don't think internally any of the staff would think any differently from some kind of grandiose vision statement, you know, I think it's kind of implied or inferred.' (Respondent #3, 2020)*

Another thought the opposite and understood these kinds of statements to be useful in cascading standards and expectations to staff which should come from the top.

> *'Is the vision, always still consistent? And we still talk about that a lot at [REDACTED CO NAME] on all the calls, you know, that the mission is really important. And it is more important, actually, we believe. If you get that, right, the money will follow. But it's not about money first. Nobody is there to make money first.' (Respondent #1, 2020)*

More cynically, another saw Example A as just surfacing the corporate social responsibility agenda, expressing the legal responsibilities of the Board but not necessarily having any meaningful impact on how the company may behave.

Everyone questioned the difference between publicly stated principles and what the reality on the ground inside the company.

> *'So I think this document is nice for the public to be aware, and then to create accountability with the public. Hey, we're saying we do this. You can keep us in check on that. But unless, internally, I don't know how it works, which is, which is where the heart of the matter is.' (Respondent #7, 2021)*

Independent third-party review was discussed by one respondent as being the only valid way to ensure that publicly stated principles were applied in practice.

> *'Unless something is externally validated, all it is at that stage is a marketing statement.' (Respondent #10, 2021)*

6.2.1.3 Internal company processes – staff survey

Doubts expressed about the influence of public statements on internal behaviour. Company culture grown from within.

When asked about any policies or practices which might fall under the term 'ethics' in their business, one respondent talked about an internal staff survey delivered every month by HR for staff to feedback anonymously. Alongside general questions about their experience at work there were also questions like 'Do you understand the mission of the company?'

> *'I think that the pertinent question is, you know, I've seen questions on it that ask, do you feel the company delivers on ethics, you aspire to yourself, we do ask the team how do they feel and all that stuff gets reported up to the board, actually, and any investors that we have, that we follow up these kind of things? And just in that sense, I think the team knows and understands what we do, and that we have safeguards and processes.' (Respondent #3, 2020)*

They felt that this was an effective way to monitor how aligned staff were to the company ethos, and that public ethics or vision statements would have little influence on how staff understood the company culture.

> *'It feels like you're checking a box? You know, I've got one. So we have one, I don't think internally any of the staff would think any differently from some kind of grandiose vision statement, you know, I think it's kind of implied or inferred.' (Respondent #3, 2020)*

Another commented that:

> *'This stuff needs to be within the DNA, it needs to be steel threaded into the whole approach. The companies and organisations that leverage this this kind of capability [AI], should have in their DNA, you shouldn't need to comply with it [ethics statements]. It's fundamentally core to what you build.' (Respondent #4, 2021)*

6.2.1.4 Transparency and explainability

Explainability not a core concern except for products where transparency of process was important to customers.

Explainability was seen as important to most respondents for communicating with their customers about their products,

> *'just being able to do something is fine, but quite often it's not - you need to be able to explain it.' (Respondent #8, 2021)*

One respondent agreed strongly with the transparency and explainability statements in Example A, as these principles were an important consideration which had driven their own product development but did not see any need for such public statements about their own product. They did spend time explaining in their product materials how the integrity of creators IP was maintained

in their system, and how collaborators were justly remunerated for their work, but did not feel their own product warranted overt ethical statements.

One respondent had written their own framework for explaining their principles for secure communications. Key parts of this included transparency in the form of open-source code and using third-party review to validate their claims. Several respondents said they thought very carefully about the explainability of their products, but that transparency beyond describing product features often conflicted with IP concerns.

> *'And then there's a different thing about how transparent you can be bearing in mind that the devil is in the detail and that's where a lot of the IP is going to be anyway. So I question the amount of transparency that's going to be put in there.' (Respondent #8, 2021)*

Explainability for end users was discussed by the respondent who built the retail recommender application. They thought that it was important that end users had reasonable high-level explanations, particularly around how their data was being used in the product. Any more detailed explanation of how the product made the recommendations they considered overkill, referring to cookie banners and privacy notices as ways in which they thought user was not served well by explainability.

Respondents in other sectors felt that explainability was important for their customers and a necessary part of the sales process. Products where security was paramount (like biometrics or medical data) felt keenly the need to be able to explain their processes.

6.2.1.5 Explainability as a driver for design - music

For the respondent producing tailored music for advertising and TV the most important consideration was how to fairly attribute the creative input of human composers with the machine manipulation of their input to produce tailored music segments.

> *'When you build the engine, the tricky thing is, is balancing of the amount of creative process that a human being is involved in and how much machine is actually doing.' (Respondent #8, 2021)*

In order to maintain control over the IP of composers and fairly reimburse them for their work the respondent chose not to use a machine learning approach because of the explainability issues.

> *'If you use fully generated machine learning, and you tried to generate a song based on whatever they say, you know, it learns from, you got massive legal issues, because unless you prove that the song like the algorithm hasn't memorised when it was learning, like some chord progressions, or whatever, it's very difficult to prove that whatever it generated, actually, you have the intellectual property rights to it, someone else might.' (Respondent #8, 2021)*

They chose to use a constraint programming approach to solve the problem to avoid the issue of explainability in an ML application. They identified this a problematic area for using these techniques in general.

> *'I think a lot of people don't realise there's a big hype on deep learning. It's a very, it's an amazing set of tools, and techniques, but I think there's going to be people are going to be stepping back away from it, because it doesn't allow you to do this thing that we did, for example, where you can track things you can actually explain whatever you do. And I think it's a big one, like, ethic wise, business wise, it's a really big thing.' (Respondent #8, 2021)*

6.2.1.6 Transparency of data use

High level principles describing data use not necessarily reflected in actual collection and processing of data, or how data was shared with law enforcement agencies.

Two of the respondents were particularly interested in the statements about data use in Example A. They expressed scepticism about what this sort of high level statement would actually mean in practice.

> *'It seems like a lot of marketing terms that look impressive, but they don't necessarily make me feel any different about how IBM would use my data… At the end of the day, is the way that the data is processed. And that's usually buried down in conditions at the bottom of some footer.' (Respondent #10, 2021)*

Another discussed at length the statement in principle No.2 about access to data by government agencies.

> 'Government Access To Data IBM has not provided client data to any government agency under any surveillance program involving bulk collection of content or metadata.' (IBM, 2018)

Their own product (a secure messaging platform) had caused them to think long and hard about providing data for governments or law enforcement agencies.

> *'My sceptical view of it bearing in mind that they are a huge organisation that are very closely linked to the government departments, the military industrial complex… how you do know what is set up in the background, they wouldn't be allowed to tell you, even if they had.' (Respondent #10, 2021)*

For this particular respondent the issue of protection of privacy was a key ethical concern and had caused much deliberation within their company. Their product was a secure messaging and video

conferencing platform, built with no backdoors or access to customer data. Even if they were subpoenaed for data

> *'we're collecting no personal data about the user from their application, we're only storing very, very basic trend, log files, which are just for debugging purposes, and wiped automatically after seven days.' (Respondent #10, 2021)*

Ensuring no user data was accessible or retained was the key selling point for their product, but of course a completely secure platform raises the question of who their customers are and what they are using the product to protect.

They went on to muse on gag orders and specific requests from government and law enforcement agencies for customer data from companies.

> *'We have, for example, on our web page, there's a transparency report. And that report lists, any court orders that we receive as an organisation, you kind of go great, and people like it, and it's the de facto setting in the industry today. There's no regulation that says we have to do that, right? And so we could just choose to ignore it. And there's a hypothetical, that if we're told not to publish, no one would know, you have a blank page.' (Respondent #10, 2021)*

The company had been through a process of trying to vet prospective customers to screen out criminal activity but had decided in the end that it was just unworkable and to allow anyone to download their app and use their service. Being able to assure their customers that there was no possible way for their personal data to be retrieved was their USP, but they were very aware of the moral dilemma of providing privacy for both criminal and legitimate interests.

> *'That's what judges decide. And as a small technology organisation, we're just not in a position to be able to police that.' (Respondent #10, 2021)*

In their opinion the ethical dilemmas of protecting privacy versus access to data for the wider social good (like catching criminals or protecting national security) is still not resolved in any satisfactory way, but that just the fact that data exists means that it is not, and can never be, secure or private.

> *'It makes me laugh, because you can be speaking to the same, almost the same people, sometimes they want their communications to be secure and protected from other state entities or other bad guys. And yet they want backdoor or the ability to access messages almost built into the system. You cannot have this piece of cake and eat it as well. If you build systems you can get access to, other people have access to your systems too.' (Respondent #10, 2021)*

6.2.2 Responses to Examples B and C

Checklists generally were perceived as useful in principle, not likely to be used in practice, either because existing practices were sufficient and/or the resource overhead would be to high.

Respondents tended to treat these two examples as similar in that they covered very similar topics although structured differently. Comments across the two examples are presented together as they often overlapped in responses. Generally, the EU checklist (Example B) was seen as a useful tool for thinking about all the issues and creating a debate at the beginning of projects, with the IAF template (Example C) being seen as more structured which appealed to some respondents. Several respondents, especially those from the smaller companies, commented that the 'overhead' in terms of time and resource would be too high for them to implement these types of processes. The respondents with senior management experience gained in roles in large corporate environments understood the rationale and structure of such documents and how these might fit into a product development process. The start-up respondents (who were also all hands-on developers) were much less familiar with this kind of approach to documenting processes.

> *'This is just coming from my experience. Wherever I have worked, I would work with probably one or two people working for me, because I worked in small teams as a data scientist. So it was a more hands on approach, rather than, like processes and stuff. So I don't have much experience with this kind of stuff. At the moment, the company is too small, I think, to build any kind of framework, because, you know, the overhead would be too big in terms of like, what do you get? Probably, we get to like, you know, 100 developers, maybe but not at the moment.' (Respondent #8, 2021)*

6.2.2.1 Data risk perception in different sectors

Products with higher risk for privacy and security had enhanced assessment already built into production process. Lower risk products considered not to need detailed ethics assessments.

The responses also differed across the set of respondents depending on the nature of their product and the regulatory environment they operated within. The companies operating in spaces where data security was paramount (biometric data, patient data, secure messaging) felt they addressed many of the issues highlighted in the examples as part of their development processes already.

> *'If I look at this [Example B], these are all absolutely built into all of these things are things that we would develop with in conjunction with our customer.' (Respondent #1, 2020)*

The start-up using patient data felt that the clinical safety regulation their product operated under covered all the necessary processes already.

> *'Its asking potentially useful questions. But it's part of everything that we have to do, anyway, as part of our clinical safety. This is part of what is our responsibility. So it seems like more paperwork for the sake of paperwork for us, but that might not be the case for others.' (Respondent #7, 2021)*

Another respondent felt that the level of attention and resource given to assessing and mitigating ethical impacts and privacy issues was dependant on the type or data you were working with.

> *'I think it depends what kind of data you work with it is personal sensitive data, I think everyone should focus on it, as you just don't know. And it can cost you a lot. If it has no sensitive data, like, you know, for example, in our case, it's not that important. It's more about, you know, working with the legal team, in order to understand the legal consequences of something.' (Respondent #9, 2021)*

They also noted though, in a previous role in global derivatives marketplace which dealt with highly valuable and sensitive financial data, that despite there being strict legal process for working on the data

> *'even then, they weren't followed… some of it, which was really sensitive, they were very strict about it. They had legal teams, which will tell you what you can do and what you can't. And basically, if they didn't tell you [explicitly] that you can't do something, you're allowed to do it. You know, it's just too much stuff.' (Respondent #8, 2021)*

Respondents who felt their data was inherently low risk (GIS mapping for example) thought that both Example A and B would be overkill for their application and would present serious challenges to resources for a small company.

> *'The intention of that is brilliant [Example B]. But that would be so onerous process to go through every time you start a project or start using a different data sets and find a thing you know, it brings to mind a little bit the Data Ethics Canvas [a framework from the ODI], which I think is great, but if you start going through that before every project, you'll never get your project started.' (Respondent #5, 2021)*

6.2.2.2 Metrics and tick-box mentality in Example B

Checklists run the risk of becoming a form of 'ethicswashing'. Translation of ethical principles into metrics allows for superficial governance.

One respondent, an experienced senior manager, expressed concerns that checklists in Example A or B would be converted into metrics and used as a form of ethicswashing.

> *'It always worries me when people take things like this, and they use it as a criteria. Somebody turns it into a metric. Yeah, we're 77% compliant, which is 10%, higher than the industry norm of our levels. So therefore, we can just sort of shut up and not think about it anymore.'*

They also thought that Example B would work effectively as a one-off process to uncover where further work was needed, but not as an ongoing process.

> *'So if you've got a team that's serious about wanting to do this. Then you take this, you drop it in and you go right, read through this, how much of this do you genuinely believe that you would comply with? And that gives you an assessment of where you need to do some work, then it is of use. But once you've done that, it ceases to have any use, you can play this card once. Otherwise it becomes a tick box and somebody will turn it into a metric.' (Respondent #4, 2021)*

Note also how the respondent refers to the necessity of pre-existing commitment on behalf of the development team. The tools in themselves do not engender this commitment, and without it the respondent thought the real impact of their use would be superficial and performative.

6.2.2.3 Business ethics and trust

Strong ethical behaviours within a company viewed as beneficial for customer trust and the basis for successful business. Recognition not all actors adhere to this as personal ethics also important.

Another respondent reflected on the outcomes of using tools like Example A and B. They described a positive feedback loop between behaving in a reliable and trustworthy manner within a company that then builds trust from customers and results in a positive impact on sales. They believed this was the sort of approach their own company followed but recognised that not all actors would do so.

Ethical behaviour was couched in terms of the beneficial outcomes for the actor and their organisation. Doing business in an honourable way was, for this respondent, an important personal value, but they also viewed it as an efficient way to conduct a successful business.

> *'I'm a big fan of checklists and return standards, but again, it comes back down to that behavioural thing. Why would I do it? You can't just have a checklist unless you're following the steps and behaving in the right way that checklist is motivating you to do. Then you get the feedback loop in terms of a positive response from your customers in that they buy more because they trust you.*

> *They tell the story that says yes, I'm doing the right thing. Otherwise, why would you bother?' (Respondent #1, 2020)*

They were also keenly aware of the differing motivations of competitor companies in his own market.

> *'There will always be organisations that go "screw standards, I just want to make a quick buck." Do a bit of fraud, get rich quick and then bugger off.'*
>
> *'I think there are weak spots around the system. And there are players there who are definitely not thinking ethically, but thinking about making loads of money, buying big yachts and private jets. So yeah, they think differently to us.' (Respondent #1, 2020)*

6.2.2.4 External audit, assurance, and legal compliance

Weakness in ethics assessments that did not have any form of 3rd party assurance or oversight by regulator.

All the participants discussed data protection regulation and understood that their handling of personal data was important in order to comply with the law. Key aspects of their product R&D directly addressed the principles of GDPR like the conditions under which data was collected and consent, security of personal data, data minimisation and anonymisation.

The participants with long careers and experience in large corporate environments all discussed aspects of enforcement, believing it was important for regulators to enforce rules and fine companies.

> *'I tend to think you probably don't want to create another regulator [to regulate AI]. But you do want to strengthen the existing regulators. And so you want to make it really clear to everybody that this is serious, and that the penalties for failure are huge. Now, they should be really, really high.' (Respondent #1, 2020)*

Another experienced respondent discussed the issue of the voluntary nature of assessments and audits like those in Example B and C and how this was a flawed approach without external audit and assurance.

> *'So then you have your internal audit, audit plan, but then unless you have external influence coming in there, whether that be auditing or penetration testing, you can't be sure [of standards] because it's about everybody marking their own homework. Yeah, and the big auditing firms where they say, "I've audited you against this standard." We'll put it this way, we all know the vulnerabilities of the large audit firm model, which is completely different to somebody externally looking and assessing whether it be that code, that practice, that system, that back end, it needs somebody outside of the*

organisation [to audit], then you really start seeing results.' (Respondent #6, 2021)

6.2.2.5 Real world constraints – time, budget, resource

Constraints of production timelines, budget, and staff perceived as limiting factors for applying assessments.

The length and complexity of Example A and B led all the respondents to comment on how these sorts of processes require time, budget and staff resources which small companies and especially start-ups do not have and cannot afford to buy in.

'We've got six employees and only four are full time. I just think we all sort of work absolutely maximum capacity all the time. And I think if someone was taken off to do this it would be quite difficult for us all to claw that time back.' (Respondent #5, 2021)

One respondent reflected on the pressure to deliver on customer deadlines, and thought that having a process in place like Example C might prevent the problem of risks being identified late in the product development lifecycle. The budget and delivery pressures in later stages of development mean a greater likelihood of pushing products through despite weaknesses in design or potential negative impacts.

'It's in the last days, and you're up against the budget. And you've got to get the thing out. That's where the pressure is. And you haven't got time. But that's when you want to step back for a little bit and say, Are we still on track with this? Because you all start with the best intentions but you ain't got the money to hold on, we're not going to meet this deadline and then we won't get paid. That's the real pressure. So that's why actually keeping on track with this as you go through is good, so you don't get to that? Because that is a really difficult decision at that point. So who's going to tell the customer that we don't have the release?' (Respondent #1, 2020)

6.2.2.6 Challenges for start-ups

Start-ups especially aware of the constraints they face with resource. New founders often lack experience and knowledge.

The smaller the company, the smaller the resource of skills and knowledge they have to draw on to meet the ethical and privacy challenges of their products, for early-stage start-ups even compliance with data protection regulation is a tough ask. One respondent was keenly aware of how tough it

had been for a two person start-up to get up to speed with compliance with data protection regulation, let alone go further and consider a wider set of ethical issues around their product.

> *'I think giving clear pipelines, particularly for small companies where they don't have an information and data governance officer or they don't have a legal department, they haven't got all those things.' (Respondent #9, 2021)*

They also thought, when looking at Example A and B, that although the checklists were useful for thinking through potential issues, they would like a more formalised pipeline to work through where they could feel confident by the end that they were compliant. They made specific reference to how difficult they found it understand data protection, and that although they found resources on the ICO website useful, they still found it all confusing and were not sure if they had met all the requirements to comply with data protection regulation. For them this was the most important consideration.

> *'I think this checklist sort of system [Example B] is great. And it obviously list all the possible things you could think of. But it'd be nice if this evolved into something more structured but it'd be nice if you almost had an account with, for example, the ICO, you had a way of knowing if you've gone through things and done them correctly, and it's all associated. And then if an update does happen, you know, are you need to also do this new thing now to comply. And it's glaringly obvious when you're missing something, whereas at the moment, I think it's not very clear if you're compliant or not, with the regulations, and there's so many kind of different things you need to consider. You can't just put a tick on it.' (Respondent #9, 2021)*

Most of the participants remarked on the more structured approach of Example C, compared to Example B and thought that it was a better, more workable approach than the more open checklist approach. Example C allowed for it to translate into a management process to track and enact governance, especially for less experienced founders.

> *'This structure is good. Obviously, this one looks a little bit daunting to look at to start with. But yeah, I think having structure and having all these points that you almost just have to complete is much more useful than just having a free form. Having a structured assessment, it's definitely much more useful to know what to do.' (Respondent #9, 2021)*

6.2.2.7 Identifying unintended consequences

Considering wider or unintended consequences often beyond the scope or focus of companies. Focus on product, production and specific market excluded wider reflection.

When looking in detail at some of the questions in Example B and C all participants reflected on how hard it is for companies to think about anything beyond their immediate focus of developing and selling products.

> *'I think the questions I found that was interesting, one of the things I don't think start-ups think of is the unintended consequences. They're so focused on the way that they see it from their own perspective, they've lived it, they focused on it, this is their bubble. That's what they can think about. And sometimes I think it's very difficult for them to snap out of that way of thinking or seeing it from an external perspective.' (Respondent #10, 2021)*

One respondent in the biometric market, who was overall quite dismissive of the need to use structured ethics assessments like Example B and C, did discuss one of the unintended consequences of biometric authentication for fintech. This was the potential for exclusion of some groups by the move to a cashless society where all payments are managed through a phone or card.

> *'There are, you know, parts of society, for whatever reasons, are excluded from the banking world in a card and cashless world. And technologies like ours, are accelerating that cashless society. And it's only when somebody actually sat me down and said, 'I am actively doing cash to make sure that the vulnerable in society are catered for' that it dawned on me. But there you have an example where we don't quite see the ethical implications of what we're doing. Because it's not necessarily obvious, I can talk all about privacy, that I understand the GDPR and stuff, but I didn't quite think of truly that if I supply technology to society that then is used everywhere, that somehow somebody is excluded because of their circumstances. So you don't always see the impact.' (Respondent #3, 2020)*

6.2.3 Example D Consequence Scanning

Workshop approach appreciated by developers in preference to a more formal process. Considered useful at beginning of projects.

The participants who were developers themselves expressed a positive response to the agile workshop approach of Example D.

> *'I like that. I think the workshop style of doing things is, I feel that's a really good approach, because it's not, you're not putting just the pressure or the responsibility on one person, you're asking your whole sort of team to think about it and look at it and take responsibility for it.' (Respondent #5, 2021)*

One respondent working with geospatial data said they had used the ODI's Data Ethics Canvas (ODI, 2018) as part of their business development process as an afternoon workshop for all the staff. They were not using it against any specific project but as a means of thinking about how different

staff members on both the technical and business side would approach use cases. Another respondent thought that Example D was similar to the process they would use during product development but with a focus on privacy concerns rather than wider ethical considerations.

> *'So we do tend to whenever we're kind of thinking about or, or coming up with a concept for a new feature, we do ask ourselves, how does this benefit the user? What kind of information are they having to hand over? Have we asked them for any more information than we really need. And then is that information justified for the benefit that we're giving? So it's kind of more focused on just the privacy aspect and how that benefit comes to the user.' (Respondent #9, 2021)*

The respondent did express concerns that continually reflecting on these issues could slow down development, and that it was not necessary for everyone to be engaged in group decision-making on every feature or change.

> *'If you have all these different things that you've then got to deal with at the start of each sprint, or each project, it does kind of slow you down. And obviously as a start-up, and you're trying to be lean, so that you can move at speed, that sort of thing out. So, it might be that this is a bit too much for every kind of iteration or every change. And instead, there should be a kind of a responsibility chain, I suppose of the person who's come up with a feature idea or that the manager of that team needs to do that. The compliance of it and make sure that that works. And you have maybe some oversight, someone independent who oversees and agrees with that, but not having everyone every time having to have a whole discussion about it, because it would probably slow things down.' (Respondent #9, 2021)*

6.3 Processes deployed to manage ethical issues

Confidence in own processes in mitigating ethical risks. Focus on privacy, data security and compliance with regulation relevant to application. Products working with high risk public sector or medical application had well-developed processes for addressing ethical risks.

In response to reviewing the example tools, respondents also discussed the processes in their own companies. All the respondents felt that they addressed ethical issues directly related to their products but did not necessarily use a separate formal process for this and, with one exception (see 6.3.1), did not have any processes or documentation in place like those shown them in the examples. Most of the respondents felt that their product development processes did capture potential issues, with these mostly focused on privacy and data security issues, and compliance with relevant regulation (e.g. financial services regulation, or clinical safety), and was legal (e.g. met the license conditions for data sets.)

6.3.1 Ethical Impact Assessment development – public sector data

High risk public sector data project developed Ethical Impact Assessment process integrated with data science production process.

One respondent recounted his experience at a research focused organisation where they developed over time their own Ethical Impact Assessments for their data science projects. The respondent recognised Example C as having close correlation to their own practice. They developed a two-tier process for ethical assessments which comprise a set of templates for recording the research process, evolved from both DPIAs and research ethics frameworks.

They described needing an agile approach to development of algorithms and models that provided actionable solutions in the software development process. Their projects were primarily for the government and non-profit sector addressing issues with very sensitive data and complex relationships between the procuring and funding agencies. At the start of their journey in this space it was felt there was too much listing of a wide variety of concerns, but what was needed was to be able to transform those concerns into an actionable process for the developer team.

> *'In terms of the risk assessments, and the DPIA's and the EIAs, we actually do those in quite a rigid way. In every project we ever do, it's our first step, and we have templates to be filled out… We have two different kinds of processes, one is kind of an assessment of research ethics, that we will kind of conduct our gathering information and analysing information in our research ethically. And then there's another process that sort of says what we will build will be ethical, and they are assessed in two parts. So one is a kind of assessment of our own personal conduct to our project. And the other one is an assessment, the product that gets produced, I think, is largely the difference.' (Respondent #2, 2020)*

Using the Cross Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000) as the model for their work process, they then injected the data protection and ethical considerations into the pipeline.

> *'So at each step, we would say, right, business case is this okay? Are there any ethical concerns before we begin? Okay, fine. Now, let's get we've got this data now. Okay. Are there any ethical concerns that come around having identified we're gonna use this data, no, carry on and so we tried to kind of embed elements of ethics into the machine learning process.' (Respondent #2, 2020)*

The respondent also described how they would undertake extensive participation exercises with domain experts and target users for projects, conducting interviews and group mapping exercises to establish concerns and identify problematic areas.

> *'So we didn't want to assume that we can envision every possible bad outcome ourselves, so we knew about algorithms but we didn't have, by any means, the expertise of [staff in statutory services] in real life in this situation.' (Respondent #2, 2020)*

Frequently concerns raised by stakeholders during the participation exercises were outside the remit of the data science team (for example worries about inter-agency information sharing and its consequences). Often these concerns and ethical problems were not part of the tool under development but were part of the wider system and processes the tool would be embedded in, and therefore lay outside of the developers ability or responsibility to solve.

The feedback from users and domain experts was important for delineating accountability – what was the responsibility of the designers, and what was the responsibility of the users of the system. Part of the ethical concern, and an important aspect to get right for the respondent, was ensuring that users understood what the tool could and could not do, and where the responsibility lay for actions taken from the insights it produced.

> *'For us, the challenge was defining what is what is our responsibility in what is the responsibility of the person using the tool. So we have responsibilities as the builder of the tool. And the people using the tool have their own. Communication of limitations and results [of the algorithm] was a primary concern.' (Respondent #2, 2020)*

The process of thinking through ethical concerns for a project surfaced a raft of concerns but they recognised the need for a filter process, to separate out where the developer versus the end user responsibility lay.

> *'We cast a wide net, sticky notes and end user sessions, concerns, interviews, then that had to be filtered down to where we transform those into user stories and feature requests.' (Respondent #2, 2020)*

An ethical concern would be translated into

> *'an actionable user story which can appear on our on a development board and up on the backlog. So it was that funnelling process from casting the wide net to getting everyone's concerns and then thinking about which ones of those needed to make their way into, into the Agile process. So that's how we ended up.' (Respondent #2, 2020)*

This was the result of a long process of iteration for the data science team, part of which was the realisation that they could surface a variety of ethical concerns which lay outside of the responsibility of the developers, and actually lay with the nature and operation of the statutory agencies themselves, and their relationships to each other.

> *'We were always trying to kind of engage with lots of ethical discussions all the time. And then kind of coming to the conclusion, well this is kind of beyond our*

control, we're kind of talking about the ethics of [statutory government] services themselves which is not within the scope of this project of rewrite.' (Respondent #2, 2020)

6.3.2 Agile process for medical application

To meet clinical safety requirements devised ethical checks throughout agile production process.

The respondent developing a system using patient records thought Example D was the closest to the kind of process they deploy in their development pipeline, while the documentation for this process was similar to Example C but with a focus on clinical safety. This includes consulting users (note that patients were not mentioned, only GPs), and documenting design decisions and features on a clinical risk register.

'The way we do it is we keep things agile. So every feature that we do, it starts from user, so we speak to the users. In our case, that's currently GPs. And from there, we iterate with designs, with rapid feedback with them. Once they're happy that it solves the problem that they're facing, then we start development. But at the same time we let my co-founder know so we have a planning meeting to say, "Hey, this is this is the problem, this is how we're solving it. What are the clinical implications of this? How can we ensure clinical safety?" And then he says some stuff, which is really useful. And then then we go ahead and develop it. And then once it's developed we go through the feature with him, and take him through that journey. And we asked the same question, what are the clinical risks? Is this clinically safe, and then make sure that he's happy with it… We document everything in our clinical risk register. And on that, we get details all of our features, where there were risks, the way we handled it, the way we're thinking of improving it, and then that that cycle, we do that on a sprint basis.' (Respondent #7, 2021)

6.3.3 Data audit procedure – GIS projects

Developed internal process and new staff responsibility for assessing data sets at beginning of projects.

The respondent in the working with geospatial data described the role of their Operations Manager who brought in a data audit procedure for new projects to identify data for new projects. This checklist process covers sourcing data, coverage, comprehensiveness, quality and accuracy, provenance including the supplier, how data was produced and any licencing and restrictions on use.

'I think that's really kind of helped us start to understand not just our own data, but how our own data sits with other data sets and how data that we bring in needs to be treated.' (Respondent #5, 2021)

6.4 Regulation, standards, and compliance

Overall positive attitude to regulation but should be applied fairly to all actors. Adopting standards should be rewarded.

In the last part of the interview respondents were asked their opinion of regulation as it currently applied to them. Most respondents in more established SME's were confident that they were compliant with existing regulations in the jurisdictions they sold into. For most respondents their key focus was on data protection regulations. One respondent had strong views on the need for regulators and courts to be more proactive in enforcing the law to protect citizens.

'The regulator and the judiciary and the recourse to the law are failing. You know, GDPR is a massive, massive failure.' (Respondent #6, 2021)

Another, when asked if there should be separate regulation for AI also indicated the need for effective regulation with commensurate penalties for wrongdoing.

'I tend to think you probably don't want to create another regulator. But you do want to strengthen the existing regulators. And so you want to make it really clear to everybody that this is serious, and that the penalties for failure are huge. Now, they should be really, really high.' (Respondent #1, 2020)

For companies selling into international markets the feeling was that they would adhere to the highest regulatory standards as this was the most cost-effective and efficient approach. The current suggestions from the UK government for amending the UK GDPR set out in the National Data Strategy (DCMS, 2020) propose paring back current privacy protections and changing the role of the ICO to better serve economic and social goals. Several respondents described working to the regulation in their biggest markets or the jurisdiction with the highest standards (GDPR).

'From what I can see, we will definitely establish our own set of rules. I suspect, any large player, like a bank, which is trading in Europe, assuming that they continue to trade in Europe will have to be GDPR compliant. And I think, I think to be honest, the very best players will continue to comply to the highest standards.'

'Companies can't even build three different sets of products now, you know, complying to different standards, they build them to the highest standard and then they can export them anywhere. And that will be the same I think with data and but to me, it should be about the penalties when people make the wrong judgments.' (Respondent #1, 2020)

Another respondent who worked on smart city projects in the intersection between government policy and private enterprise was keen to promote technical standards such as ISO/BSI. Adopting appropriate technical standards should then lead to greater investment, alongside reducing the risk of such projects.

> *'I'm trying to lobby the government through the BSI. I've also been supporting them consulting with and advising cities that you should be asking for government funding to make yourself standards compliant. And because you've done that you should get greater access to bigger pots of cash because you will have more chance of successful less risky delivery. In terms of cybersecurity, data leakage, making use of that data with AI, wherever you're getting the data from in terms of sensors, be it from the internet of things or from people themselves, you can be trusted because you've got the right processes.' (Respondent #6, 2021)*

6.4.1 Regulatory environment – secure messaging

Encrypted messaging market constantly reviewing regulatory landscape. Not clear consensus across jurisdictions on the encryption and secure messaging. Tensions between protecting privacy and protecting criminal activity.

The respondent in the secure messaging market was very aware of having to horizon scan constantly in the regulatory space to keep up with changes to rules in different jurisdictions. The company itself was registered in Switzerland to take advantage of strong Swiss protections on the right to privacy (motivated by the secrecy of the Swiss banking sector).

> *'It feels with the compliance and the regulations, sometimes it feels like you're, you're standing on a piece of ground, and slowly more and more of that ground gets kind of cut away. And you have to maintain quite close monitoring on what the regulations are in different countries and how things change. Because operating in the position we do there is a real fine line between regulation that's trying to, on the surface of it, protect the world from the bad guys, and then that erosion of individual's privacy and that general view that people have the right to privacy. And yes, it's a fine line. People want to know how secure and private the system is, but they also want to know what you're doing to stop bad people going on it. You have to explain to people that you can't have it both ways. You need to make organisations decide where they stand and be able to justify and fight for that position.' (Respondent #10, 2021)*

6.4.2 Working with regulations – start-ups

Start-ups face challenges of knowledge, experience, time, and resource to achieve compliance with existing regulation.

Start-ups face particular challenges when working with regulation and achieving compliance as they often lack knowledge and experience, and additionally do not have the financial or staff resources to draw on for expertise in compliance issues. The start-ups interviewed with young founders faced a steep learning curve in this space and felt there needed to be clearer guidance and practical tools available to work through compliance processes.

6.4.3 Medical sector

Start-ups in heavily regulated sectors like medical applications struggled with the approval process and the lack of clear guidance. Suggest an agile toolkit for developers.

Actors in heavily regulated sectors like finance and health must ensure their products meet the standards for technology in their sector. The health sector and any technology that uses patient data and/or is a medical device must conform to clinical safety regulation and needs approval before a product can go to market. As one respondent recounts, this can be an onerous task.

> *'There are specific laws with medical data. But also there is there's more guidance or rules set by NHS digital, and the Health Record system that we're partnering with.'*
>
> *'We've had to undergo almost a year of approval, because we have to prove that we're processing patient data for the explicit needs that require it, we have to show that we are clinically safe in the way that we're handling the data. We have to show that we have the appropriate security approaches involved as well. So we have to do pen tests, we have to get certain certifications. And so there is a lengthy process to be able to get to that point. So part of that is GDPR kind of based, but there's also additional, because it's medical records, you've got an additional set of regulation for that.' (Respondent #7, 2021)*

The respondent described how difficult it was to deal with the process of applying to NHS Digital, and how unclear the process and pathways to approval were to navigate. They also described how assessors did not necessarily have the technical knowledge to be able to understand the products and applications they were tasked with approving which the respondent identified as a risk.

> *'The only way we were able to navigate [the NHS Digital approval process] is because we were fortunate enough to be introduced to the CEO of another health tech company, who is familiar with that process. To navigate through the system was a huge piece of work. And the other issue, is that there are a lot of people who are not technical. And it's very hard to speak to someone technical. So the people who are making the decisions typically make decisions without someone technical looking at it or speaking to us. Which in my view, is more of a risk than what the company are going to do.' (Respondent #7, 2021)*

They also wanted better guidance from NHS Digital on how to embed clinical safety in their development process, perhaps in the form of an agile toolkit encompassing the clinical safety and patient data constraints.

> *'I think if NHS digital were to create an agile toolkit that they would recommend to all their partners that might help some of their partners. The way we figured out how we wanted to do ours was based off their rough, confusing guidelines. And also think we managed to speak to someone who was working on the NHS app. And it sounded like they were on the right sort of trail with that, so then we just combined that with how we normally do agile, just kind of fit in or make clinical safety a big part of it.' (Respondent #7, 2021)*

6.4.4 Retail sector

Start-up in retail sector struggled with UK GDPR compliance. Suggest checklist process that would be easy for small companies to understand if they have achieved compliance.

The main ethical concern for the start-up with a product recommender tool for retail platforms was complying with data protection regulation. As a start-up with young founders this was challenging to navigate.

> *'So, it's a challenge, to be sure, and we've had to do a lot of like, research and understanding, especially around GDPR, and that sort of thing? To make sure that everything that we do is kind of compliant and making sure that people are aware of how we process the data and how we work with the information they're giving us.' (Respondent #9, 2021)*

When asked if they had documentation of their application, they thought that the squeeze on staff and resources meant they only had capacity to meet their compliance obligations for data protection regulation and not to go beyond what was strictly necessary. Small companies do not have the resource to hire in expertise for aspects of the business, like data protection or assessing impacts of projects.

> *'The difference here, like you said, between having a dedicated member of staff or a consultant, their whole job is to make sure you comply with that. And, and it's the difference when you're a start-up, and you've got one person sometimes who's got to deal with all of it, as well as all the other stuff that their company needs.' (Respondent #9, 2021)*

The respondent, who had taken his company through two accelerators, thought that not enough training and attention was paid to training and support for aspects of the business like GDPR compliance, or considering ethical impacts.

> *'I think a lot of start-ups definitely underestimated [compliance and regulation], and there's not very much focus on it from the accelerators, or from investors or that sort of thing, because obviously it's not a money maker. It's just something you have to do to comply with the law. And it's not a particularly glamorous thing to have. You get someone who comes in and does a talk about it or something, but there's not usually a big focus on it. And there's no "You need to make sure you get this right, or this could happen."' (Respondent #9, 2021)*

The respondent also described the balancing act between investing time and resources in examining potential ethical and compliance risks. Small companies cannot afford, on the one hand to fall foul of regulations, or have the capacity to weather a negative impact on reputation.

> *'And then there's obviously the other end of the spectrum where you probably never get anything done. Because you've written all this documentation, and then you realise that your idea is not very good. And you've just wasted half a year or something, being the most perfectly legal company. And I think it's just the support, I suppose, that needs to be there to help start-ups and smaller companies have the confidence that they can just have a template, a very comprehensive template that is that just covers everything they need.' (Respondent #9, 2021)*

The respondent also described how current guidance for data protection was confusing to understand for those new to the subject (like young founders) and left them feeling slightly uneasy whether they had achieved a suitable level of compliance with the law.

> *'I don't ever feel 100% confident and just using a random template that they've got on there [the ICO website]. And I never thought that was a clear kind of flow of you need to do this, this, this and this, and then you've got everything covered, I had to go through all the different pages, and then there'll be another thing and you need to also do this checklist to make sure you've done another kind of assessment over here. It's always a bit confusing. Something that has a kind of nice checklist flow going through, potentially, if you have more of that sort of thing from the ICO, it might be useful.' (Respondent #9, 2021)*

It should be noted that throughout this respondent returned to the topic of data protection and did not really see how there any other ethical issues or challenges to be addressed in the company or product. When asked, the respondent was not aware of how the underlying natural language

processing model they used was created, or how bias might be embedded. They considered their application to be low risk as they did not store any personal data on customers using the application, only high-level analytics for use by the retail platforms the application was embedded in.

6.5 Innovation and regulation

Overall positive response to regulation and that did not stifle innovation. Start-ups keen for better guidance.

In the final section of the interviews, respondents were asked about the relationship they saw, if any, between the freedom to innovate, and the effect of regulation. Most respondents did not think the current regulatory landscape was a barrier to innovation, particularly the older participants with lengthy experience across their industry and from within larger organisations.

> *'I don't see that kind of thing, you look at the amount of kind of disruptive tech start-ups that there are out there that are really doing well, and that are getting their seed funding going on to A rounds and B rounds and really making a difference. They wouldn't get to that stage if they were being stifled in any way, shape, or form.' (Respondent #5, 2021)*

One respondent thought regulation that, for example, restricted the use of certain data because of privacy concerns was actually a driver for innovation.

> *'I think the other thing is, if you're innovating, and you can't be successful, because you can't get your hands on a specific piece of data, I think you're coming up short anyway. The good innovators will look at it and go, right, we got all this data here. But we haven't got these pieces of data, actually, what can we infer from what we've got, that will start to give us some of the answers that we don't have at the moment, and they'll start to fill the gaps themselves? Yeah, you know, coming up with a new method of inferring something from another data set or putting two datasets together to pull out a different insight. So I think, to say that it was stifling innovation, you could almost argue it's the opposite. It's making people do things to fill the gaps that they perceive to exist at the moment.' (Respondent #5, 2021)*

For another respondent it was important not to consider 'regulation for regulation's sake' but to build a culture within a company where the right set of behaviours and controls were exhibited, and risk was properly managed. They refer to the safety and governance procedures in the airline and rail industry to illustrate how standards and regulation mitigate the risks of inherently dangerous technologies.

> *'The reason why we have less deaths now than 50 years ago on the railways and in the airline industry is because of the culture that is pervasive, that's no blame, that seeks to continuously improve and also provides methods that people can whistleblow without fear or favour. People are held accountable as part of their fiduciary requirements to ensure that if they are in a job they will go to jail if they cock it up. Bad shit will happen especially if you're pushing the edge of stuff. You know, it's to be expected that's the risk. So it's all about how do I mitigate that risk?' (Respondent #6, 2021)*

Regulation makes for better, more reliable products, and innovation a safer bet from a business risk point of view, even though it incurs greater costs to production and sales. One of the risks the one respondent recounts is not only the dangers of producing damaging products, but also causing societal and customer pushback.

> *'Our industry has just been in thrall to the new, that's because the people that create it need to sell it to recover the investment and they have a notional value on it's worth. But that value is really hard to calculate, especially in the world of AI. Because you're treading on someone's toes. So the question is, how do you do it without making them feel uncomfortable? You've got to bring people along. And that takes effort. And that means it's time and it means costs. So that slows down innovation. So it's greed, that slows down innovation. It's not regulation, regulation makes innovation safer. A safer gamble, surely.' (Respondent #6, 2021)*

The younger founders were less positive about regulation in general but did highlight the need for guidance to support companies who do not have the expertise or resources to define the appropriate ethical policies and governance practices themselves.

> *'I think there's a lack of knowledge that's out there. Concerning this area, and especially with regard to start-ups and with everything else that needs to be done [in a start-up], but providing direction, because these are such big questions that need answering and areas that need that need focusing on where there's not a lot of knowledge out there. And I would say that not a lot of people or companies think about it early on in the process. And that it does need some regulation or guidance or something to comply to, to provide the start-ups with the direction that they need. Because they haven't really gotten the time to be defining what the policy or the procedures should be. So the guidance for them for that and taking that decision out of their hands, then I think makes that process a lot easier.'* (Respondent #10, 2021)

> *'I think that having impact assessments by the EU, and that sort of thing, is probably a good idea. But then it does, sometimes stifle the innovation and that sort of thing, because people struggle to comply with that sort of thing. And there's just not great guidance.' (Respondent #9, 2021)*

6.6 Key themes from interviews

6.6.1 Privacy and data protection emerged as the central concern across all the respondents

Participants understood as 'ethics' in the context of their products through the lens of data protection – privacy and security of data. Ethical behaviour was understood to equate to compliance with data protection regulations and ensuring their products could be trusted by their customers (security, robustness). This is also shown in a recent survey by Morley et al. (2021, p. 3) which similarly found the understanding of ethical design to be primarily understood through the lens of data protection principles.

Design decisions of products were often heavily influenced by data protection and cybersecurity risks. Respondents described how they tried to avoid collecting or processing personal data wherever possible. This could be conceived as positive, but for example in the biometrics applications the main purpose was to ensure that it was the customer for their systems who bore the responsibility for the data, not the builder of the system. Design was therefore focused on shifting accountability for risks from the vendor to the customer. In other examples, the design of the product was intended to push back against previous data practices, for example in the retail application, was intended to move away from the ubiquitous data collection and tracking of the conventional ad-tech system.

6.6.2 Confidence expressed in existing process

Respondents all expressed confidence in their own processes in mitigating ethical risks as they perceived them. Products with using public sector or medical data had the most well-developed processes for addressing ethical risks, driven by the sensitive nature of the personal data they were processing and by specific regulation in their sector.

6.6.3 Desire for better guidance

Start-ups in heavily regulated sectors like medical applications struggled with the relevant clinical approval processes pointed to the lack of clear guidance. They would appreciate more structured processes like an agile toolkit for developers. The start-up in the retail sector struggled with UK GDPR compliance. They suggested checklist process that would be easy for small companies to follow which would reassure them that they had achieved compliance.

6.6.4 Business and reputational risk was a concern for vendors

Business and reputational risk was a concern for vendors when considering how their customers might use their system. This took different forms. Firstly, to ensure their product was not responsible for privacy and security breaches (privacy and security by design). Secondly, to consider the nature of the customer and the uses to which their product, insights and data might be put. For some use cases this led the vendors to be very circumspect about who they would sell to, but in some cases to acknowledge they could not control this this. Another approach was to limit the inclusion of personal data in a product, or leaving the responsibility for data processed, collected, or stored by a system to sit with the customer. For some vendors despite acknowledging the risks of misuse of their product, the inability to complete effective due diligence on customers resulted in choosing not to conduct any form of vetting.

6.6.5 Public sector actors concerned with perception in use of citizen data

Public sector purchasers of AI systems put the privacy and security of citizen data at the forefront. Value for money and possible political damage from ill-conceived projects was also a key consideration. Standard local government procurement procedures were understood by the respondents in this space to be a bare minimum and need further processes to understand fully the nature and desirability of a private sector technology partner.

6.6.6 Perception of wider ethical principles

Bias, discrimination, and exclusion was only mentioned by a few respondents and was not a core ethical consideration for the design of their products. Explainability was also not a core concern beyond the need to explain products to customers in such a way as to secure sales.

6.6.7 Perceptions of ethical statements as a tool

Public-facing statements of ethical principles (as in Example A) were unanimously viewed as a branding exercise, and potentially a form of empty virtue signalling. Public-facing ethics statements were generally perceived to function for marketing and branding purposes.

There was clear identification of the potential gap between public statements and internal company behaviour. All respondents felt their own company culture was ethical, and that company culture is grown from within and led by senior management. Strong ethical behaviours within a company were viewed as beneficial for customer trust and the basis for a successful business. There was recognition from respondents that there were actors in their sector who they considered to lack an

ethical approach, and actors who could abuse the affordances of their products. Respondents' personal ethics informed the way they built their products and ran their companies.

6.6.8 Perceptions of checklists as a tool

Checklists generally like Example B and C were perceived as useful in principle, but not likely to be used in practice, either because existing practices were considered sufficient and/or the resource overhead would be too high. The constraints of production timelines, budget, and staff were generally perceived as limiting factors for applying assessments. Start-ups especially aware of the constraints they face with resource including their lack of business experience and knowledge.

Concerns were expressed that checklists run the risk of becoming a form of 'ethics theatre', where the translation of ethical principles into metrics and checkboxes enables superficial governance practices especially if these are voluntary internal practice without any form of 3rd party assurance or oversight by a regulator.

Products judged to have higher risk for privacy and security had enhanced assessment already built into production process for example in the biometrics applications. Lower risk data use was considered not to need detailed ethics assessments. No one reported using DPIA's.

6.6.9 Perceptions of developer workshop materials as tools

Example D appealed to respondents who were developers, but only useful if they perceived a particular project had already apparent risks that would justify the time and resource. The respondent who had overseen the development of a full ethical impact assessment process was working with highly sensitive personal data which required an enhanced level of oversight. This wove ethical considerations, including participation processes and user studies, into their existing data science production process. Other respondents also thought that if they were going to use an ethical tool, one which fitted into their existing production process (agile) would be more helpful than the more paperwork heavy tools like Example B and C.

6.6.10 Lack of consideration for wider consequences

Considering the wider impacts or unintended consequences of products was considered beyond the scope or focus of companies. They were focused on their specific use cases and products, the production process and specific market they were selling into. Wider impacts (like, for example, the wider implications of ubiquitous use of biometrics for identity, or the nefarious use of encrypted messaging) was considered the remit of government and regulators not companies.

6.6.11 Overall positive response to regulation

Overall, the response was positive to regulation and it was not felt that it had a negative impact on innovation, in fact for some respondents regulation was actually a stimulus. Providing whatever rules were in place were fairly and effectively enforced the attitude from respondents was supportive. This mirrors the findings in a report from the Ada Lovelace Institute report 'Regulate to innovate'.

> 'Far from being an impediment to innovation, effective, future-proof regulation will provide companies and developers with the space to experiment and take risks without being hampered by concerns about legal, reputational or ethical exposure.' (Farmer, Strait and Parker, 2021, p. 7)

Start-ups, in particular, felt keenly the need for better guidance and tools to implement regulatory requirements and feel confident they were compliant. Respondents who were in complex ethical spaces (like biometrics or secure messaging) felt that society needed to make decisions and provide them with regulation which they could follow.

6.6.12 Missing ethical considerations

If we consider what was *not* discussed by the respondents during the interviews we can see how understanding on the ground departs from the ethical principles being proposed by a range of commentators in the AI ethics space. In Table 11 seven key ethical categories are listed. Respondents focused strongly on the principles of 'technical robustness and safety' and 'privacy and data governance', and to a lesser extent on the principles of 'human agency and oversight' and 'transparency'.

The principles that received limited attention were those of 'diversity, non-discrimination and fairness', very few respondents mentioned bias, with no discussion of inclusive design or the potential for exclusion in their products. Most importantly this principle also covers stakeholder participation. This was discussed by only one of the respondents when describing the design process for a tool for statutory agencies. Even then the definition of the stakeholders was still narrow. 'Users' were generally defined as the direct users or purchasers of the products, not any individuals or groups who may be affected by processes or decisions enacted by the systems. As noted in the analysis of AI ethics tools in Section 5.9, the inclusion of voices (Voiceless, Vested Interests and Users) from beyond the narrow scope of vendors and buyers of systems and their vested interests is also absent from proposed tools.

Respondents for the most part did not report on wider the principle of ‘societal wellbeing’, or if they did this was considered outside of their scope to address. No mention was made of ‘environmental wellbeing’ either or any reference to environmental impacts or sustainability of their products (despite some the respondents working on tools in this area).

The last set of principles in Table 11 come under the heading ‘accountability’. This covers ‘auditability’, which except for the respondent working in the medical sector, was not a consideration. There was no mention of internal process for the principle of ‘reporting negative impacts’ or any reference to internal process to encourage reporting by staff or to protect whistle-blowers. Respondents were, however, aware of the problems of the chains of accountability and liability across the supply chain and some reported struggling with how to effectively undertake risk mitigation.

Chapter 7 Discussion

Reviewing the landscape of tools in Chapter 5, there are three key areas where tools are being developed – impact assessment, audit and technical/design tools. As Figure 14 illustrates these approaches target different stages of AI system development and provide different outcomes. Ex ante impact assessments are used at the early stages of use case development, and for procurement processes. These provide a predictive decision-making tool for whether a proposed AI system should progress to development, be deployed or purchased and what are the possible impacts of its use. Ex post impact assessment is used as a post-deployment tool to capture the impacts of a system, often in comparison to a particular set of stakeholders, or issues like impact on human rights or democracy.

Audit tools showed an equal level of presence in this study to impact assessment which can be used for assurance of production and monitoring purposes. Audit processes traditionally follow well-defined systematic processes that require third party verification. There is some confusion in the current landscape between a technical intervention (often called an audit e.g. for fairness or bias), and what is more generally understood business practice of formal auditing (Carrier and Brown, 2021). In this study I have differentiated between those tools that more closely resemble other comparable audits, and categorised tools for specific aspects of the assessment of data training sets or models as technical tools – not audits. Technical tools do have an important role to play in addressing ethical issues in AI systems, but ultimately need to be part of a wider governance process. The documentation produced by these tools should form part of impact assessment and audit processes in order that all ethical aspects of a product can be captured (not just a focus on e.g. metrics for fairness (Lee and Singh, 2021)). In Figure 14 technical and design tools have been incorporated into the model as an input to the category of auditable artifacts which are necessary for evidence in both impact assessments and audits.

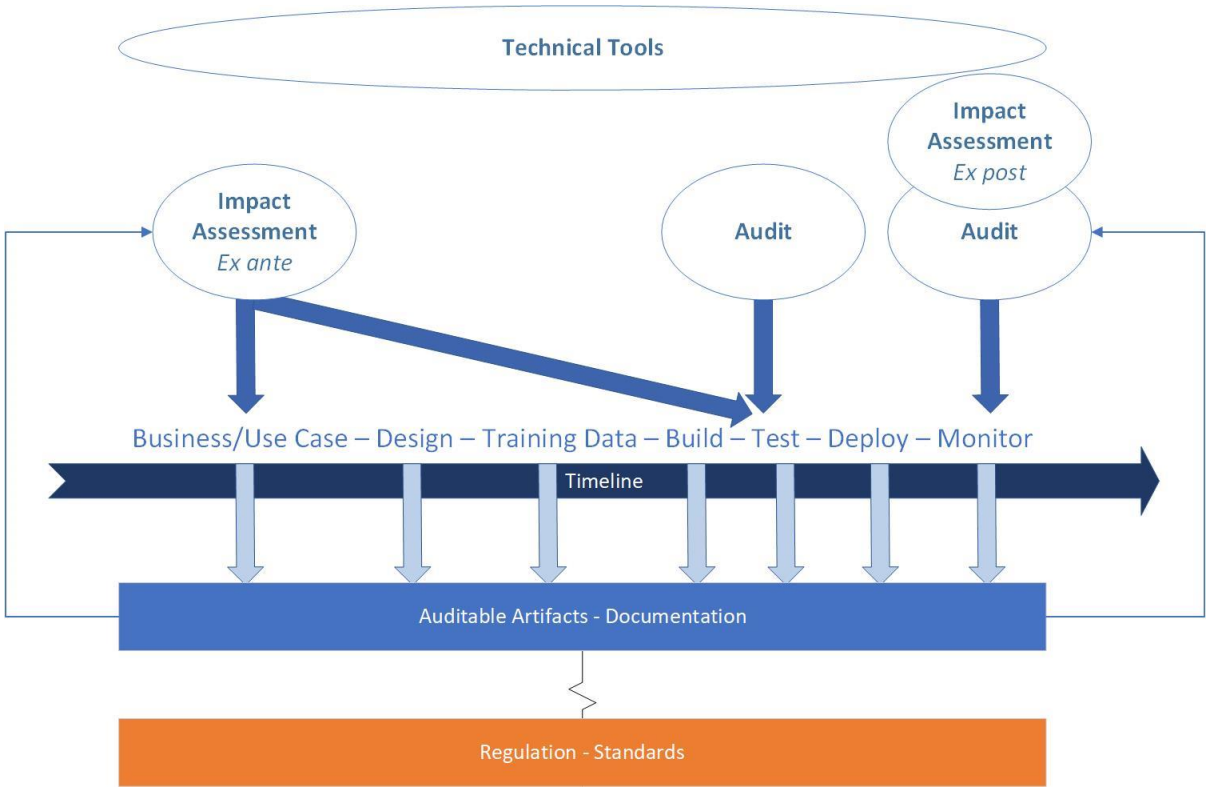


Figure 14 Process model for application of AI ethics tools to the development pipeline

The discussion of these tools with respondents in the tech industry in Chapter 6 shows that there is very limited adoption of the proposed principles and ethics tools analysed in Chapter 5. Actors in this sector are confident they have the right processes in place already to meet any ethical challenges in their products. It should be noted though that their perceptions of ethical challenges tended to be somewhat narrow, focused on traditional data protection principles and security issues. An interesting accountability gap was noted during these interviews between the vendor and buyer of products. Vendors identified reputational blow back risks from their product when deployed by the customer prompting some to engaged in vetting of customers where possible, or to admit that the ability to influence how products were used out in the wild was beyond their control.

Given the limitations identified in the interview data collection process (Section 3.9, Limitation 4) the evidence that the interview set provides does not provide a complete answer to RQ0 'Are the AI ethics tools being proposed fit for purpose for use by SMEs?' as participants frequently only had a brief overview of the example tools supplied. This therefore meant they were not giving responses to the application or practice of applying the tools, but their responses gave more of an indication to their views and position in relation to such tools in general. To gain more insight to better address RQ0, a different approach to engagement with the tools would be beneficial, as suggested in Section 3.9, Limitation 4 and 7, to deepen the responses to how these tools might be practically applied from both a senior management and developer perspective. Another approach

to answering RQ0 which was not planned in this research would be to pilot a tool in a company for a specific business use and gather data on the process to understand how the tool worked in practical application, rather than just relying on participants giving an opinion on how they imagine a tool may or may not work.

This study contributes to the discussion about ethical AI by clarifying the different themes emerging in this landscape. It also serves to illustrate how complex this landscape is, as others have noted (Morley *et al.*, 2019; Mulgan, 2019; Vakkuri *et al.*, 2019; Carrier and Brown, 2021; Lee and Singh, 2021; Schiff *et al.*, 2021), which provides a barrier to those developing or purchasing AI systems when it is unclear which tool is appropriate for their purposes. Addressing ethical issues systematically requires resource and time, familiarity with assessment/audit regimes and the ability to use the outputs of these tools to make judgements. This was a concern voiced in the industry interviews where business pressures and lack of resource limit investment in ethical assessment processes. Even with the aid of procedures and processes to surface ethical risks, there are still difficult judgements to be made in the real world. Competing claims between different actors, balancing protection and benefits and differing ethical viewpoints mean that even the most rigorously applied tools will still require complex human judgements. As Floridi (2017) observes 'there is no ethics without choices, responsibilities, and moral evaluations, all of which need a lot of relevant and reliable information and quite a good management of it.' Ethical tools can though, provide a reliable evidence base on which to make decisions, but without robust oversight may result in procedures that produce a checklist mentality and performative gestures that constitute 'ethics washing' (Kitchin, 2016; Bietti, 2020; Raab, 2020).

An important finding from the document analysis also puts in plain sight the fact that these tools are emerging in a landscape where currently there are no specific regulatory regimes or legislation for AI systems. In Figure 14 the base-level – Regulation and Standards – has no direct connection to the other processes and artifacts illustrated as there are currently no specific regulations. This means that these tools are for voluntary self-regulation without external governance mechanisms where third-party agents can interrogate the process and decisions. As Raab notes, 'an organisation or profession that simply marks its own homework cannot make valid claims to be trustworthy' (Raab, 2020, p. 13). Impact assessment and audit practices in other domains as discussed previously sit within national and international regulation and provide for external verification and assurance. Metcalf et. al. (2021) conclude that historically impact assessments are tools for evaluation that operate within relationships of accountability between different stakeholder groups. As this study

reveals there is currently a focus on a narrow group of internal stakeholders, with little transparency or accountability to wider stakeholders. In order for those who build AI products and services, and those who buy them, to provide credible and trustworthy governance of this technology, external verification, means of redress and contestation by different stakeholder groups, and methods of control for wrongdoing are required.

There are moves now to draft legislation to address the specific problems AI systems can produce with the EU leading the global pack with its recently published 'Proposal for a Regulation laying down harmonised rules on artificial intelligence' (European Commission, 2021). It proposes a risk-based approach to AI regulation, proposing an audit regime which will strengthen enforcement and sets out 'new requirements for documentation, traceability and transparency… the framework will envisage specific measures supporting innovation, including regulatory sandboxes and specific measures supporting small-scale users and providers of high-risk AI systems to comply with the new rules' (European Commission, 2021, p. 10). China is also working on these challenges with new regulation being proposed for data protection which includes processing using AI techniques, and specific new regulation for applications like facial recognition and autonomous vehicles (Lee *et al.*, 2021; Webster, 2021). The respondents from industry in this study were certainly not opposed to regulation and in some cases felt strongly that firm regulation and appropriate guidance made a positive contribution to innovation.

In the US a surprisingly strongly worded blog by the Federal Trade Commission (FTC) states that companies building or deploying AI should be 'using transparency frameworks and independent standards, by conducting and publishing the results of independent audits, and by opening your data or source code to outside inspection… your statements to business customers and consumers alike must be truthful, non-deceptive, and backed up by evidence' (Jillson, 2021). The post makes reference to a range of existing laws which might be applied to AI products and warns 'keep in mind that if you don't hold yourself accountable, the FTC may do it for you' (Jillson, 2021). As Joanna Bryson argues 'All human activity, particularly commercial activity, occurs in the context of some sort of regulatory framework' (Bryson, 2020, p. 8). Providing assurance of the safety, security and reliability of a project, product or system is the basis for the impact assessment and audit traditions discussed in this paper, the practices of which can be usefully applied to the domain of AI. It should also be noted that these traditions sit within established legal and regulatory frameworks. AI will need a similar regulatory ecosystem, which are being considered in multiple jurisdictions but are yet to be formally adopted.

The findings of this study also serve to illustrate the confusion in language and approach of the key features of impact assessment and audit. The latest thinking emerging from the UK Centre for Data

Ethics and Innovation (CDEI) echoes the findings in this research, recognising the need for clarification around AI ethics methodologies in practice (CDEI, 2021). The CDEI categorises the difference between impact assessment and audit and assurance in a similar way to the mapping in Figure 14, which they divide into compliance assurance (audit), and risk assurance (impact assessment) which are used at different stages of the process and meet different needs.

> 'The current discourse sometimes mistakenly calls on risk assurance tools like impact assessments to achieve the goals of compliance, leading to complex and burdensome efforts to address common challenges. Meanwhile, sometimes compliance mechanisms like audits are discussed as if they can achieve loftier goals - an exercise which may be better suited to Risk Assurance tools like impact assessments' (CDEI, 2021).

Clarifying the types of tools appropriate for which assessment and governance outcomes and implementing well-regulated compliance regimes for producers of AI systems would be a great step towards effectively operationalising the ethical principles and concerns motivating the production of AI ethics tools. A note of caution though on how effective regulation might be, see for example the recent European Parliament resolution on UK protection of personal data where concern is expressed 'about the lack and often non-existent enforcement of the GDPR by the UK when it was still a member of the EU; points, in particular, to the lack of proper enforcement by the UK Information Commissioner's (ICO's) Office in the past' (European Parliament, 2021, p. 6).

The document analysis highlighted gaps in the range of stakeholders included in AI ethics tools. As Figure 8 illustrates, current tools, not surprisingly, are designed for use by those in the production process of AI systems and the key decision-makers around that process. Participation in these tools was found to be limited beyond these core stakeholders, except for tools explicitly focused on participatory design processes (see for example (Madaio *et al.*, 2020)). There is a long tradition in HCI of Participatory Design (PD), and Human/Ethically/Value-centred Design (Simonsen and Robertson, 2012), which have been wrestling with the problem of inclusion and participation in the process of design and production of ICT systems (Beck, 2002). Participatory processes have also been addressed in pTA where governance of emerging technology includes deliberative public forums (Westin, 1971; CSPO, 2021), and research organisations like the Ada Lovelace Institute enabling 'informed and complex public dialogue about technology, policy and values, and represent the voice of the public in debates around data and AI' (Ada Lovelace Institute, 2020).

Including wider stakeholders presents challenges at the level of companies producing AI systems, as it is time and resource heavy and requires particular skillsets not necessarily present in developer teams. In the industry interviews it was clear that the perception of who was a stakeholder was narrowly defined. Participation is also about power, who has the power to decide, who is invited to

the table, whose views and goals take precedence. As Beck pointed out in the field of participatory design 'rather than participation, concern with power and dominance needs to be stated as the core of the research field' (Beck, 2002, p. 77). Who should decide on the design and use/non-use of AI systems is often framed as a 'project of expert oversight', giving little or no input to those stakeholders subject to AI systems (Greene, Hoffmann and Stark, 2019), and where the process can become a form of 'participation washing' (Sloane *et al.*, 2020). This is where informed public debate must feed into regulation and the law, to ensure appropriate governance is in place to protect rights and represent the views of all stakeholders in a society.

It is interesting to note the level of cynicism expressed by interviewees over public statements of ethics principles, and how aware respondents were of these being used as a branding and marketing exercise. Respondents felt that without robust third-party assurance, ethics statements were an empty gesture. Similar criticisms have been levelled at ESG (Environmental, Social and Governance) metrics and statements which are produced by third-party analytics companies and are supposed to inform investors in making sustainable and socially responsible investments. Lack of regulation in this sector has led to charges of ESG metrics being nothing more than 'an opaque system that sanctifies and rewards the most rudimentary business practices' (Simpson, Rathi and Kishan, 2021). It has been suggested that AI ethics be included in ESG metrics, alongside privacy and data protection which are already one of the KPIs (CFA Institute, 2021). Similar to the ESG market, the global business intelligence and auditing industry is busy absorbing the AI ethics discussion and creating toolkits and business opportunities for auditing and assurance (PricewaterhouseCoopers, 2019; Deloitte, 2021). The 'big 4' accounting firms made over 55bn in revenue in 2020 for auditing and assurance services (Statista, 2021), and AI ethics auditing and assurance services are being developed as new market in this sector. Some of the respondents in this study were concerned that translating AI ethics into metrics and KPIs would result in forms of corporate 'ethicswashing' and become an enabler for 'business as usual' without really addressing the negative impacts.

Chapter 8 Conclusion

This thesis set out to answer the question:

RQ0: Are the AI ethics tools being proposed fit for purpose for use by SMEs?

After an initial case study to explore data flows and potential ethical challenges in geospatial products (Chapter 4) the following research questions were formulated to answer the overall question RQ0:

RQ1: What practical tools are being proposed to operationalise AI ethics?

RQ1 was answered with a systematic collection of AI ethics documents which proposed practical tools (excluding those which only proposed principles). This showed that out of n=169 documents, only n=39 consisted of practical tools (see Table 13).

RQ2: What features do these tools have when compared to existing practices in other domains?

RQ2 was addressed by reviewing best practice and tools in other domains and developing a set of typologies (see Section 3.3), to understand the features of the AI ethics tools, and identify any gaps (see Chapter 5).

RQ3: How are these tools understood and used by senior decision-makers in SMEs?

RQ3 was answered by conducting 10 interviews with senior decision-makers (Chapter 6) to understand the current position in SMEs.

This study identified a preponderance of work on ethical principles for AI, and far fewer proposals for operationalising these principles for industry. This gap has been identified elsewhere (Morley *et al.*, 2019, 2021; Ada Lovelace Institute and DataKindUK, 2020; Raab, 2020), so it is hoped that work will continue to develop tools that are practical and applicable, particularly for SMEs. CEOs and founders in SMEs report lack of finance, staff and expertise as barriers to implementing additional AI ethics processes. This indicates that uptake of the principles and the application of related tools is limited, and very context dependant. The respondents in the interviews considered they understood the risks and were taking all necessary steps to produce responsible products. It was clear from the interviews that respondents understood of the scope of AI ethics to be focused on privacy and security issues, with many of the wider harms not being recognised (see Section 6.6.12). It is possible that AI ethics itself runs the risk of narrowing its concerns to a limited set of technical concerns to be mitigated by technocratic actors, excluding wider participation and influence of non-

technical stakeholders (Crawford and Calo, 2016; Hagendorff, 2021). It is certainly the case that this study identified gaps in both the landscape of tools and in practitioners in the field, where deeper questions of influence on social and democratic structures, or environmental impact were overlooked. This exclusion of stakeholders affected by the negative externalities can also be traced throughout the history of impact assessment and audit in other domains. Environmental impact assessments have long been criticised for serving those stakeholders who wield power and marginalising those without (Beck, 2002). Taking the best practice forward, while acknowledging the dangers of AI ethics practice falling into a process of empty virtue signalling will be an ongoing challenge.

8.1 Future work

AI ethics is still a nascent field and presents many opportunities for further research and development. Future work expanding on the studies in this thesis could consist of:

i. Development of an AI ethics tool specifically for SMEs.
ii. Workshops for developers using example tools to understand how these might usefully be adapted to meet the needs of the AI production pipeline.
iii. In-house piloting of AI ethics tools to understand how they might roll out in the live production process.
iv. Developing the typologies created in this thesis to produce a framework for analysing suggested tools and approaches in a comparable manner and use this to recommend tools for specific purposes/use cases.
v. Younger, less experienced founders felt particularly exposed to risks, and lacked the knowledge or experience to easily put governance measures in place. There is a need for education and support materials to be disseminated to accelerators to support start up founders.
vi. Work on specific ethics tools for high-risk applications like biometrics, and for specific sectors like law enforcement. As one of the interview respondents noted:

'I think used in the right way, biometrics and facial recognition offers massively better security now, across the board, you know, against attacks and very dangerous attacks in public, then potentially it could be used in the right way. But to me, it seems crazy. We have 45 police forces all trying different things with different providers. That is crazy to me, there ought to be a national framework for this.'

(Respondent #1, 2020)

8.2 Afterward

‘Real stupidity beats artificial intelligence every time.’ (Pratchett, 1997, p. 223)

Appendix A Source documents for analysis

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
1	Risks, Harms and Benefits Assessment	2017	UN Global Pulse	UN Global Pulse	https://www.unglobalpulse.org/policy/risk-assessment/	27/06/2018
2	AI and Big Data: A blueprint for a human rights, social and ethical impact assessment	2018	Mantelero, Alessandro	Computer Law & Security Review	10.1016/j.clsr.2018.05.017	17/05/2019
3	ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY	2018	Reisman, Dillon; Schultz, Jason; Crawford, Kate; Whittaker, Meredith	AI Now Institute	https://ainowinstitute.org/aiareport2018.pdf	24/06/2019
4	An Ethical Toolkit for Engineering/Design Practice	2018	Vallor, S; McKenna, D	Markkula Center for Applied Ethics, Santa Clara University	https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/	14/09/2019
5	Ethical Data and Information Management: Concepts, Tools and Methods	2018	O'Keefe, Katherine; Brien, Daragh O.	Kogan Page Ltd.	978-0-7494-8205-3	15/01/2020
6	Ethical OS	2018	Institute for the Future; Omidyar Network	Ethical.os	https://ethicalos.org/	13/06/2019
7	Ethics & Algorithms Toolkit (beta)	2018	GovEx; City and County of San Francisco; Harvard DataSmart; Data Community DC	Ethicstoolkit.ai	https://ethicstoolkit.ai/	27/01/2020

Appendix A

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
8	AI Fairness 360	2019	IBM Research	IBM	aif360.mybluemix.net/resources	12/01/2020
9	AI Procurement in a Box	2019	World Economic Forum	World Economic Forum	https://www.weforum.org/reports/ai-procurement-in-a-box/	13/10/2020
10	AI-RFX Procurement Framework	2019	The Institute for Ethical AI & Machine Learning		https://ethical.institute	18/06/2019
11	Algorithmic Impact Assessment (AIA)	2019	Secretariat, Treasury Board of Canada	Government of Canada	https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html	27/06/2019
12	Codex for Data- Based Value Creation	2019	Swiss Alliance for Data-Intensive Services Expert Group	Swiss Alliance for Data-Intensive Services	www.data-service-alliance.ch/codex	16/03/2020
13	Consequence Scanning – doteveryone	2019	Doteveryone	Doteveryone.org	https://doteveryone.org.uk/project/consequence-scanning/	18/06/2019
14	IBM Watson OpenScale	2019	IBM	IBM	https://www.ibm.com/uk-en/cloud/watson-openscale	13/11/2020
15	IEEE SA - The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	2019	IEEE Standards Association	IEEE	https://standards.ieee.org/industry-connections/ecpais.html	30/08/2019
16	Judgment Call the Game: Using Value Sensitive Design and	2019	Ballard, Stephanie; Chappell, Karen	Proceedings of the 2019 on Designing	10.1145/3322276.3323697	16/11/2020

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
	Design Fiction to Surface Ethical Concerns Related to Technology		M.; Kennedy, Kristen	Interactive Systems Conference		
17	Model Cards for Model Reporting	2019	Mitchell, Margaret; Wu, Simone; Zaldivar, Andrew; Barnes, Parker; Vasserman, Lucy; Hutchinson, Ben; Spitzer, Elena; Raji, Inioluwa Deborah; Gebru, Timnit	asXiv Working Paper	10.1145/3287560.3287596	25/09/2019
18	Model Ethical Data Impact Assessment	2019	IAF	Information Accountability Foundation	http://informationaccountability.org/publications/	08/12/2019
19	ODI Data Ethics Canvas	2019	ODI	ODI	https://theodi.org/article/data-ethics-canvas/	27/06/2019
20	Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector	2019	Leslie, David	The Alan Turing Institute	https://zenodo.org/record/3240529	13/01/2020
21	A Proposed Model AI Governance Framework - Second Edition	2020	PDPC Singapore	Personal Data Protection Commission Singapore	https://www.pdpc.gov.sg/resources/model-ai-gov	12/01/2020
22	AI Blindspot: A Discovery Process for preventing, detecting,	2020	Calderon, A; Taber, D; Qu, H; Wen, J	MIT	https://aiblindspot.media.mit.edu/	09/11/2020

Appendix A

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
	and mitigating bias in AI systems					
23	Algorithm Register	2020	City of Amsterdam	City of Amsterdam	https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/grip-op-algoritmes/	27/11/2020
24	Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment	2020	EU HLEG AI	European Commission	https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence	30/08/2020
25	Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing	2020	Raji, Inioluwa Deborah; Smart, Andrew; White, Rebecca N; Mitchell, Margaret; Gebru, Timnit; Hutchinson, Ben; Smith-Loud, Jamila; Theron, Daniel; Barnes, Parker	FAT* '20 Barcelona	https://dl.acm.org/doi/pdf/10.1145/3351095.3372873	16/11/2020
26	Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI	2020	Madaio, Michael A.; Stark, Luke; Wortman Vaughan, Jennifer; Wallach, Hanna	Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems	10.1145/3313831.3376445	08/10/2020
27	Corporate Digital Responsibility	2020	Lobschat, Lara; Mueller, Benjamin; Eggers, Felix; Brandimarte, Laura;	Journal of Business Research	10.1016/j.jbusres.2019.10.006	28/01/2020

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
			Diefenbach, Sarah; Kroschke, Mirja; Wirtz, Jochen			
28	Data Ethics Framework	2020	DCMS	Gov.uk	https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-legislation-and-codes-of-practice-for-use-of-data	13/10/2020
29	Datasheets for Datasets	2020	Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; Daumé III, Hal; Crawford, Kate	arXiv:1803.09010 [cs]	arXiv:1803.09010 [cs]	12/06/2020
30	Empowering AI Leadership	2020	World Economic Forum	World Economic Forum	https://spark.adobe.com/page/RsXNkZANwMLEf/	30/09/2020
31	Fairlearn: A toolkit for assessing and improving fairness in AI	2020	Bird, Sarah; Dudík, Miroslav; Edgar, Richard; Horn, Brandon; Lutz, Roman; Milan, Vanessa; Sameki, Mehrnoosh; Wallach, Hanna; Walker, Kathleen; Design, Allovus	IBM	https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf	13/10/2020
32	IEEE Draft Model Process for Addressing Ethical Concerns During	2020	IEEE Standards Association	IEEE	https://standards.ieee.org/project/7000.html	04/06/2020

Appendix A

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
	System Design P7000/D3					
33	IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being Std 7010	2020	IEEE Standards Association	IEEE	https://standards.ieee.org/industry-connections/ec/autonomous-systems.html	30/08/2020
34	Responsible AI	2020	TensorFlow	Tensorflow.org	https://www.tensorflow.org/resources/responsible-ai	02/11/2020
35	Standard Clauses for Municipalities for Fair Use of Algorithmic Systems	2020	City of Amsterdam	City of Amsterdam	https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/grip-op-algoritmes/	27/11/2020
36	Toward situated interventions for algorithmic equity: lessons from the field	2020	Katell, Michael; Young, Meg; Dailey, Dharma; Herman, Bernease; Guetler, Vivian; Tam, Aaron; Binz, Corinne; Raz, Daniella; Krafft, P. M.	Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency	10.1145/3351095.3372874	28/01/2020
37	Value-based Engineering for Ethics by Design	2020	Spiekermann, Sarah; Winkler, Till	IEEE pre-print	arXiv:2004.13676 [cs]	06/10/2020
38	Welcome to the Artificial Intelligence Incident Database	2020	Partnership on AI	The Partnership on AI	https://incidentdatabase.ai/	21/11/2020
39	White Paper on Data Ethics in Public Procurement of AI-	2020	Hasselbalch, Gry; Olsen, B; Tranberg, P	DataEthics.eu	https://dataethics.eu/wp-content/uploads/dataethic	25/08/2020

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
	based Services and Solutions				s-whitepaper-april-2020.pdf	

Appendix B Interview questions

Introduction

Appendix A Ethics Application 55529

Introducing the research project purpose and interviewer.

Providing the interviewees the opportunity to introduce themselves, their job roles and their company. (What is your job/role/position? How would you describe what you do? Tell me about your work etc.)

Ethical Issues

1. Do you think there are any ethical challenges raised by your own business/product/service?

2. If yes, do you have any process for considering the ethical impacts of the product or service? Formal, informal? What form does this take?

3. How are any conflicts resolved?

4. Is your main concern compliance?

5. Are there any differences in the ethical challenges arising from data-driven technologies currently than in the past? Can traditional approaches solve these problems (e.g. data protection, risk assessment, audit)?

Models proposed to meet ethical challenges

Talk participant through the 4 examples –

A. IBM (high-level corporate statements) (IBM, 2018)
B. Trustworthy AI (high-level statements with an assessment checklist) (High Level Expert Group on AI, 2019)
C. IAF (audit document with risk assessment element) (IAF, 2019)
D. Doteveryone (Agile event) (Doteveryone, 2019)

1. Do you think high level principles are effective (e.g. IBM)? Do you have any principles like these? Do you think they affect the way a company operates? Virtue-signalling?
2. Do you think lists of questions are useful (e.g. Trustworthy AI)? How could they be deployed in your company?
3. What do you think of the more structured audit approach of IAF?
4. Would you use the approach the Doteveryone Agile tool takes? How do you think the results could be managed?
5. What sort of costs do you think these processes would place on your business activities?

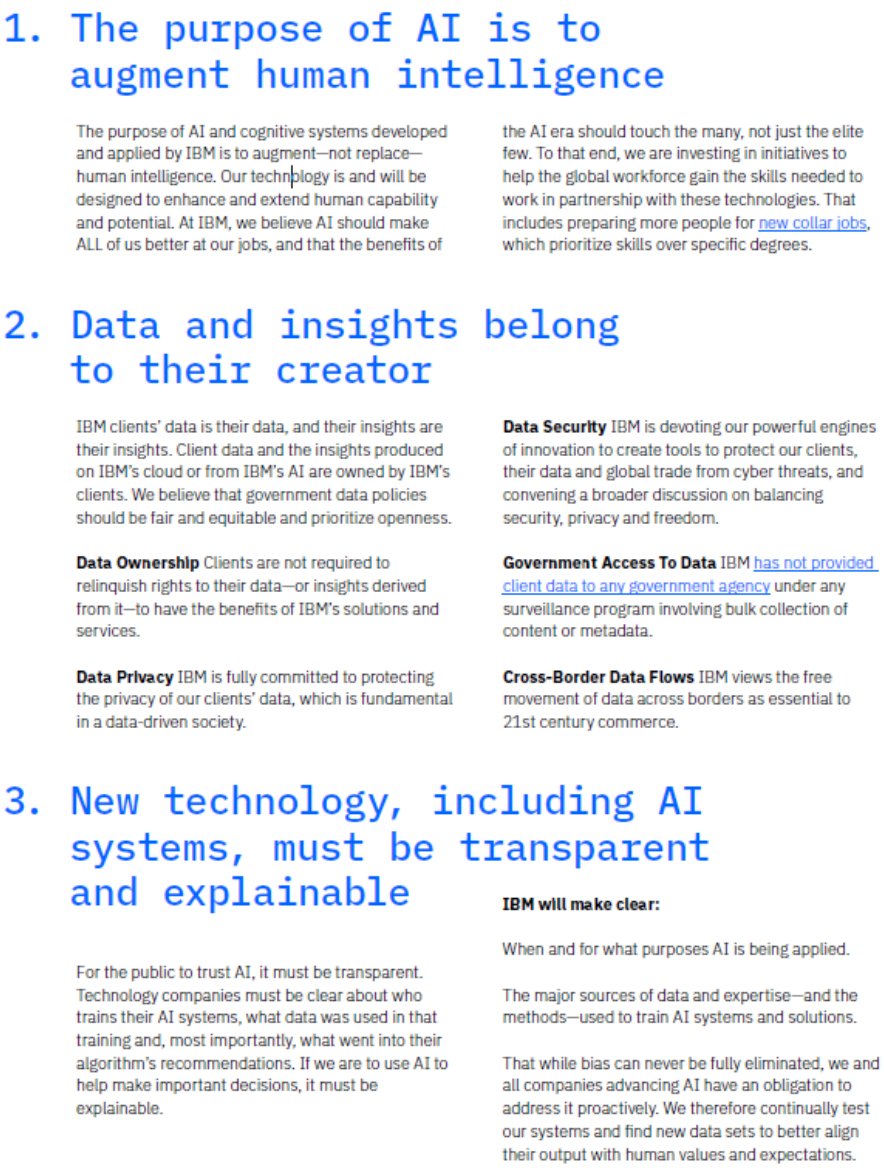
6. Do you think they might stifle innovation?
7. Do you use anything similar at the moment?

Closing discussion

1. Any further thoughts or comments on managing ethical risks not covered in our discussion above?

Appendix C Snapshots of example tools

C.1 Example A IBM Principles



1. The purpose of AI is to augment human intelligence

The purpose of AI and cognitive systems developed and applied by IBM is to augment—not replace—human intelligence. Our technology is and will be designed to enhance and extend human capability and potential. At IBM, we believe AI should make ALL of us better at our jobs, and that the benefits of the AI era should touch the many, not just the elite few. To that end, we are investing in initiatives to help the global workforce gain the skills needed to work in partnership with these technologies. That includes preparing more people for new collar jobs, which prioritize skills over specific degrees.

2. Data and insights belong to their creator

IBM clients' data is their data, and their insights are their insights. Client data and the insights produced on IBM's cloud or from IBM's AI are owned by IBM's clients. We believe that government data policies should be fair and equitable and prioritize openness.

Data Ownership Clients are not required to relinquish rights to their data—or insights derived from it—to have the benefits of IBM's solutions and services.

Data Privacy IBM is fully committed to protecting the privacy of our clients' data, which is fundamental in a data-driven society.

Data Security IBM is devoting our powerful engines of innovation to create tools to protect our clients, their data and global trade from cyber threats, and convening a broader discussion on balancing security, privacy and freedom.

Government Access To Data IBM has not provided client data to any government agency under any surveillance program involving bulk collection of content or metadata.

Cross-Border Data Flows IBM views the free movement of data across borders as essential to 21st century commerce.

3. New technology, including AI systems, must be transparent and explainable

For the public to trust AI, it must be transparent. Technology companies must be clear about who trains their AI systems, what data was used in that training and, most importantly, what went into their algorithm's recommendations. If we are to use AI to help make important decisions, it must be explainable.

IBM will make clear:

When and for what purposes AI is being applied.

The major sources of data and expertise—and the methods—used to train AI systems and solutions.

That while bias can never be fully eliminated, we and all companies advancing AI have an obligation to address it proactively. We therefore continually test our systems and find new data sets to better align their output with human values and expectations.

C.2 Example B EU Trustworthy AI Assessment List

TRUSTWORTHY AI ASSESSMENT LIST (PILOT VERSION)

1. Human agency and oversight

Fundamental rights:

- ✓ Did you carry out a fundamental rights impact assessment where there could be a negative impact on fundamental rights? Did you identify and document potential trade-offs made between the different principles and rights?
- ✓ Does the AI system interact with decisions by human (end) users (e.g. recommended actions or decisions to take, presenting of options)?
  - Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?
  - Did you consider whether the AI system should communicate to (end) users that a decision, content, advice or outcome is the result of an algorithmic decision?
  - In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent?

Human agency:

- ✓ Is the AI system implemented in work and labour process? If so, did you consider the task allocation between the AI system and humans for meaningful interactions and appropriate human oversight and control?
  - Does the AI system enhance or augment human capabilities?
  - Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?

Human oversight:

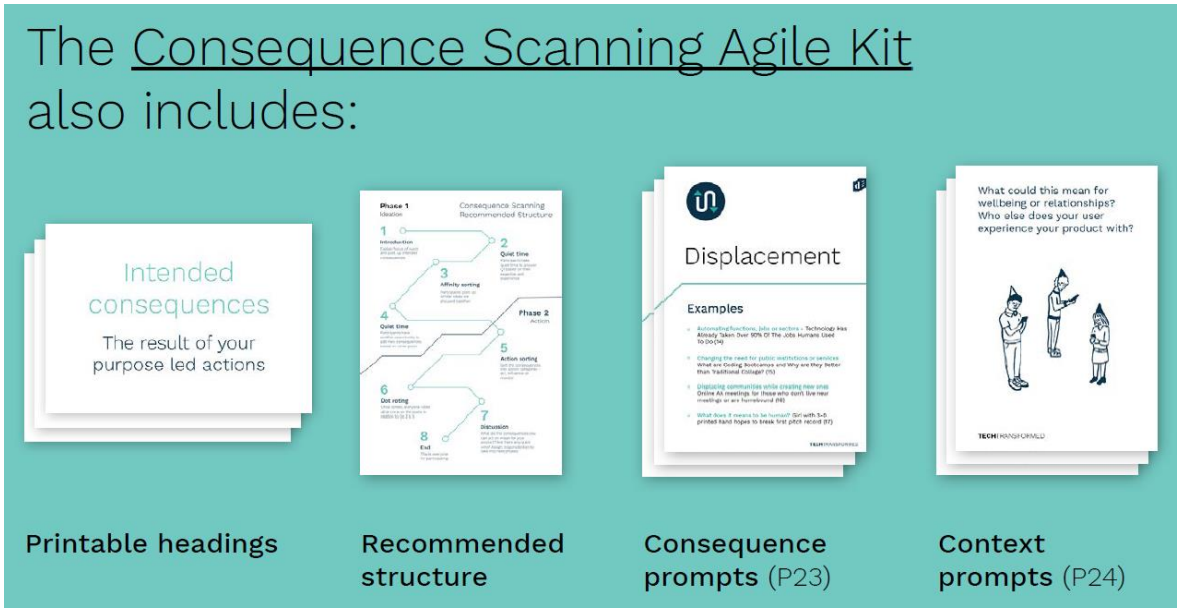
- ✓ Did you consider the appropriate level of human control for the particular AI system and use case?
  - Can you describe the level of human control or involvement?
  - Who is the "human in control" and what are the moments or tools for human intervention?
  - Did you put in place mechanisms and measures to ensure human control or oversight?
  - Did you take any measures to enable audit and to remedy issues related to governing AI autonomy?
- ✓ Is there is a self-learning or autonomous AI system or use case? If so, did you put in place more specific mechanisms of control and oversight?
  - Which detection and response mechanisms did you establish to assess whether something could go wrong?

C.3 Example C IAF Ethical Data Impact Assessment

Model Ethical Data Impact Assessment January 2019

EDIA Question	Answer/Notes
Section 1: Purpose of the Activity	
A. Business objective and purpose of the data activity	
1. What is the business need/goal/objective for this data activity? *If the purpose of the activity is to solve a question/problem, what particular question/problem is the activity trying to solve? Does the activity fit within a larger theme of work that is currently being contemplated or undertaken?*	
2. Is this activity an expansion of a previous activity? If yes, determine whether a previous assessment has been done. If a previous assessment has been done, what has changed in this data activity and why (refer to previous assessment)? *Does the activity fit within a larger theme of work that is currently being contemplated or undertaken?*	
B. Accountability for the data activity	
1. Who has ultimate decision-making authority for the data activity? *Who else needs to be involved in making the decision regarding the activity?*	
2. Who is accountable for the various phases of the data activity? *Who are the leaders that are responsible for the activity?*	
C. Legal and Other obligations regarding data collection, analysis and use(s)	
1. What laws apply to the collection, analysis and use(s) of data?	
2. Does the data activity comply with all organizational policies and self-regulatory commitments?	

C.4 Example D Doteveryone Consequence Scanning



List of references

Abrams, M. (no date) 'Fourth Privacy Legislative Wave | The Information Accountability Foundation - IAF'. Available at: http://informationaccountability.org/fourth-privacy-legislative-wave/ (Accessed: 19 August 2019).

Ackerman, F. and Heinzerling, L. (2001) 'Pricing the Priceless: Cost-Benefit Analysis of Environmental Protection', *University of Pennsylvania Law Review*, 150(5), pp. 1553–1584.

Ada Lovelace Institute (2020) *Our Strategy*. Available at: https://www.adalovelaceinstitute.org/about/ (Accessed: 26 May 2021).

Ada Lovelace Institute and DataKindUK (2020) *Examining the Black Box: Tools for assessing algorithmic systems*. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/ (Accessed: 23 February 2021).

Affordable Warmth Solutions CIC (2019) *WHF Latest Update*, *Affordable Warmth Solutions*. Available at: https://www.affordablewarmthsolutions.org.uk/warm-homes-fund/whf-latest-update (Accessed: 14 February 2019).

AGI Standards Committee (2018) *UK GEMINI (GEo-spatial Metadata INteroperability iNItiative)*. Available at: https://www.agi.org.uk/agi-groups/standards-committee/uk-gemini (Accessed: 10 May 2019).

Agile Manifesto (2001) *Manifesto for Agile Software Development*. Available at: https://agilemanifesto.org/ (Accessed: 13 August 2021).

AlgorithmWatch (2020) *AI Ethics Guidelines Global Inventory by AlgorithmWatch*, *AI Ethics Guidelines Global Inventory*. Available at: https://inventory.algorithmwatch.org (Accessed: 11 August 2020).

Arena, M., Arnaboldi, M. and Azzone, G. (2010) 'The organizational dynamics of Enterprise Risk Management', *Accounting, Organizations and Society*, 35(7), pp. 659–675. doi:10.1016/j.aos.2010.07.003.

Aven, T. (2016) 'Risk assessment and risk management: Review of recent advances on their foundation', *European Journal of Operational Research*, 253(1), pp. 1–13. doi:10.1016/j.ejor.2015.12.023.

AWC CIC (2019) *New Gas Connection Application » Affordable Warmth Solutions*, *Affordable Warmth Solutions*. Available at: https://www.affordablewarmthsolutions.org.uk/new-gas-connection-application/ (Accessed: 20 May 2019).

Ayling, J.A. (2017) *How is the new EU General Data Protection Regulation impacting on data management in the voluntary sector?* MSc Dissertation. University of Southampton Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science.

Badr, W. (2019) *Evaluating Machine Learning Models Fairness and Bias.*, *Medium*. Available at: https://towardsdatascience.com/evaluating-machine-learning-models-fairness-and-bias-4ec82512f7c3 (Accessed: 13 November 2020).

List of references

Bantilan, N. (2017) 'Themis-ml: A Fairness-aware Machine Learning Interface for End-to-end Discrimination Discovery and Mitigation', *arXiv:1710.06921 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1710.06921 (Accessed: 13 November 2020).

Baxter, G. and Sommerville, I. (2011) 'Socio-technical systems: From design methods to systems engineering', *Interacting with Computers*, 23(1), pp. 4–17. doi:10.1016/j.intcom.2010.07.003.

Beck, E. (2002) 'P for Political: Participation is Not Enough', *Scandinavian Journal of Information Systems*, 14(1). Available at: https://aisel.aisnet.org/sjis/vol14/iss1/1.

Beck, P.U. (1992) *Risk Society: Towards a New Modernity*. London: SAGE.

BEIS (2018) *Fuel poverty statistics*, *GOV.UK*. Available at: https://www.gov.uk/government/collections/fuel-poverty-statistics (Accessed: 19 May 2019).

Bellamy, R.K.E. *et al.* (2018) 'AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias', *arXiv:1810.01943 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1810.01943 (Accessed: 27 May 2021).

Bengtsson, M. (2016) 'How to plan and perform a qualitative study using content analysis', *NursingPlus Open*, 2, pp. 8–14. doi:10.1016/j.npls.2016.01.001.

Bernard, H.R. and Gravlee, C.C. (2014) *Handbook of Methods in Cultural Anthropology*. Rowman & Littlefield.

Biernacki, P. and Waldorf, D. (1981) 'Snowball Sampling - Problems and Techniques of Chain Referral Sampling', *Sociological Methods & Research*, 10(2), pp. 141–163.

Bietti, E. (2020) 'From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona, Spain: Association for Computing Machinery (FAT* '20), pp. 210–219. doi:10.1145/3351095.3372860.

Bird, S. *et al.* (2020) *Fairlearn: A toolkit for assessing and improving fairness in AI*. Microsoft. Available at: https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf (Accessed: 13 October 2020).

Birhane, A. *et al.* (2021) 'The Values Encoded in Machine Learning Research', *arXiv:2106.15590 [cs]* [Preprint]. Available at: http://arxiv.org/abs/2106.15590 (Accessed: 25 July 2021).

Bloomfield, R. *et al.* (2019) 'Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy', *Computer*, 52(9), pp. 82–89. doi:10.1109/MC.2019.2914775.

Brundage, M. *et al.* (2020) 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims', *arXiv:2004.07213 [cs]* [Preprint]. Available at: http://arxiv.org/abs/2004.07213 (Accessed: 16 November 2020).

Bryson, J.J. (2020) 'The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation', in Bryson, J. J., *The Oxford Handbook of Ethics of AI*. Edited by M. D. Dubber, F. Pasquale, and S. Das. Oxford University Press, pp. 1–25. doi:10.1093/oxfordhb/9780190067397.013.1.

Business Roundtable (2020) 'Our Commitment', *Business Roundtable - Opportunity Agenda*. Available at: https://opportunity.businessroundtable.org/ourcommitment/ (Accessed: 5 February 2021).

CAHAI (2020) *The feasibility study on AI legal framework CAHAI*. Council of Europe. Available at: https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da (Accessed: 9 December 2021).

Carrier, R. and Brown, S. (2021) *Taxonomy: AI Audit, Assurance, and Assessment*, *ForHumanity*. Available at: https://forhumanity.center/blog/taxonomy-ai-audit-assurance-and-assessment (Accessed: 26 April 2021).

CDEI (2021) *Types of assurance in AI and the role of standards*, *Centre for Data Ethics and Innovation Blog*. Available at: https://cdei.blog.gov.uk/2021/04/17/134/ (Accessed: 26 May 2021).

CFA Institute (2021) *ESG Investing and Analysis*, *CFA Institute*. Available at: https://www.cfainstitute.org/en/research/esg-investing (Accessed: 22 December 2021).

Chapman, A. *et al.* (2020) 'Capturing and querying fine-grained provenance of preprocessing pipelines in data science', *Proceedings of the VLDB Endowment*, 14(4), pp. 507–520. doi:10.14778/3436905.3436911.

Citron, D.K. and Solove, D.J. (2021) *Privacy Harms*. SSRN Scholarly Paper ID 3782222. Rochester, NY: Social Science Research Network. doi:10.2139/ssrn.3782222.

Clarke, R. (2009) 'Privacy impact assessment: Its origins and development', *Computer Law & Security Review*, 25(2), pp. 123–135. doi:10.1016/j.clsr.2009.02.002.

Clarke, R. (2011) 'An evaluation of privacy impact assessment guidance documents', *International Data Privacy Law*, 1(2), pp. 111–120. doi:10.1093/idpl/ipr002.

Clarke, T. (2005) 'Accounting for Enron: shareholder value and stakeholder interests', *Corporate Governance: An International Review*, 13(5), pp. 598–612. doi:10.1111/j.1467-8683.2005.00454.x.

CNIL FR (2018) *Guidelines on DPIA | CNIL*. Available at: https://www.cnil.fr/en/guidelines-dpia (Accessed: 20 November 2018).

Coates, J.F. (1974) 'Some methods and techniques for comprehensive impact assessment', *Technological Forecasting and Social Change*, 6, pp. 341–357. doi:10.1016/0040-1625(74)90035-3.

Crawford, K. (2021) *Atlas of AI: power, politics, and the planetary costs of artificial intelligence*. New Haven: Yale University Press.

Crawford, K. and Calo, R. (2016) 'There is a blind spot in AI research', *Nature*, 538(7625), pp. 311–313. doi:10.1038/538311a.

CSPO (2021) 'Participatory Technology Assessment | CSPO', *Consortium for Science and Policy Outcomes*. Available at: https://cspo.org/areas-of-focus/pta/ (Accessed: 12 February 2021).

Cullen, P. (2015) 'What Does Information Accountability 2.0 Look Like in a 21st Century Data World? | The Information Accountability Foundation - IAF'. Available at: http://informationaccountability.org/what-does-information-accountability-2-0-look-like-in-a-21st-century-data-world/ (Accessed: 15 September 2019).

Cullen, P. (2018) 'Evolving Accountability to Ethical Data Stewardship – A Key Part of Wave Four Privacy Laws | The Information Accountability Foundation - IAF'. Available at: http://informationaccountability.org/evolving-accountability-to-ethical-data-stewardship-a-key-part-of-wave-four-privacy-laws/ (Accessed: 15 September 2019).

Cullen, P. (2019) 'Evolving Ethical Data Impact Assessments | The Information Accountability Foundation - IAF', *IAF Blog*, 22 January. Available at: http://informationaccountability.org/evolving-ethical-data-impact-assessments/ (Accessed: 9 August 2019).

Cycyota, C.S. and Harrison, D.A. (2006) 'What (Not) to Expect When Surveying Executives: A Meta-Analysis of Top Manager Response Rates and Techniques Over Time', *Organizational Research Methods*, 9(2), pp. 133–160. doi:10.1177/1094428105280770.

Dasgupta, P. (2021) *Final Report - The Economics of Biodiversity: The Dasgupta Review*, *GOV.UK*. Available at: https://www.gov.uk/government/publications/final-report-the-economics-of-biodiversity-the-dasgupta-review (Accessed: 4 March 2021).

DCMS (2020) *National Data Strategy*, *GOV.UK*. Available at: https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy (Accessed: 23 February 2021).

Deloitte (2021) *Trustworthy Artificial Intelligence (AI)™*, *Deloitte United States*. Available at: https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html (Accessed: 22 December 2021).

Diakopoulos, N. (2016) 'Accountability in algorithmic decision making', *Communications of the ACM*, 59(2), pp. 56–62. doi:10.1145/2844110.

D'Ignazio, C. and Klein, L. (2018) 'Chapter Seven: The Power Chapter'. Available at: https://bookbook.pubpub.org/pub/7ruegkt6 (Accessed: 9 May 2019).

Donaldson, T. and Preston, L.E. (1995) 'The Stakeholder Theory of the Corporation: Concepts, Evidence, and Implications', *The Academy of Management Review*, 20(1), p. 65. doi:10.2307/258887.

Doteveryone (2019) 'Consequence Scanning – doteveryone'. Available at: https://doteveryone.org.uk/project/consequence-scanning/ (Accessed: 18 June 2019).

ECNL and Data & Society (2021) *Recommendations for incorporating human rights into AI impact assessments | ECNL*. Available at: https://ecnl.org/publications/recommendations-incorporating-human-rights-ai-impact-assessments (Accessed: 9 December 2021).

Edwards, M.M. and Huddleston, J.R. (2009) 'Prospects and Perils of Fiscal Impact Analysis', *Journal of the American Planning Association*, 76(1), pp. 25–41. doi:10.1080/01944360903310477.

Eubanks, V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

European Commission (2016) *SME definition*, *Internal Market, Industry, Entrepreneurship and SMEs - European Commission*. Available at: https://ec.europa.eu/growth/smes/sme-definition_en (Accessed: 8 August 2021).

European Commission (2021) *Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future*. Proposal. Brussels: European Commission. Available

at: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence (Accessed: 21 May 2021).

European Commission HLEG AI (2020) *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*, *Shaping Europe's digital future - European Commission*. Available at: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (Accessed: 30 August 2020).

European Council and Parliament (2016) *REGULATION (EU) 2016/679 General Data Protection Regulation*. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679 (Accessed: 23 September 2017).

European Parliament (2021) *The adequate protection of personal data by the United Kingdom*. Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0262_EN.html (Accessed: 26 May 2021).

Farmer, H., Strait, A. and Parker, I. (2021) *Regulate to innovate*. Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/regulate-innovate/ (Accessed: 30 November 2021).

Fereday, J. and Muir-Cochrane, E. (2006) 'Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development', *International Journal of Qualitative Methods*, 5(1), pp. 80–92. doi:10.1177/160940690600500107.

Financial Reporting Council (2020) *Auditors I Audit and Assurance I Standards and Guidance for Auditors I Financial Reporting Council*. Available at: https://www.frc.org.uk/auditors/audit-assurance/standards-and-guidance (Accessed: 26 April 2021).

Fjeld, J. *et al.* (2020) *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for Ai*. SSRN Scholarly Paper ID 3518482. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=3518482 (Accessed: 27 January 2020).

Floridi, L. (2017) *Why Information Matters*, *The New Atlantis*. Available at: http://www.thenewatlantis.com/publications/why-information-matters (Accessed: 14 October 2020).

Floridi, L. *et al.* (2018) 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds and Machines*, 28(4), pp. 689–707. doi:10.1007/s11023-018-9482-5.

Foden, C. (2019) *Our structure*, *City of Lincoln Council*. Available at: https://www.lincoln.gov.uk/council/structure (Accessed: 10 January 2021).

Freeman, R.E. (2010) *Strategic Management: A Stakeholder Approach*. Cambridge University Press.

Gebru, T. *et al.* (2020) 'Datasheets for Datasets', *arXiv:1803.09010 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1803.09010 (Accessed: 3 December 2020).

Glucker, A.N. *et al.* (2013) 'Public participation in environmental impact assessment: why, who and how?', *Environmental Impact Assessment Review*, 43, pp. 104–111. doi:10.1016/j.eiar.2013.06.003.

Greene, D., Hoffmann, A.L. and Stark, L. (2019) 'Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning', in. *Hawaii International Conference on System Sciences*. doi:10.24251/HICSS.2019.258.

Hagendorff, T. (2019) 'The Ethics of AI Ethics -- An Evaluation of Guidelines', *arXiv:1903.03425 [cs, stat]* [Preprint]. Available at: http://arxiv.org/abs/1903.03425 (Accessed: 12 January 2020).

Hagendorff, T. (2021) 'Blind spots in AI ethics', *AI and Ethics* [Preprint]. doi:10.1007/s43681-021-00122-8.

Hayne, C. and Free, C. (2014) 'Hybridized professional groups and institutional work: COSO and the rise of enterprise risk management', *Accounting, Organizations and Society*, 39(5), pp. 309–330. doi:10.1016/j.aos.2014.05.002.

Hennen, L. (2012) 'Why do we still need participatory technology assessment?', *Poiesis & Praxis*, 9, pp. 27–41. doi:10.1007/s10202-012-0122-5.

High Level Expert Group on AI (2019) *Ethics guidelines for trustworthy AI*. Text. Brussels: European Commission. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (Accessed: 23 May 2019).

Hirsch, D. *et al.* (2020) *Business Data Ethics: Emerging Trends in the Governance of Advanced Analytics and AI*. Ohio State Legal Studies Research Paper No. 628. The Ohio State University, p. 107. Available at: https://cpb-us-w2.wpmucdn.com/u.osu.edu/dist/3/96132/files/2020/10/Final-Report-1.pdf.

Hoffmann, A.L. (2019) 'Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse', *Information, Communication & Society*, 22(7), pp. 900–915. doi:10.1080/1369118X.2019.1573912.

Hutchinson, B. and Mitchell, M. (2019) '50 Years of Test (Un)fairness: Lessons for Machine Learning', in *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAT* '19), pp. 49–58. doi:10.1145/3287560.3287600.

IAF (2019) 'Model Ethical Data Impact Assessment', *IAF Publications*. Available at: http://informationaccountability.org/publications/ (Accessed: 12 August 2019).

IAIA (2009) *Technology Assessment*. Available at: https://www.iaia.org/wiki-details.php?ID=26 (Accessed: 26 January 2021).

IBM (2018) *IBM'S Principles for Data Trust and Transparency*, *THINKPolicy*. Available at: https://www.ibm.com/blogs/policy/trust-principles/ (Accessed: 21 January 2020).

ICO (2020) *Guidance on the AI auditing framework Draft guidance for consultation*. Information Commissioner's Office, p. 105. Available at: https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

ICO UK (2018) *Data protection impact assessments*. Available at: https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/ (Accessed: 7 June 2018).

IMA Europe (2020) *Life Cycle Assessment | IMA Europe*, *Industrial Mineral Association - Europe*. Available at: https://www.ima-europe.eu/eu-policy/environment/life-cycle-assessment (Accessed: 6 May 2021).

Institute for the Future and Omidyar Network (2018) 'Ethical OS'. Available at: https://ethicalos.org/ (Accessed: 21 June 2019).

International Organisation for Standardization (2021) *ISO - ISO 14000 family — Environmental management*, *ISO*. Available at: https://www.iso.org/iso-14001-environmental-management.html (Accessed: 26 July 2021).

International Organization for Standardization (2021a) *ISO - Certification*, *ISO*. Available at: https://www.iso.org/certification.html (Accessed: 15 July 2021).

International Organization for Standardization (2021b) *ISO - Standards*, *ISO*. Available at: https://www.iso.org/standards.html (Accessed: 15 July 2021).

Jillson, E. (2021) *Aiming for truth, fairness, and equity in your company's use of AI*, *Federal Trade Commission*. Available at: https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai (Accessed: 20 April 2021).

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399. doi:10.1038/s42256-019-0088-2.

Kaissis, G.A. *et al.* (2020) 'Secure, privacy-preserving and federated machine learning in medical imaging', *Nature Machine Intelligence*, 2(6), pp. 305–311. doi:10.1038/s42256-020-0186-1.

Kazim, E. and Koshiyama, A. (2020) *AI Assurance Processes*. SSRN Scholarly Paper ID 3685087. Rochester, NY: Social Science Research Network. doi:10.2139/ssrn.3685087.

Kemp, D. and Vanclay, F. (2013) 'Human rights and impact assessment: clarifying the connections in practice', *Impact Assessment and Project Appraisal*, 31(2), pp. 86–96. doi:10.1080/14615517.2013.782978.

Kende-Robbe, C. (2003) *Poverty and Social Impact Analysis : Linking Macroeconomic Policies to Poverty Outcomes: Summary of Early Experiences*, *IMF*. Available at: https://www.imf.org/en/Publications/WP/Issues/2016/12/30/Poverty-and-Social-Impact-Analysis-Linking-Macroeconomic-Policies-to-Poverty-Outcomes-16248 (Accessed: 12 February 2021).

Kind, C. (2020) 'The term "ethical AI" is finally starting to mean something', *VentureBeat*, 23 August. Available at: https://venturebeat.com/2020/08/23/the-term-ethical-ai-is-finally-starting-to-mean-something/ (Accessed: 23 August 2020).

Kiran, A., Oudshoorn, N.E.J. and Verbeek, P.P.C.C. (2015) 'Beyond checklists: toward an ethical-constructive technology assessment', *Journal of responsible innovation*, 2(1), pp. 5–19. doi:10.1080/23299460.2014.992769.

Kitchin, R. (2016) 'The ethics of smart cities and urban science', *Phil. Trans. R. Soc. A*, 374(2083), p. 20160115. doi:10.1098/rsta.2016.0115.

Kitchin, R. (2019) 'The ethics of smart cities'. Available at: https://www.rte.ie/brainstorm/2019/0425/1045602-the-ethics-of-smart-cities/ (Accessed: 7 May 2019).

Krippendorff, K. (2013) *Content Analysis: An Introduction to Its Methodology*. SAGE.

Lee, A. *et al.* (2021) *China's Draft Privacy Law Adds Platform Self-Governance, Solidifies CAC's Role | DigiChina*, *Stanford DigiChina Cyber Policy Unit*. Available at:

https://digichina.stanford.edu/news/chinas-draft-privacy-law-adds-platform-self-governance-solidifies-cacs-role (Accessed: 21 May 2021).

Lee, M.S.A., Floridi, L. and Singh, J. (2021) 'Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics', *AI and Ethics* [Preprint]. doi:10.1007/s43681-021-00067-y.

Lee, M.S.A. and Singh, J. (2021) 'The Landscape and Gaps in Open Source Fairness Toolkits', in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery (CHI '21), pp. 1–13. doi:10.1145/3411764.3445261.

Madaio, M.A. *et al.* (2020) 'Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery (CHI '20), pp. 1–14. doi:10.1145/3313831.3376445.

Mantelero, A. (2018) 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment', *Computer Law & Security Review*, 34(4), pp. 754–772. doi:10.1016/j.clsr.2018.05.017.

Mayring, P. (2019) 'Qualitative Content Analysis: Demarcation, Varieties, Developments', *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 20(3). doi:10.17169/fqs-20.3.3343.

Mehrabi, N. *et al.* (2021) 'A Survey on Bias and Fairness in Machine Learning', *ACM Computing Surveys*, 54(6), p. 115:1-115:35. doi:10.1145/3457607.

Met Office (2019) *Met Office DataPoint*, *Met Office*. Available at: https://www.metoffice.gov.uk/datapoint (Accessed: 16 May 2019).

Metcalf, J. *et al.* (2021) 'Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '21), pp. 735–746. doi:10.1145/3442188.3445935.

Ministry of Housing, Communities and Local Government (2019) *Domestic Energy Performance Certificate Register*. Available at: https://www.epcregister.com/ (Accessed: 15 May 2019).

Mishan, E.J. and Quah, E. (2020) *Cost-Benefit Analysis*. Routledge.

Mitchell, M. *et al.* (2019) 'Model Cards for Model Reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pp. 220–229. doi:10.1145/3287560.3287596.

Mittelstadt, B.D. *et al.* (2016) 'The ethics of algorithms: Mapping the debate', *Big Data & Society*, 3(2), p. 2053951716679679. doi:10.1177/2053951716679679.

Mökander, J. and Floridi, L. (2021) 'Ethics-Based Auditing to Develop Trustworthy AI', *Minds and Machines* [Preprint]. doi:10.1007/s11023-021-09557-8.

Morgan, R.K. (2012) 'Environmental impact assessment: the state of the art', *Impact Assessment and Project Appraisal*, 30(1), pp. 5–14. doi:10.1080/14615517.2012.661557.

Morley, J. *et al.* (2019) 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices', *Science and Engineering Ethics* [Preprint]. doi:10.1007/s11948-019-00165-5.

Morley, J. *et al.* (2021) 'Operationalising AI ethics: barriers, enablers and next steps', *AI & SOCIETY* [Preprint]. doi:10.1007/s00146-021-01308-8.

Moses, K. and Malone, R. (2004) *Development of Risk Assessment Matrix for NASA Engineering and Safety Center NASA Technical Reports Server (NTRS)*, *NASA Technical Reports Server (NTRS)*. Available at: https://ntrs.nasa.gov/citations/20050123548 (Accessed: 27 May 2021).

Mulgan, G. (2019) *AI ethics and the limits of code(s)*, *nesta*. Available at: https://www.nesta.org.uk/blog/ai-ethics-and-limits-codes/ (Accessed: 16 September 2019).

National Crime Agency (2021) *Our leadership*, *National Crime Agency*. Available at: https://www.nationalcrimeagency.gov.uk/who-we-are/our-leadership (Accessed: 10 January 2021).

Norval, C. *et al.* (2021) 'Data protection and tech startups: The need for attention, support, and scrutiny', *Policy & Internet*, 13(2), pp. 278–299. doi:10.1002/poi3.255.

ODI (2018) 'The Data Ethics Canvas – The ODI'. Available at: https://theodi.org/article/data-ethics-canvas/ (Accessed: 27 June 2019).

Office for Statistics Regulation (2021) *Ensuring statistical models command public confidence*. Office for Statistics Regulation, p. 76. Available at: https://osr.statisticsauthority.gov.uk/wp-content/uploads/2021/03/Ensuring_statistical_models_command_public_confidence.pdf.

Ofgem (2016a) *Energy Company Obligation*, *Ofgem*. Available at: https://www.ofgem.gov.uk/environmental-programmes/eco (Accessed: 13 May 2019).

Ofgem (2016b) *Ofgem scheme connecting more households to gas grid*, *Ofgem*. Available at: https://www.ofgem.gov.uk/news-blog/our-blog/ofgem-scheme-connecting-more-households-gas-grid (Accessed: 10 May 2019).

Ofgem (2017) *Decision to change the criteria for the Fuel Poor Network Extension Scheme*, *Ofgem*. Available at: https://www.ofgem.gov.uk/publications-and-updates/decision-change-criteria-fuel-poor-network-extension-scheme (Accessed: 10 May 2019).

OGC (2019) *Geography Markup Language | OGC*. Available at: http://www.opengeospatial.org/standards/gml (Accessed: 10 May 2019).

Ordnance Survey (2019a) *AddressBase products*. Available at: https://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-products.html (Accessed: 10 May 2019).

Ordnance Survey (2019b) *Policy statement – OS MasterMap Topographic Identifiers*. Available at: https://www.ordnancesurvey.co.uk/about/governance/policies/os-mastermap-toids.html (Accessed: 10 May 2019).

Palm, E. and Hansson, S.O. (2006) 'The case for ethical technology assessment (eTA)', *Technological Forecasting and Social Change*, 73(5), pp. 543–558. doi:10.1016/j.techfore.2005.06.002.

Patton, M.Q. (2014) *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. SAGE Publications.

Pearce, D.W. (2016) *Cost-Benefit Analysis, 2nd edition*. Macmillan International Higher Education.

Penrose, E. (1995) *The Theory of the Growth of the Firm*. 3rd edn. Oxford: Oxford University Press. doi:10.1093/0198289774.001.0001.

Plessis, J.J. du, Hargovan, A. and Harris, J. (2018) *Principles of Contemporary Corporate Governance*. Cambridge University Press.

Power, M. (2009) 'The risk management of nothing', *Accounting, Organizations and Society*, 34(6), pp. 849–855. doi:10.1016/j.aos.2009.06.001.

Pratchett, T. (1997) *Hogfather*. Random House.

PricewaterhouseCoopers (2019) *Responsible AI Toolkit*, *PwC*. Available at: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html (Accessed: 26 April 2021).

PwC UK (2013) *Understanding a financial statement audit*. UK: PricewaterhouseCooper. Available at: https://www.pwc.com/gx/en/audit-services/publications/assets/pwc-understanding-financial-statement-audit.pdf.

Raab, C.D. (2020) 'Information privacy, impact assessment, and the place of ethics', *Computer Law & Security Review*, 37, p. 105404. doi:10.1016/j.clsr.2020.105404.

Radford, J. and Joseph, K. (2020) 'Theory In, Theory Out: The Uses of Social Theory in Machine Learning for Social Science', *Frontiers in Big Data*, 0. doi:10.3389/fdata.2020.00018.

Renn, O. (2008) *Risk Governance: Coping with Uncertainty in a Complex World*. Earthscan.

Richards, D. (1996) 'Elite Interviewing: Approaches and Pitfalls', *Politics*, 16(3), pp. 199–204. doi:10.1111/j.1467-9256.1996.tb00039.x.

Roessler, B. (2008) 'New Ways of Thinking about Privacy', in *The Oxford Handbook of Political Theory*. Oxford: Oxford University Press. Available at: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199548439.001.0001/oxfordhb-9780199548439-e-38 (Accessed: 17 September 2017).

Rotem, N. and Locar, R. (2019) *Report: Data Breach in Biometric Security Platform Affecting Millions of Users*, *vpnMentor*. Available at: https://www.vpnmentor.com/blog/report-biostar2-leak/ (Accessed: 24 August 2021).

Rusby, R. (2015) *The Interpretation and Evaluation of Assurance Cases*. Technical Report SRI-CSL-15-01. SRI International, Menlo Park CA 94025, USA: Computer Science Laboratory.

Ryan, M. and Stahl, B.C. (2020) 'Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications', *Journal of Information, Communication and Ethics in Society*, ahead-of-print(ahead-of-print). doi:10.1108/JICES-12-2019-0138.

Schiff, D. (2020) 'AI Ethics Global Document Collection'. IEEE. Available at: https://ieee-dataport.org/open-access/ai-ethics-global-document-collection-0 (Accessed: 23 February 2021).

Schiff, D. *et al.* (2021) 'AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection', *IEEE Transactions on Technology and Society*, pp. 1–1. doi:10.1109/TTS.2021.3052127.

Schopfel, J. (2010) 'Towards a Prague Definition of Grey Literature', in *Conference Proceedings on Grey Literature*. *Twelfth International Conference on Grey Literature*, Amsterdam: Open Grey (The Grey Journal), p. 51. Available at: http://www.greynet.org/images/Contents_TGJ.V6.N1.pdf (Accessed: 12 July 2020).

Shannon, V., Green, B. and Raicu, I. (2018) *Ethics in Technology Practice: A toolkit*, *Markkula Center for Applied Ethics at Santa Clara University*. Available at: https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/ (Accessed: 19 August 2019).

Shearer, C. (2000) 'The CRISP-DM Model: The New Blueprint for Data Mining', 5(4), pp. 13–22.

Silverman, D. (2016) *Qualitative Research*. SAGE.

Simonsen, J. and Robertson, T. (2012) *Routledge International Handbook of Participatory Design*. London: Routledge.

Simpson, C., Rathi, A. and Kishan, S. (2021) 'Sustainable Investing Is Mostly About Sustaining Corporations', *Bloomberg.com*, December. Available at: https://www.bloomberg.com/graphics/2021-what-is-esg-investing-msci-ratings-focus-on-corporate-bottom-line/ (Accessed: 21 December 2021).

Singh, A. *et al.* (2018) *PriMP Visualization - Principled Artificial Intelligence Project*, *Harvard Law School, Berkman Klein Center for Internet and Society*. Available at: https://ai-hr.cyber.harvard.edu/primp-viz.html (Accessed: 24 June 2019).

Sirur, S., Nurse, J.R.C. and Webb, H. (2018) 'Are We There Yet? Understanding the Challenges Faced in Complying with the General Data Protection Regulation (GDPR)', in *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*. New York, NY, USA: Association for Computing Machinery (MPS '18), pp. 88–95. doi:10.1145/3267357.3267368.

Sloane, M. *et al.* (2020) 'Participation is not a Design Fix for Machine Learning', *arXiv:2007.02423 [cs]* [Preprint]. Available at: http://arxiv.org/abs/2007.02423 (Accessed: 26 May 2021).

Smith, K.B. (2002) 'Typologies, Taxonomies, and the Benefits of Policy Classification', *Policy Studies Journal*, 30(3), pp. 379–395. doi:https://doi.org/10.1111/j.1541-0072.2002.tb02153.x.

Solove, D.J. (2001) 'Privacy and Power: Computer Databases and Metaphors for Information Privacy', *STANFORD LAW REVIEW*, 53, p. 71.

Solove, D.J. (2006) 'A Taxonomy of Privacy', *University of Pennsylvania Law Review*, 154(3), p. 477. doi:10.2307/40041279.

Stanley, M. (2020) *UK Civil Service - Grades and Roles*, *Understanding Government*. Available at: https://www.civilservant.org.uk/information-grades_and_roles.html (Accessed: 10 January 2021).

Starr, C. (1969) 'Social Benefit versus Technological Risk', *Science*, 165(3899), pp. 1232–1238.

Statista (2021) *Big Four: revenue by function 2020*, *Statista*. Available at: https://www.statista.com/statistics/250935/big-four-accounting-firms-breakdown-of-revenues/ (Accessed: 22 December 2021).

Stephanidis, C. *et al.* (2019) 'Seven HCI Grand Challenges', *International Journal of Human–Computer Interaction*, 35(14), pp. 1229–1269. doi:10.1080/10447318.2019.1619259.

Stewart, B. (1996) 'Privacy impact assessments', *Privacy Law & Policy Reporter*, 39(3(4)). Available at: http://www.austlii.edu.au/au/journals/PLPR/1996/39.html (Accessed: 17 February 2021).

STOA (2021) *Centre for AI | Panel for the Future of Science and Technology (STOA) | European Parliament*. Available at: https://www.europarl.europa.eu/stoa/en/centre-for-AI (Accessed: 11 February 2021).

Suter, G.W., Barnthouse, L.W. and O'Neill, R.V. (1987) 'Treatment of risk in environmental impact assessment', *Environmental Management*, 11(3), pp. 295–303. doi:10.1007/BF01867157.

Taylor, J. (2019) *Major breach found in biometrics system used by banks, UK police and defence firms*, *the Guardian*. Available at: http://www.theguardian.com/technology/2019/aug/14/major-breach-found-in-biometrics-system-used-by-banks-uk-police-and-defence-firms (Accessed: 24 August 2021).

TensorFlow (2020) *Responsible AI*, *TensorFlow*. Available at: https://www.tensorflow.org/resources/responsible-ai (Accessed: 2 November 2020).

The Danish Institute for Human Rights (2016) *Human rights impact assessment guidance and toolbox - road-testing version*, *The Danish Institute for Human Rights*. Available at: https://www.humanrights.dk/business/tools/human-rights-impact-assessment-guidance-and-toolbox (Accessed: 3 February 2020).

Thompson, K.M., Deisler, P.F. and Schwing, R.C. (2005) 'Interdisciplinary Vision: The First 25 Years of the Society for Risk Analysis (SRA), 1980–2005', *Risk Analysis*, 25(6), pp. 1333–1386. doi:10.1111/j.1539-6924.2005.00702.x.

UK Space Agency (2018) *Ground-breaking satellite projects will transform society*, *GOV.UK*. Available at: https://www.gov.uk/government/news/ground-breaking-satellite-projects-will-transform-society (Accessed: 15 May 2019).

UN Environment (2018) *Assessing Environmental Impacts A Global Review Of Legislation - UNEP-WCMC*, *UNEP-WCMC's official website - Assessing Environmental Impacts A Global Review Of Legislation*. Available at: https://www.unep-wcmc.org/assessing-environmental-impacts--a-global-review-of-legislation (Accessed: 12 February 2021).

Vakkuri, V. *et al.* (2019) 'Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study', *arXiv:1906.07946 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1906.07946 (Accessed: 7 August 2020).

Veale, M., Van Kleek, M. and Binns, R. (2018) 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM (CHI '18), p. 440:1-440:14. doi:10.1145/3173574.3174014.

Wang, P. (2019) 'On Defining Artificial Intelligence', *Journal of Artificial General Intelligence*, 10(2) 1-37, p. 37.

Webster, G. (2021) *Translation: Personal Information Protection Law of the People's Republic of China (Draft) (Second Review Draft) | DigiChina*, *Stanford DigiChina Cyber Policy Unit*. Available at: https://digichina.stanford.edu/news/translation-personal-information-protection-law-peoples-republic-china-draft-second-review (Accessed: 21 May 2021).

Westin, A.F. (1967) *Privacy and Freedom*. Ig Publishing.

Westin, A.F. (1971) *Information Technology in a Democracy*. Harvard University Press.

Whittlestone, J. *et al.* (2019) 'The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions', in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Honolulu, HI, USA: Association for Computing Machinery (AIES '19), pp. 195–200. doi:10.1145/3306618.3314289.

van Wynsberghe, A. (2021) *Europe's trustworthy AI meets AI ethics*, *Deloitte Netherlands*. Available at: https://www2.deloitte.com/nl/nl/pages/risk/articles/europes-trustworthy-ai-meets-ai-ethics.html (Accessed: 13 January 2021).