Cambridge Judge Business School

Working Paper No. 01/2022

MANAGING NEW TECHNOLOGY: THE COMBINATION OF MODEL RISK AND ENTERPRISE RISK MANAGEMENT

Eleanor Toye Scott, Philip Stiles & Pradeep Debata

UNIVERSITY OF CAMBRIDGE
Judge Business School

Cambridge Judge Business School Working Papers

These papers are produced by Cambridge Judge Business School, University of Cambridge. They are circulated for discussion purposes only. Their contents should be considered preliminary and are not to be quoted without the authors' permission.

Cambridge Judge Business School author contact details are as follows:

Dr Philip Stiles
Cambridge Judge Business School
University of Cambridge

Email: p.stiles@jbs.cam.ac.uk

Please address enquiries about the series to:

Research Manager
Cambridge Judge Business School
University of Cambridge
Trumpington Street
Cambridge CB2 1AG
UK

Tel: +44 1223 760546
Email: research-support@jbs.cam.ac.uk

MANAGING NEW TECHNOLOGY: THE COMBINATION OF MODEL RISK AND ENTERPRISE RISK MANAGEMENT

Eleanor Toye Scott, Philip Stiles, and Pradeep Debata
Cambridge Judge Business School
University of Cambridge
Trumpington Street
Cambridge, CB2 1AG
England

Tel: + 44 (0)1223 339620

e.toyescott@jbs.cam.ac.uk
p.stiles@jbs.cam.ac.uk
p.debata@jbs.cam.ac.uk

We confirm that there were no conflicts of interest in the writing of this article.

Abstract

Artificial intelligence (AI) and machine learning (ML) offer organisations expanding opportunities for greater control and efficiency and more timely and accurate results, but at the same time bring escalating emergent risks. Two significant and complementary approaches to the organisational challenges posed by AI and ML are model risk management (MRM) and enterprise risk management (ERM). In this review we identify the key literature on technology risk and organisations and use it to consider how effectively MRM policies nested within an ERM approach can resolve the risk conundrum created by the growing complexity of algorithmic technologies. We develop here a framework of the elements of MRM and ERM and the links between them. We first look at MRM and highlight four areas of model development (data, design, implementation and performance) and their associated risks. We then consider ERM and how digital technology implementation affects the entire organisation. We highlight the need to move away from a measurement and compliance approach to risk towards a broader and more proactive approach, aimed at organising technology risk, to which our MRM/ERM framework contributes. We argue that careful attention to the roles, aspirations and incentives of human operators and other stakeholders will be critical in making this transition successfully. Our review has implications for future research in several areas, including the design of human-machine hybrid systems, development of organisational best practice in managing risks arising from algorithmic bias, the design of effective government regulation for artificial intelligence and machine learning, and the role of algorithms in regulatory regimes.

Introduction

As digital technologies and services have blossomed over the last two decades, a number of tensions at the heart of automation have come into sharp focus: our growing ability to design and build increasingly sophisticated algorithmic systems creates huge potential for gaining greater control over our environment through larger datasets, more accurate and timely results, greater flexibility and scalability and the ability to solve more challenging problems; but at the same time, there is an apparent loss of human control, because as the technology becomes more complex, it often also becomes less transparent and accountable. Second, the widespread use of advanced technology, big data, predictive analytics and automation promises gains regarding the way organisations manage risk. But emergent technology also increases the risks organisations face, due to issues of complexity, lack of transparency and susceptibility to cyber-attacks.

Business products and services, with decision engines which are driven by increasingly complex algorithmic technologies, require sophisticated models which may operate without day-to-day human intervention, but which do require effective *model risk management* (MRM) to ensure that products and services are produced and/or delivered as intended and that the business of the organisation and its stakeholders are protected from costly errors (Kosoff, 2016; Shi, Young, Lantsman, & Wei, 2016). Model risk includes the development, validation, implementation and on-going monitoring of each model and its place in the general model portfolio, as well as the risks arising from the model portfolio in aggregate. MRM also needs to be integrated within the firm's existing overarching risk management approach, or its *enterprise risk management* (ERM) so that it can effectively interact with various other business risks that the firm faces which need to be managed, monitored and controlled. However, research on how these two forms of risk management combine is scarce. In this article, we develop a framework of the elements of model risk management and enterprise risk management and the linkages between them.

This is important as the increasing scale of digital technology use entails that problems with technology risk and governance models can be highly significant. For example, high profile failures such as the UK 2020 A level algorithm scandal (Haines, 2020; Hern, 2020), a racially biased algorithm used in the US criminal justice system (Dieterich, Mendoza, & Brennan, 2016; Larson, Mattu, Kirchner, & Angwin, 2016), and the Facebook

Cambridge Analytica scandal (Cadwalladr & Graham-Harrison, 2018; HouseofCommons & DCMSCommittee, 2019) have highlighted the concern over inadequate risk and governance approaches within digital technology introduction.

A recent review of the field argued that "to date, most research in management and economics has emphasized the benefits of using algorithms to improve allocation and coordination in complex markets, facilitate efficient decision-making within firms, and improve organizational learning" (Kellogg, Valentine, & Christin, 2020: 366). There is good evidence to suggest that algorithmic models can have a positive impact on the efficiency and innovation of organisations (e.g., (Athey & Stern, 2002; Hall, Horton, & Knoepfle, 2021; M. Liu, Brynjolffson, & Dowlatabadi, 2018a). However, risk management and governance methods for such applications are not yet well-established, leading to problems like algorithmic bias, lack of transparency, poor data management, proliferation of misinformation and increased vulnerability to cyber-attack (CDEI, 2020; Cresci, 2020; Doshi-Velez & Kortz, 2017; Hackett, 2015; He, Frost, & Pinsker, 2020; J. R. Jackson, 2018; WEF, 2020).

Discussion of risk often appears in predictions for societal level outcomes, such as the impact on employment and labour markets (Autor, Levy, & Murnane, 2003; Butcher, 2013; Campa, 2019; Goos & Manning, 2007; McIntosh, 2013), surveillance (Cobbe & Morison, 2019; Myers West, 2019; Zuboff, 2015, 2019), and the nature of capitalism (Manzarolle and Smeltzer 2011, Gillespie 2014, van Dijck 2014). While these bodies of work are important, we focus here on the issue of how risk plays out *within* organisations as new algorithmic capabilities are introduced, and how these new capabilities interact with organisational governance and culture.

We respond to calls for "further work to develop understanding of how risks are managed for the entire organisation" (Bromiley, McShane, Nair, & Rustambekov, 2015: 317) by considering the following question: how effectively MRM policies nested within an ERM approach can resolve the risk conundrum created by increasingly complex algorithmic technologies.

The review is organised as follows. In the next section, we examine the principal areas of technology and risk. We then explore the issue of model risk management and highlight the four aspects of our guiding framework and the risks which arise in connection with each aspect. Through this analysis we present a case that designing, implementing and

managing models effectively requires as much attention to human, organisational and governance factors as it does to the design of the algorithms themselves. At each stage, we identify relevant debates, highlight the risks attached to the activity and give illustrative case studies. In the second part of the literature review we highlight enterprise risk management and how technology implementation affects the entire organisation.

Technology, organisations and risk

Technology is integral to all organisations and pervades organisational cultures, structures and processes. Research on risk in technology at the organisational level has tended to focus on technology project risk, highlighting the design and implementation of IT and IS (Jørgensen & Jordan, 2016). The body of work predominantly highlights risk analysis, which treats risks as quantifiable and seeks "to calculate risks - typically in statistical terms as the probability of an event multiplied by the magnitude of the resulting gains or losses – so that appropriate actions [can] be taken." (Hardy, Maguire, Power, & Tsoukas, 2020: 1033). Risk management follows once the risk has been assessed, and, according to the Society for Risk Analysis, is concerned with "exploring opportunities on the one hand, and avoiding losses, accidents, and disasters on the other" to achieve "the proper risk level" (SRA, 2019: 5). Such judgements determine the risk "tolerance" or "appetite" of the organization (Bromiley et al., 2015; Power, 2009). With emerging technologies, however, risk maybe uncertain and so mechanisms for observation and monitoring may be also unclear. Further, as scientists and industry may not agree, either on methods or on data for assessing potential risks and consequences, various interpretations of the science may emerge, together with ambiguity and possible divergent perceptions of risks and benefits associated with the technologies (Falkner & Jaspers, 2012; Renn, 2011).

Although complex systems will give rise to risk incidents, in line with "normal accident theory" (Perrow, 1984), the proliferation and scale of technology deployment may give rise to greater incidence of normal accidents (Agarwal, Argarwal, Kayyali, & Stephens, 2020). Technological risks are becoming more prominent within organisations, and more dangerous, primarily because of the scale, complexity and interconnectivity of devices and models (Bevan, Ganguly, Kaminski, & Rezek, 2016; CDEI, 2020).

To address such issues, research has focused on two broad areas: (1) Design and development principles concerned with the creation and implementation of specific pieces of technology or models. This would relate to the quality of data used in the models, the quality

of the model specification, the accuracy (and avoidance of bias) of the model and the adaptiveness and system protection of the model. This is often bracketed under the heading "model risk management" (Brotcke, 2020). (2) Controls over the models and their aggregation into the organisation-wide risk governance framework. This is usually labelled "enterprise risk management" which is concerned with the systematic and holistic nature of risk within an organization and how particular models or technologies integrate with the wider set of technologies within the organization (Bromiley et al., 2015).

This approach is in line with recent work on risk governance, which describes risk governance as "the confluence of all analyses and actions relative to the development of a given technology", (Renn, 2008). This would include

(i) framing the technology in the context of its possible deployment and applications, benefits, and risks for various stakeholders;
(ii) assessing those benefits and risks (including assessment of perception and concerns);
(iii) evaluating other aspects that decision makers will consider before making decisions, such as the existence of specific economic, political or societal interests, or also certain issues of national security or ideology, that must be considered;
(iv) identifying various risk management options, which can be combined to establish a strategy for the development (or not) of the technology; and
(v) communicating about risk and benefits (Renn, 2008).

Figure 1 shows how we conceptualise model risk as a subset of operational risk, which is one of a (non-exhaustive) set of risk areas managed at enterprise level. Model risk is further classified into individual model risk, i.e., the areas of risk which must be assessed and managed with respect to a single model, and aggregate model risk, which concerns the additional risk arising from the whole model portfolio.

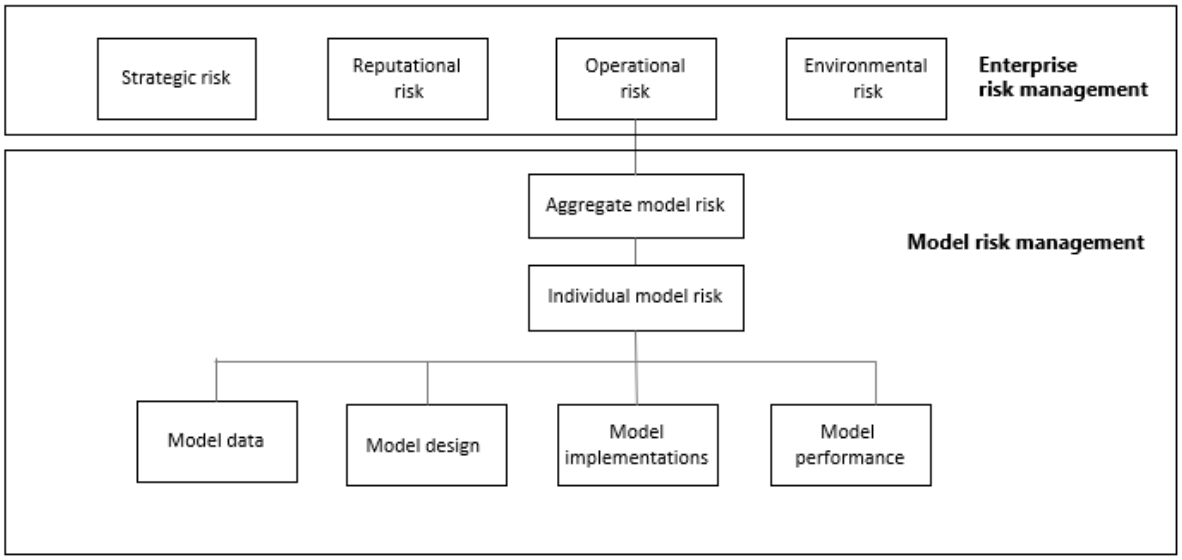

Figure 1. Model Risk Management in relation to Enterprise Risk Management

2.1 Model risk management (MRM)

MRM is most developed in the financial services industry. This industry, having faced a systemic collapse in 2008 which was only prevented by state intervention, has since been addressing the origins of that crisis and reappraising and updating its approach to the downside of risk (Arner, Avgouleas, Busch, & Schwarcz, 2019; Haag et al., 2010; Palermo, Power, & Ashby, 2017). As early adopters of advanced technologies, financial services companies and their regulators are now at the forefront of risk management practice (BankofEngland, 2019; Cohn, 2019; FSB, 2017; IIA, 2013).

According to a recent report from McKinsey, the number of models in use at large financial institutions is rising as much as 10 to 25 percent annually (Crespo, Kumar, Notebook, & Taymans, 2017), as organisations use increasingly advanced analytics techniques, including machine learning, to achieve higher performance standards. A typical model for a bank might be, for instance, a specified calculation for assigning consumer credit scores, into which an individual's data could be entered, the calculation run, and a decision provided on whether the loan would be granted. The promise and wider application of models have brought into focus the need for an efficient MRM function, to ensure the development and validation of high-quality models across the whole organisation.

Financial institutions typically create a policy definition of a model based on the mathematical approach taken to solving business problems, usually broad enough to meet the regulatory expectations of SR 11-7 but narrow enough not to include every spreadsheet used

at the bank (Bridgers, Lee, & Kosoff, 2020). SR 11-7 and another key US banking regulation, OCC 2011-12, treat the term "model" as any quantitative method, system or approach which applies statistical, mathematical, economic or financial theories, assumptions and techniques to process input data into quantitative estimates (Yoost, 2013). According to Hill, "the great majority of financial models in use today remain essentially static collections of… code… designed to perform essential and well-defined computational tasks" (Hill, 2020: 26), although he sees an increasing role for AI/ML in the next generation of smarter models.

In applying the concept of MRM to technology risk in industries outside financial services, we define a model pragmatically as any attempt to predict or forecast what will happen given a specific set of variables and a relevant set of data. For an airline, this might mean modelling seat prices and loads on aircraft, for a restaurant it might mean predicting patterns of customer demand for different menu items, while for a telecoms operator it could mean predicting how many faults may arise under certain weather conditions. These processes are becoming less likely to be based on human judgement, and even automated processes are now less likely to be directly supervised by a human operator, instead perhaps diagnosing their own faults, or perhaps reporting to and overseen by an algorithmic "manager". As a result, new opportunities for efficiency and growth arise, as do new types of risk.

Model risk is only one part of technology risk, which can include cyber-attacks, unstable systems requiring expensive down-time, vendors failing to deliver a service, legacy systems that are difficult to maintain or integrate with newer systems, but which may be impossible to discard, outdated hardware, technology being used for purposes that it was not designed for, and so on. These other technology risks will be discussed in relation to enterprise risk management.

2.2 Enterprise risk management

Authors and regulators disagree on exactly what constitutes ERM. Addressing the variety of definitions and implementations of ERM, Power (2007) urged caution, asserting that ERM is an "umbrella concept" and managers should not "...assume that ERM refers unequivocally to a coherent set of practices." As regulators pressure organisations to integrate risk management into corporate governance, new risk categories and definitions have been created, leading to the "risk management of everything" (Power, 2004) which Power ultimately concluded had resulted in the "risk management of nothing" (Power, 2009).

Power's scepticism about the proliferation of new risk management structures, categories and processes is understandable, particularly in the immediate aftermath of the 2008 financial crisis. However, Power and colleagues have recently taken a more moderate and cautiously optimistic view (Hardy et al., 2020), suggesting that while it may be impossible to calculate and manage precisely every individual risk within an organisation, it is nevertheless worthwhile to consider how risks are *organised*; a process which, they argue, "involves far more than calculating them before they arise; it also means containing risks that do arise and reflecting on how to improve how they are organised in the future" (Hardy et al., 2020: 1033). This process of organising risk, we suggest, constitutes the main business of ERM.

Research on ERM, then, is in its infancy but several core issues have been identified (Bromiley et al., 2015): First, risk is most efficiently managed at the level of the enterprise rather than at the level of individual parts of the business. Second, major decisions at a strategic level by an organisation involve risk and must be managed. The competency in managing such risks may give an organisation a competitive advantage, allowing it to improve its position over rivals.

Organisations by nature manage risks and have a variety of existing departments or functions ("risk functions") that identify and manage specific risks. However, each risk function varies in capability and in how it coordinates with other risk functions. A central goal and challenge of ERM is improving this capability and coordination, while integrating the output to provide a unified picture of risk for stakeholders and improving the organisation's ability to manage the risks effectively.

ERM is evolving to address the needs of various stakeholders, who want to understand the broad spectrum of risks facing complex organizations to ensure they are appropriately managed. Regulators and debt rating agencies have increased their scrutiny on the risk management processes of companies.

For our review, we turn first to model risk management.

1. Technology and model risk management

The design and implementation of individual algorithmic models can be conceptualised as a stage process, comprising four key stages: Data, design, implementation and performance. There is also an overarching requirement for management of the aggregate model risk arising from the model risk portfolio, which applies at all four stages of model

development and use. As shown in Figure 1, MRM can be considered a subset of operational risk, which is one of a range of risk areas managed at enterprise level.

Models are dependent on the availability of suitable **data**. Organisations may develop a model to exploit data that they already have, or they may have an idea for a model but need to gather the required data. Either way, data acquisition and management are critical parts of the model development process. **Design** encompasses an organisation's awareness that existing models require updating, the specification of goals and requirements for a new model, prototyping, "sandbox" testing to create a proof-of-concept, and in some cases, small-scale pilot live deployments to get user feedback and debug any issues that arise in a real-world environment. **Implementation** is the process of deploying the model at scale in the real world. At this stage, the model needs to be integrated carefully into its environment, and any arising issues dealt with. **Performance** must then be monitored and managed over the lifetime of the model, including regular recalibration and audits to ensure that the model continues to be fit for purpose.

Model risk management is a process by which the risks associated with these stages and their overall combination are assessed, for each model, and for the organisation's portfolio of models. In this section, we examine the four stages and their associated risks, followed by a short overview of model portfolio management.

3.1 Stage 1: Data

Organisations today typically have access to very large quantities of consumer, employee and market data – so-called Big Data. Indeed, it was only after these very large data pools became available as "data exhaust" from the Internet in the 2000s that ML algorithms (conceived in the 1980s) finally had sufficient learning power to achieve anything useful (Gray & Suri, 2019; LeCun, 2018; Marcus & Davis, 2019; Myers West, 2019; Zuboff, 2015). The current business environment and the unfolding technological possibilities create new opportunities, with attendant risks, for organisations which have access to large quantities of consumer and employee data, not least through the development of models, including ML models, for prediction and decision-making.

Figures 2a and 2b below illustrate a possible organisational flow through the process of defining, acquiring, testing and protecting data to be used in an algorithmic prediction or decision model. Figure 2a gives an overview of the end-to-end process, while Figure 2b

illustrates the tests which need to be applied to determine whether the available data is adequate for use in the model.

1. Principles	Specify data definitions	Specify logical, ethical and regulatory limits to use
2. Viability tests	See Figure 2b. Proceed if data passes viability tests	
3. Protection	Specify who can enter, maintain and edit data	Specify and implement mechanism for **logging data changes**
	Specify and implement **authorisation procedure** for data entry and editing	Specify and implement procedure to **regularly identify and safely discard** unneeded/unauthorised personal data
	Specify and implement **other security procedures** as needed to prevent human- or machine-caused data corruption	

Figure 2a. Data considerations

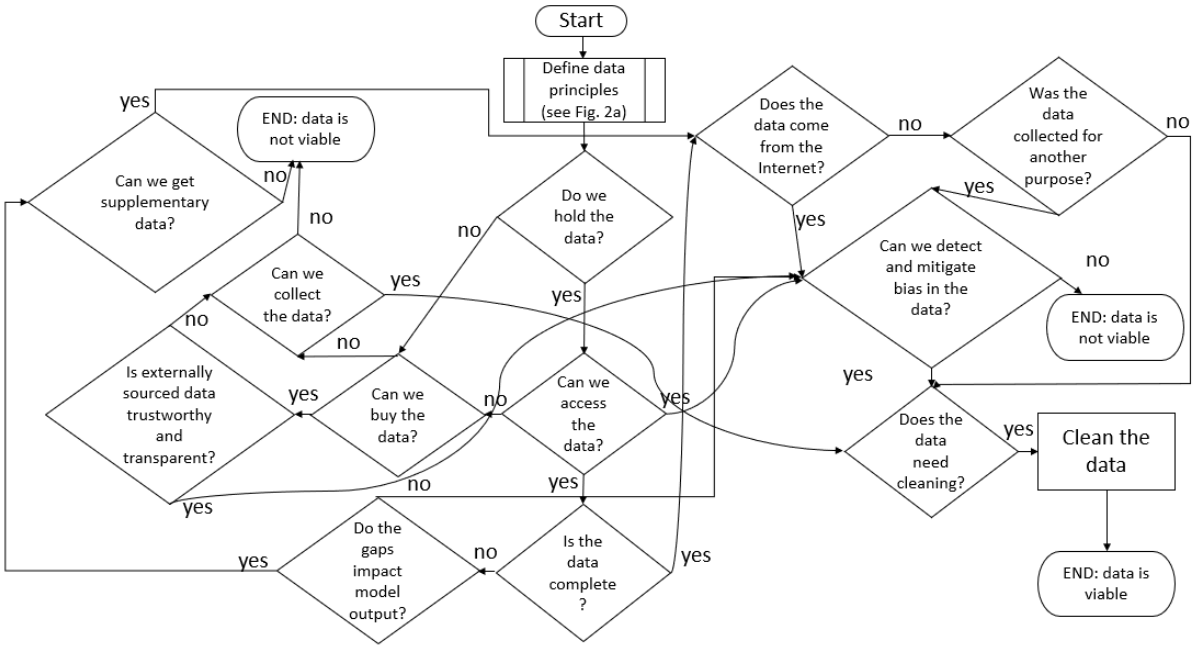

Figure 2b. Data viability tests

3.1.1 Data risks

3.1.1.1 Data accessibility

In practice, organisations are sometimes unable to exploit useful information that they hold because of organisational, regulatory or IT infrastructure limitations. For example, a large North American hospital developed a prototype AI-enabled system to improve patient care, but were unable to implement the project because the data required was distributed

across twenty legacy systems, and retrieval would be too complex (Vial, Jiang, Giannelia, & Cameron, 2021). It is therefore critical that early consideration is given to how data will be accessed at scale when the model goes live. Vial, Jiang et al. (2021) point out that data accessibility is often treated as a technical problem which can be addressed in the later phases of the project life cycle, but in fact it is a complex management problem which happens to involve technology. Failing to address data accessibility at an early stage can result in unexpected extra costs or even jeopardise the project.

3.1.1.2 Data acquisition

When an organisation acquires personal data, whether directly from customers or employees, or by buying a database from a third party, data protection law requires that records are kept demonstrating what individuals have consented to, what they were told, and when and how they consented (ICO, 2021a, 2021b). In practice, however, many organisations are still failing to obtain adequate consent for their data-gathering activities (CDEI, 2020).

3.1.1.3 Data sharing and sale

There is a longstanding trend towards monetising data, established by Google, Facebook and other large Silicon Valley companies (Zuboff, 2015, 2019), which means that there is a large market for data. For instance, a 2018 Gartner survey of IoT projects showed that 35% of respondents were either already selling or planning to sell data collected by their products and services (Preimesberger, 2018).

In this context, public sector organisations in particular may fail to fully appreciate the value of the datasets they hold (for example, medical records), and sell them or share them with commercial partners or suppliers at a fraction of their true value (CDEI, 2020). Conversely, some organisations may be unwilling to explore data-sharing arrangements because of the complexity of data protection regulations and may miss out on valuable opportunities to collaborate with external parties.

3.1.1.4 Data quality and hidden data bias

Key determinants of data quality are accuracy, completeness and timeliness, as well as accessibility by machines (Vial et al., 2021). In addition, the data must be targeted to the use case, since this reduces the quantity of data required and improves the chances that the algorithm will learn to perform the required task accurately (Azulay, 2019). Poor-quality or

unsuitable data may compromise the ability of an algorithmic model to produce useful results.

A more insidious problem is the existence of undesirable and undetected bias in data sets. For example, data sets used by face recognition algorithms have been found to be over-representative of white male faces, causing poor performance on recognition of black and female faces, with worst performance on faces which were both black and female (Buolamwini & Gebru, 2018). In the context of employment, algorithmic models used to sift CVs and select possible applicants for short-listing can suffer from a similar problem – if models are trained only on historical data about who has been hired in the past, this can unwittingly entrench existing bias towards employing men and white people, and undermine attempts to improve the gender balance and ethnic diversity of new hires (O'Neill & Mann, 2016). Therefore, training data for these algorithms need to be adjusted to be more diverse and representative of the wider population.

However, where algorithms are operating on data drawn directly from the Internet, it may be much more difficult to find and correct the sources of bias in the data. One research group attempted to show an Internet advertisement for STEM careers to men and women in equal numbers but found that the ad was still shown to more men than women, for reasons which were initially unclear (Lambrecht & Tucker, 2019). Eventually the authors discovered that Internet advertising to women is more expensive because they tend to be more responsive, and that the ad-serving algorithm was optimising for cost. Data bias is thus a difficult problem to detect and solve; but organisations must address it, both for ethical reasons, and because otherwise they face reputational risk as well as the risk of failing in their legal obligations to avoid unfair discrimination. (Bias is discussed further in the Case Example for this section, and in the Design and Implementation sections below.)

3.1.1.5 Data integrity

Data integrity covers system control and access (who can access the data, for which purposes, and how is access tracked and logged), data protection, controls and regulatory compliance (focusing on defining the data and restricting the ability to delete or modify it), and audit trails, metadata and review (focusing on completeness and appropriateness of audit trails and required reviews of data and metadata).

In highly regulated industries like pharmaceuticals, the need for data integrity is well-established, but in practice even in this environment, companies sometimes struggle to ensure

data integrity requirements are met – Lippke et al. (2020: 51) report that from 2014 to 2016, inspections of pharmaceutical facilities revealed "a pattern of repeated failures to follow data integrity requirements". As a result, data integrity has become an area of increased regulatory focus across the pharmaceutical industry. And as other industries and public services become more data-focused and deal with larger quantities of data, data integrity is becoming a more widespread concern.

3.1.1.6 Data breaches

As the volume of data which organisations are gathering has mushroomed, so too have the burdens of managing it responsibly. Breaches of organisation's databases are very serious events, which can affect millions of individual customers or employees, attract significant fines from regulators, and damage a company's reputation and share price – for example, consumer credit agencies Experian and Equifax have both suffered major breaches in the past decade, and British Airways was fined £183 million by the regulator in 2019 for a breach of their customer database (Gressin, 2017; Hackett, 2015; Sweney, 2019). However, many organisations, particularly those with low digital/data maturity, are retaining personal data which they do not need, placing their customers' and employees' data at unnecessary risk (CDEI, 2020).

3.1.1.7 Limitations of regulation

While GDPR is helpful in providing guidelines, it also poses significant challenges for organisations of all types and may at times hamper innovative and socially valuable projects. The costs associated with compliance with GDPR are large – one survey is reported to have estimated that the minimum cost of compliance for anyone doing business with any EU resident is $1 million just for changes to IT systems (Downes, 2018). Organisations which analyse customer data may be prevented by GDPR from outsourcing analysis functions or sharing data for research purposes or may find the costs of compliance for these activities too high to be worthwhile, and this may impose unnecessary limits on what they can learn from the data.

Another concern is that like all regulations, the GDPR rules are static until they are revised; but they were drafted in a rapidly changing technological context, which may already be unable to cover the ways that data is being shared with Internet of Things (IoT) devices (Sullivan, 2018). There is thus a risk that new data-related activities may either be

unintentionally left unregulated or alternatively, may be unnecessarily banned or restricted by regulations intended for a different purpose.

3.1.1.8 Trust and reputational risks

Despite GDPR, there are still risks around loss of personal privacy and autonomy, as organisations gather ever more detailed and holistic data on their employees, service users and customers, and use the information to shape their behaviour. While in the short-term, closer organisational control over employees and greater influence over customers is likely to boost profits, over time it can also erode trust, and may bring the organisation into suppressed or open conflict with its employees (Kellogg et al., 2020). Employee mental health and experience of work suffers from reduced autonomy and increased monitoring (Brione, 2017; Hannif, Cox, & Almeida, 2014), and organisations lose many of the benefits of employee initiative, goodwill and collaboration (Gray & Suri, 2019). The risk of losing customer and service user trust through data collection and analysis is perhaps more subtle, but nonetheless can cause companies significant damage, as both US retail chain Target and Facebook can testify (Cadwalladr & Graham-Harrison, 2018; Duhigg, 2012; OxfordAnalytica, 2018)

3.1.1.9 Systemic and market risks from data monopolies

The Cambridge Analytica/Facebook scandal in 2018 demonstrated both the value and scale of personal data held by Facebook and the risks arising from this dataset to individuals and society (Cadwalladr & Graham-Harrison, 2018). Arguably, Cambridge Analytica's abuse of so many users' Facebook data was only possible because Facebook is effectively a worldwide monopoly. The CDEI report notes that "…platforms have created and aggregated huge datasets that can be used to generate increasingly sophisticated inferences and insights about people. These datasets have enormous potential value beyond the contexts they are collected in… although the concentration of data within a small number of organisations creates system and market risks." (CDEI, 2020: 85). A similar problem exists in financial services: both the CDEI report (2020) and the Financial Stability Board report (2017) highlight the possibility of increased systemic risk in the financial sector because of large, varied and high-quality data sets being concentrated in the hands of a small number of companies.

3.1.1.10 Failure to realise return on investment

There is a risk that organisations devote increasing resources to big data analytics but fail to develop and sustain a competitive advantage from their investment. As Grover and

colleagues point out, a 2016 Gartner survey revealed that as big data investments continue to rise, results can be disappointing, with many firms struggling to achieve insights that made a real difference (Grover, Chiang, Liang, & Zhang, 2018; Heudecker & Hare, 2016). One of the reasons for this may be low data quality. Where organisations are dealing with incomplete or low accuracy data, or very broad and varied data which is inappropriate for their specific use cases, analysing even large volumes may fail to deliver business value (Azulay, 2019; CDEI, 2020). Data accessibility at scale can also be a challenging problem, as noted above (Vial et al., 2021). Finally, organisations may be struggling to recruit data scientists and other professionals with the skills to exploit the data they have at their disposal (Bayer, Sandy, Schrader, & Spiegl, 2021).

3.1.2 Case example

Racial discrimination is an unintended but frequent component of algorithmic model outputs based on Internet data. A study by Latanya Sweeney showed that online searches for names associated with Black people were more likely to bring up ads for public record checks than searches for names associated with White people. In cases where public record check ads did appear, they were much more likely to suggest that a person with a typically Black name had been arrested or had a criminal record than when they appeared for typically White names, regardless of whether the person concerned actually had such a record (Sweeney, 2013). This shows just one way that data bias embedded in the Internet can routinely disadvantage those from marginalised communities, by reinforcing negative stereotypes and unfairly arousing suspicion against individuals with certain backgrounds whenever anyone else is interested in finding out about them – whether in the context of a job search, dating, volunteering, media interest, or any other circumstances which arise.

While these outcomes have not been consciously planned by algorithm creators, they reflect both the historical data on which they are trained, which is shaped by societal biases, and the blind spots and unconscious prejudices of algorithm developers. Unfortunately, if an ML algorithm learns from biased data, it acquires the same biases in its predictions, decisions and recommendations. These biased model outputs then reinforce the same societal biases which caused the data bias – partly because people see the patterns they expect to see, and partly because people often take an algorithmic model's output on trust as being neutral and objective (J. R. Jackson, 2018).

There is little that can be done directly to eliminate societal biases in Internet data, but organisations need to be aware that these biases are likely to affect the outputs of, for example, ad-serving algorithms on the Internet; and to find ways to avoid or at least detect and counteract such bias. The bottom line, as Sweeney points out, is that if an employer disqualifies a job applicant solely on information indicating an arrest record, the company could face legal consequences under anti-discrimination laws.

3.2 Stage 2: Design

All organisations pursue their goals in a changing context, both internally and externally, which they must review for opportunities and threats, and update their offering to their service users, clients or customers accordingly – and with the ever-faster pace of change in technology and society, this process of horizon scanning, internal review and updating is now continuous in many industries (Agarwal et al., 2020; Tarafdar, Tu, Ragu-Nathan, & Ragu-Nathan, 2011). This results in the continuous creation of new decision models, either for new products and services, as is typical in financial services, where new loan terms, insurance policies or customer account terms are constantly being offered, or for operational processes within the organisation such as stock control, purchasing or forecasting of seasonal demand fluctuations.

Purpose	Is a model needed?	What is the goal of the model?	
Stakeholders	Who will use the model?	Who will be affected by the model's decisions?	
Design approach	What theories and assumptions underpin the model?	How will the model be developed and implemented?	
Standards and regulations	What standards (including regulatory) must the model meet?	Safety Accuracy Explainability	Security Fairness Usability
Monitoring and validation	How frequently will the model be monitored and revalidated?	And at what level of detail?	

Figure 3. Design considerations

3.2.1 Design risks

While the term *model risk management* is rarely used outside financial services, an increasing number of industries and sectors are reliant on models, and there are a range of

associated model risks which can cause serious problems unless they are identified and avoided or mitigated at the design stage.

3.2.1.1 Safety and security

Safety and security are critical aspects of models and must be assessed in relation to all relevant stakeholders, including anyone working directly with the model, any client, customer or service user affected by the model's decisions, the organisation itself, and in some cases, the public. For regulated industries, consideration of safety and security issues is likely to start with compliance with the regulations (NAO, 2017; Stanley & Wdowin, 2018) – but must extend beyond compliance to be effective (Kaplan & Mikes, 2012a). Model safety issues are most likely to arise at the compliance level and it is important to ensure that these risks are addressed on a model-by-model basis (Kosoff, 2016). At the same time, interdependencies between models also need to be recognised and addressed at the model portfolio level, and in the wider context of enterprise risk management (Bridgers et al., 2020; Shi et al., 2016).

Networked technologies offer huge potential but can also create new cyber-security vulnerabilities. Proposed models therefore need to be assessed on whether their implementation might offer new opportunities to hackers to steal data or to hijack or damage the organisation's facilities (Brands, 2020; CDEI, 2020; Dzinkowski, 2019; McCollum, 2020; WEF, 2020).

3.2.1.2 Accuracy

Models offer significant potential for improving accuracy of predictions and judgements relative to traditional decision-making processes or systems, especially when they include ML algorithms. McKinsey highlights improved resource-allocation decisions and greater efficiency as typical benefits from using ML (Dharasathy, Jain, & Khan, 2020).

However, the authors argue that measuring model accuracy is not necessarily straightforward, since for any moderately complex model there are typically many possible measures, as well as probabilistic outcomes. Therefore, they recommend identifying typical use cases and consulting with users, to select two or three measures which matter most for those use cases. The same article points out that where an automated model is being introduced to replace a previous system, whether automated or based on human judgements or scoring systems, it is important to establish the baseline accuracy of the current system, against which the performance of the proposed new model can be compared.

Accuracy may also need to be traded off against other requirements like fairness and explainability. In a specific context, the model which is most accurate overall may be less equitable in its outcomes for some sub-groups (Chowdhury, 2018). The need to generate useful explanations for various purposes may also affect the accuracy of models (Doshi-Velez & Kortz, 2017). In all cases, an adequate level of accuracy must be determined in relation to the context and the other priorities for the model.

3.2.1.3 Fairness

Models which are based on ML algorithms often run into difficulties with undesirable emergent bias. However, even deterministic and relatively transparent algorithms can produce damagingly biased model outputs if fairness requirements have been poorly specified, or the algorithm has not been adequately tested - as happened with UK exam regulator Ofqual in the summer of 2020 when they attempted to allocate A level grades to a cohort of students who had been unable to take their exams because of the coronavirus pandemic (see case example below).

There are several possible approaches to defining and enforcing fairness for algorithmic models, some of which have drawbacks and all of which require careful application in context. Dharasathy et al (2020) identify and discusses three options:

1. *Wilful blindness* to sensitive category data such as sex, race or other socio-economic factors. Creating a model which is merely unaware of factors related to disadvantage and discrimination can still lead to unfair outcomes, including the emergence of proxies for protected characteristics, or can cause issues with the sample data used to train the model itself.
2. *Demographic or statistical parity* in the outcomes of the model – for example, in a credit-scoring model, by ensuring that an equal proportion of loan applicants are successful from groups with and without protected characteristics. It is most likely to be effective in situations where only a single measure of fairness is being monitored, such as ensuring that the same proportion of loan approvals go to men and women respectively.
3. *Predictive equality* is an attempt to ensure that the model performs equally accurately across different groups, i.e., that the error rates for each sub-group are the same, even though the outcomes for each sub-group may be different because of other factors. In the credit-scoring example, the same proportion men's and women's

loan applications may not be accepted, but the percentage of approved applicants who end up defaulting (the false positive rate) would be the same for each group.

Given the nuances and trade-offs involved in defining fairness for a given context, it is also especially important that the teams tasked with developing algorithmic models are themselves diverse (Carini, 2019; Nunn, 2018; Snow, 2018). As we noted above, ML algorithms frequently operate on data sourced directly from the Internet which contains structural biases; understanding and solving the resulting problems is complex (Chowdhury, 2018; Guynn, 2019; Lambrecht & Tucker, 2019). Diverse development teams are perhaps more likely to spot the potential for biased outcomes, and more motivated to find solutions, since they are more likely to be adversely affected personally.

3.2.1.4 Explainability

The advent of machine learning (ML) algorithms as a ubiquitous business tool means that we are for the first time in an era when machines can autonomously adjust their goal-directed behaviour in the light of their own experience. This self-learning by machines in no sense replaces human reflective and adjustment activity (most notably in the choice of training data), nor does it directly replace human control over machines. However, it creates new challenges for the oversight of algorithmic models, especially around explainability.

Unlike deterministic algorithms, for which the programmer specifies every calculation directly, ML algorithms are usually “black boxes”, i.e., because of the way they store information distributed across many layers of nodes, no information is available about what makes them arrive at their predictions. All that is available to the users of the model is information about the inputs and the outputs (Doshi-Velez & Kortz, 2017; Marcus, 2018a; Samek, Wiegand, & Muller, 2017). As ML applications are more widely adopted by businesses and governments for critical real-world applications, understanding what they are doing and why will increasingly become a top priority. Samek, Wiegand et al. (2017) identify four major reasons why transparent ML algorithms are important:

- **Verification of the system** – to establish that an application’s prediction accuracy is not an artefact of reliance on causally unrelated but coincidentally correlated data. This is essential for healthcare and other safety-critical applications.

- **Compliance with legislation** – Courts require explanations of events and actions to reach fair judgements, so future AI systems will need to become more explainable.
- **Improvement of the system** – If we can interrogate a system, we can discover weaknesses in its processes or biases in its model or dataset. We can also compare different models and architectures which may achieve similar accuracy by relying on analysis of different data features.
- **Learning from the system** – If ML algorithms discover better ways to do things, it would be useful to understand their reasoning or analysis process. In the sciences, AI systems may acquire new insights about natural laws and processes, which human scientists would benefit from understanding.

Explanatory transparency is certainly not required (or possible) for every aspect of a model's functioning – but it must cover situations where the operator needs a background understanding of how it works, or real-time information about what the system is currently doing, as well as situations where post-hoc explanations are required, either as part of a regular audit of how a model has been behaving, or after something has gone wrong. Different types of inquiry and explanation will be relevant to different applications and audiences and for different purposes (Anderson et al., 2003; Doshi-Velez & Kortz, 2017; Rai, 2020; Ribeiro, Singh, & Guestrin, 2016). However, there are some common features which useful explanations share. Doshi-Velez and Kortz (2017) suggest that a useful explanation should be able to answer at least one of these three questions:

- What were the main factors in a decision?
- Would changing a certain factor have changed the decision?
- Why did two similar-looking cases get different decisions or vice versa?

Because of the inherent opacity of many ML models, Doshi-Velez and Kortz recommend that separate "explanation engines" be designed to unpack and present their logic for human audiences. Designing algorithms which meet these varied requirements for explanation presents some significant technical challenges, which are starting to be addressed by the machine learning community (Doshi-Velez & Kortz, 2017; Rai, 2020; Ribeiro et al., 2016; Samek et al., 2017). Doshi-Velez and Kortz (2017) and Ribeiro et al. (2016) both discuss the problem of making explanations *interpretable* to humans, i.e. that they provide qualitative understanding of the pathway between the input variables and the response.

There are trade-offs to be made when designing explanation systems. Doshi-Velez and Kortz (2017) and Rai (2020) both mention balancing the need for ML applications to be accurate against the need for them to be explicable: generating explanations has a cost in terms of time and resources, and decisions will need to be made about when explanations are necessary, and when they would slow down systems or degrade their accuracy to too great an extent to be valuable. Ribeiro et al. (2016) discuss the more subtle issue of trading off the accuracy of each *explanation* against its interpretability – since more accurate explanations are likely to be less readily interpreted. In both cases, the trade-offs would need to be decided based on the context of use and the needs of the users.

Explanation algorithms are increasingly necessary and will play an essential role in making ML applications more transparent, both to manage what should be preventable risks (like being unable to explain an ML application's behaviour to a regulator or court) and to get the right balance of explainability, efficiency and accuracy from the application to meet business goals while minimising strategy risks. At the same time, expert human "explainers" will continue to have a critical role (Dharasathy et al., 2020; O'Neill & Mann, 2016; Wilson & Daugherty, 2018) . In practice, however, organisations themselves often lack transparency and may be either unable or unwilling to investigate or explain algorithmic decisions, with significant costs to themselves and their stakeholders – see the case example below.

3.2.1.5 Usability

Usability is an ongoing concern from the design stage through the implementation and performance stages, and under the Implementation heading below, we cover some practical examples of model usability problems arising at the implementation stage. However, users need to be centred in the design process from the start to avoid "baking in" clashes between assumptions in the model design and emerging user needs at a later stage (Brione, 2017; Kochan, 2019; Tarafdar et al., 2011). If the new model is very similar to ones the users are already using all the time, the user consultation process can be routine, but nonetheless needs to be embedded in the organisation's model development practice. If the new model represents a significant departure from previous practice, consultation needs to include all groups affected by the change, and could involve informal discussions, interviews, focus groups, surveys and questionnaires, as well investigation of background documents.

3.2.2 Case example

The UK qualifications regulator Ofqual faced serious difficulties in the summer of 2020 with the algorithm they used to calculate A level grades for a cohort of students who had not been able to take their examinations because of COVID19 (Hern, 2020; Mills, 2020). While Ofqual's interim report (Ofqual, 2020) on their method for calculating 2020's A level results indicates a strong concern for fairness, their focus was on eliminating biases caused by differences between schools and between teachers; surprisingly they did not consider their own algorithmic model (and decisions about when to use it) as a potential source of bias.

The model was designed on the reasonable assumption that teacher-assessed grades would be somewhat inflated and aimed to find a fair way to bring grades into line with previous and future cohorts' exam grades. To achieve this, teachers were asked to rank their students, and the ranks were then pegged to each school's previous exam candidates' grades. One outcome of this scheme was that excellent students in schools with historically low grades could not attain top marks. Teachers were also not allowed to rank students equally, resulting in some students of equivalent ability being awarded vastly different grades based on previous cohorts' results. Small classes and small subjects (usually in independent schools) were exempted from the algorithm entirely, because of lack of representative data – meaning that these students benefitted from their teachers' optimism about their performance, while other students were marked down by the algorithm.

This model was of course developed under crisis conditions for an unprecedented situation, in a relatively short timeframe, so mistakes were perhaps inevitable. One basic error, noted by machine learning lecturer and blogger Tom Haines (Haines, 2020) occurred in testing the accuracy of the algorithm's grade predictions: in 2020, teachers were asked to rank their students, but this was not something they had done in previous years, so there was no equivalent data for 2019 – so when the algorithm was tested for accuracy on the 2019 results, the ranks used were the actual ranks of the students in the final exams, rather than their teacher estimated rank, effectively giving the algorithm part of the answer it was supposed to calculate.

The problems with the A level results model could likely have been detected earlier if it had been subject to independent scrutiny before the A level results were announced, and Ofqual has been criticised for not accepting help which was offered at an early stage by the Royal Statistical Society (RSS) (Murray, 2020). Ofqual wanted the RSS experts to sign a non-

disclosure agreement before involving them, which the RSS was unwilling to accept, because of its commitment to transparency, and so Ofqual did not take up the RSS offer. However, Ofqual's reluctance to expose the details of its model to public scrutiny is understandable, and perhaps even justified, given the political sensitivity of the A level results and the personal consequences for thousands of families, as the unfolding of the crisis emphasised. Ultimately, the political storm caused by the fiasco led to the results model being abandoned and teacher-assessed grades being used instead, an arrangement which continued in 2021.

3.3 Stage 3: Implementation

At the implementation stage, prototype models are thoroughly tested by data scientists and data engineers and then adjusted as necessary to create a proof of concept. This may be followed by a small-scale live roll-out, involving software engineers and IT staff as well as members of business functions. This again must be monitored and adjusted by software engineers in the light of real-world feedback to create a minimum viable product, which can then be put into production by IT and business staff. If data availability for a scaled-up process has not been addressed at an earlier stage, this may now rear its head as a serious problem (Vial et al., 2021).

Task requirements such as a working definition of fairness and user needs, goals and beliefs should have been specified at the design stage. However, at the implementation stage, testing and user feedback is likely to reveal gaps and mismatches between the assumptions embedded in the model and the realities of the world. The implementation phase is therefore critical for spotting and resolving areas where the model does not work as well as expected.

1. Prototype testing	2. Assessing risk and scope of model	3. Prototype evolution	4. People and processes
Is the model's performance stable?	What types and severity of risk does the model generate?	Have the use cases for the model shifted during testing?	Who is responsible for testing, monitoring and reporting on the prototype's performance?
How does the model perform in a range of likely and unlikely scenarios?	What are the boundaries on the contexts and uses for the model?	Is the model fit for its current purpose?	Who approves the model to go "live"?
How does the model perform relative to challenger models?	What are the limits on the model's decision-making capacity?	In what ways will the prototype need to be scaled up for real world use?	What is the process for handing the model forward from data scientists and engineers to software engineers, IT staff and business staff?
How does the model perform relative to industry benchmarks?	How will limits and boundaries on model use be enforced?	Have the challenges of scaling up been adequately addressed?	How are these processes, and the model itself, documented?

Figure 4. Implementation considerations

3.3.1 Implementation risks

3.3.1.1 Data accessibility

As noted above in the Data section, it is non-trivial to scale up a prototype algorithm which has been trained and tested on simulated data so that it can operate in a real-world context. Data scientists and data engineers creating the prototype therefore need to engage with IT and business staff at an early stage to address this issue, so that it does not become an obstacle to the project at the implementation stage (Vial et al., 2021).

3.3.1.2 Stability

The performance of all models (and indeed of traditional processes based on human decision-making or risk scoring) becomes less stable over time, primarily because each model is developed using data which was collected prior to its own existence. Introducing a model to process information and make decisions inevitably changes the world in which it operates. Over time, these changes systematically affect the data the model is designed to process and thus the decisions it makes.

Further, major changes in the world can affect the relevance of the historical data used to train models. The COVID19 pandemic, for instance, has radically altered historical patterns of consumption, travel, work and engagement with public services, rendering many pre-pandemic models obsolete. Equally, changes within the organisation such as new policies, or external regulatory or legal changes, can also affect model reliability.

Therefore, all models need to be reviewed at regular intervals, as well as whenever events indicate that relevant data patterns are likely to have shifted significantly. To determine the frequency of regular reviews, McKinsey recommends back-testing models on old data at different historical distances from the present – for example, data from a year ago and from two years ago. If one-year-old data gives good results but two-year-old data does not, then the model needs to be reviewed and retrained roughly every 18 months (Dharasathy et al., 2020).

3.3.1.3 Usability

The literature we reviewed offers several examples of workplaces where, for a variety of reasons, models for decision support were not being used as management intended, or as employees had been instructed, with the result that neither the employees nor the models were working optimally (Bader & Kaiser, 2019; Lebovitz, Lifschitz-Assaf, & Levina, 2019; Russell, 2012). The models concerned all aimed to support human workers by processing input data about a patient or customer and producing a prediction about that patient or customer's situation and needs. These workers were professionals with established expertise and a commitment to their role and professional goals, which they felt was undermined to some extent by models imposed by technology that they were expected to use at work. In other words, they saw the model as creating risk to the quality of their work, which they sought to avoid or mitigate.

For example, in a US study, radiologists were provided with an ML tool which could analyse medical images to assess children's bone growth and to detect lung cancer and breast cancer (Lebovitz et al., 2019). The radiologists were keen in principle to use ML to relieve them of the more routine and time-consuming parts of their job; but in practice, the judgements of the ML model frequently conflicted with the radiologists' own judgements, which introduced doubt and ambiguity into the process, and forced radiologists to re-examine the images and question both the model's judgements and their own. Critically, the radiologists had no background information on how the model was trained, which data it had

been trained on, or the specialism and number of years of experience of the radiologists who had provided the training assessments. They also did not know anything about the algorithms the model was using, and they had no means to enquire into the logic behind an assessment to get an explanation. As a result, although the radiologists used the model to provide a "second opinion", this simply made the assessment process more time-consuming, but did not add any significant value to it, because in the absence of necessary background information on the AI tool, in most cases the radiologists (not unreasonably) had more confidence in their own professional judgement.

In this study, as in the others cited, predictive models were introduced into an organisation with the goal of structuring people's decisions at work in a way which would improve its quality and efficiency, with certain organisational priorities and values in mind. Unfortunately, in all cases, management wrongly assumed that having endowed the model with these priorities and values, the model would impose the same priorities and values directly onto employees' work practices. As these studies show, employees are likely to resist, manipulate or avoid systems and models which interfere with their personal and professional priorities and values embedded in their existing work practices – especially if the new approach has not been properly explained, is inflexible and/or cannot be enquired into.

To avoid this kind of outcome, Kochan (2019) and Tarafdar et al (2011) both stress that employers need to engage with employees early and establish honest two-way communication about the issues a proposed technology is designed to address and how it will integrate with employees' current expertise and existing work practices. In all cases, the best results will be achieved when the *mental* models (Johnson-Laird, 1983) of the expected users of the system, and the ways these may differ from the proposed decision support model, are well-understood by those designing and deploying it.

At the same time, it is necessary to adjust the algorithmic model at the implementation phase to support human operators, and to be supported by them where appropriate (Daugherty & Wilson, 2018; Wilson & Daugherty, 2018).

3.3.1.4 Maintaining skills, engagement and autonomy of human operators

It is critical to understand and address human factors issues which emerge at implementation, to avoid deskilling and reducing the autonomy of human operators. Reduced autonomy and deskilling may make human jobs less satisfying, for instance as reported by Brione (2017) in a case study of a Siemens factory, where factory workers lost the ability to

plan their work, because new automatic scheduling software instructed factory teams on the order to complete their workloads for maximum productivity. More seriously, once most of the work in a complex task is automated, the risk of human error may rise through inattention or inadequate preparedness for situations when the automatic system requests human intervention or simply fails. Airline pilots can find themselves in this position, occasionally with tragic consequences (see case example below). Again, the human contribution must be recognised, tested and adjusted as necessary, as part of the development of a proof-of-concept model.

3.3.1.5 Automation is simple, organisational change is difficult

Finally, automating a process which is currently integral to someone's job is likely to be disruptive to them and to the organisation in which they work. This is not a reason to avoid automating, but it is a reason to think ahead about the impact of introducing an automated model on staff, on their jobs, workflows and work relationships, with a view to optimising the organisational and stakeholder benefits of the automated process. In other words, the risks of the proposed process itself (the model risk) may be minimal or easily controlled, but the wider enterprise risk could be significant and must not be overlooked. This type of risk is discussed more fully in the second part of the paper, looking at ERM.

3.3.2 Case example

In an example recently publicised in the media, Goldman Sachs faced legal investigation in the United States over the ML model it was using to determine credit limits for the Apple Card credit card. David Heinemeier Hanson, a high-profile tech entrepreneur, complained on Twitter that despite his wife's higher credit score, he received a credit limit twenty times higher than hers. Customer service were courteous, but unable and unwilling to challenge the judgment of the model, to which they had no access. Apple co-founder Steve Wozniak joined the Twitter conversation to say that he and his wife had had a similar experience. Eventually Hanson's wife's credit limit was "bumped" to match his, but the algorithm remained unchanged. There was no suggestion that the model was designed to discriminate against women, but in practice it did, and there was no way of finding out why, or changing its conclusion (Fingas, 2019).

This example points to some model risk management failures, and indeed enterprise risk management failures at Goldman Sachs, which go well beyond the regrettable fact that their credit limit model unintentionally discriminated against women. We do not know

whether the company had taken steps before deployment to check that their algorithm was not reproducing societal bias against women or other protected groups, but whatever testing was done failed to detect the model's sexism, and perhaps other inappropriate biases. Secondly, customer service staff should have been trained to know that this kind of issue can emerge, and once the situation arose, they should have been able to escalate the issue to an expert ML team within the organisation. The expert team should then have been able to derive an explanation of the model's behaviour. If the algorithm could not be corrected immediately, the organisation should have suspended its use and reverted to other trusted methods of credit checking until the problem could be solved. Instead, there was no process in place, and customer services had no resources to deal with a problem they did not understand. When the problem was eventually escalated, the company did not know how to fix it, or have a fallback position while they dealt with it. As a result, it was exposed to a justified Twitter rant by a very high-profile customer followed by negative media coverage and legal trouble.

3.4 Stage 4: Performance

Once a model has passed all necessary tests, checks and pilot stages, it can be put into production. At this stage, the model is operating in real time on real-world current data, and so there is actual rather than hypothetical operational risk if the model malfunctions or the real-world data diverges significantly from the training data.

Model goal	**Is the model doing the task it was designed for?**	**Is this model the best model for the current task?**	
Model assessment	Who assesses model performance?	What are the incentives of the model assessors?	How might model assessors fail?
Model thresholds	What thresholds are set for anomaly detection?	What thresholds are set for change detection?	
Validation techniques	What techniques are appropriate for validating the model?	e.g. back-testing against previous performance e.g. expert judgments e.g. reasonableness checks on model outputs	
Validation time intervals	What are the right time intervals for continuous monitoring?	What is the right time interval for periodic model revalidation?	

Figure 5. Performance considerations

3.4.1 Performance risks

3.4.1.1 Human biases and perverse incentives

Although more and smarter automation reduces human involvement in routine processes, humans are still involved in and ultimately responsible for oversight and risk management of automated models, at every level of the organisation. (See below under *Managing aggregate model risk*). While risks arising from human management are of course part of Enterprise Risk Management, discussed below, they also need to be considered as a component of model risk, including at board level.

There is a large literature on the many human cognitive shortcuts, heuristics and biases which can reduce people's effectiveness in fulfilling these responsibilities – for example, overconfidence about forecast accuracy (J. Liu, Zhou, Wan, & Liu, 2019), confirmation bias (Allahverdyan & Galstyan, 2014), anchoring estimates to readily available data (Stulz, 2009), the planning fallacy, i.e. "nothing will go wrong with this project"

(Kahneman, 2011), groupthink (Janis, 1972), normalisation of deviance (Vaughan, 1997) and organisational licensing (Pernell, Jung, & Dobbin, 2017).

These issues all require attention in the design of human work and human interaction with automated systems and models. However, they are "means" problems; that is, people are honestly working towards a common end, but achieve it more slowly, more expensively, at a lower standard of accuracy or safety, or fail to achieve it altogether, because something goes wrong with their individual or collective thinking about how to do it.

Even more serious, perhaps, is the Principal-Agent problem, which is an "ends" problem, indicating divergence between the interests and goals of the organisation (the principal) and those of the individuals or teams working for it (the agents) (Stanley & Wdowin, 2018). In the context of model risk management, it is therefore important that the personal incentives and penalties faced by everyone with responsibility for managing model risk closely reflect the incentives and penalties facing the organisation.

3.4.1.2 Coordination between human oversight and automated process management

Regular monitoring and review processes can sometimes repeatedly fail to capture serious ongoing anomalies in business processes. For example, the bank UBS suffered from a single trader's fraud, which may have continued for years before it was detected, despite their best-practice financial risk management strategies and repeated positive outcomes from regular risk reports, internal audits and external reviews; this was apparently because risk managers did not sufficiently investigate automatically reported operational anomalies, or enforce existing risk controls (Conforti et al., 2013).

Airline pilots occasionally face analogous difficulties: problems with auto-pilot systems are fortunately extremely rare, but in consequence, when they do occur, pilots may not notice when automated processes fail or may misunderstand feedback from their instruments or otherwise fail to respond appropriately (Hart, 2017). (See case example below). Similar failures have occurred with experimental self-driving vehicles, where human drivers become complacent and fail to monitor the vehicle and the road adequately, and are not able to respond appropriately in time when something goes wrong (Marcus & Davis, 2019; Smith, 2018).

Evidently, it is not enough to have automated monitoring of live models and processes and routine human risk management procedures. Human risk management must be sufficiently engaged with automatic operational processes and alerts to monitor them

adequately, and this requires carefully designed human-machine interactions (Anderson et al., 2003; Wilson & Daugherty, 2018).

3.4.1.3 Setting and adjusting time intervals for model monitoring and review

All automated decision processes and the data on which they operate need to be monitored by humans at appropriate time intervals to make sure they are still returning accurate results. In setting the level of human scrutiny required, there is a trade-off between the cost of monitoring the process and the cost of missing an error or anomaly. This trade-off should be relatively straightforward to calculate in situations in which the rate of change in the data being processed is reliably approximately constant.

However, situations sometimes arise in which the rate of data change suddenly accelerates, as can happen in financial markets (Stulz, 2009). For these kinds of scenarios, it is critical firstly that the system has been designed to flag any sudden changes to human managers, and secondly that if they occur, the rate of monitoring can be rapidly adjusted upwards to match the current rate of change.

3.4.2 Case example

Automation in the airline industry began several decades ago. According to Christopher Hart, former Chairman of the US National Transportation Safety Board, “Automation in aviation began with a simple goal: assume that automation is good and automate whenever it is technologically feasible. Unfortunately, the ‘technologically feasible’ approach did not consider the human factors issues associated with the automation.” (Hart, 2017).

Hart provides several examples of accidents where a combination of unexpected unavailability of automated information and systems and pilot inexperience and error led to tragic results. While aviation accidents are rare, and typically have complex causes, they often include variations on these themes:

(1) When the automation is running smoothly, there is a tendency for pilots to become complacent and bored, and to pay insufficient attention to what is going on. When automation is temporarily unavailable, suddenly fails, or reports a problem, pilots may not notice (Hart, 2017) .

(2) When pilots do notice a problem with the automation, they may not be prepared to take control smoothly and competently due to lack of adequate training and recent relevant experience (Hart, 2017).

(3) When pilots have taken over manual control of a plane, they may misunderstand the feedback they are getting from their instruments and from the environment, again due to lack of adequate training and experience (Hart, 2017).

(4) Sometimes when pilots attempt to take manual control of a malfunctioning aircraft, they are unable to do so, because the automated task model has been designed to override pilot input (Travis, 2019), on the assumption that the automated system is always more reliable than the human pilot.

But if one is tempted to think that these problems might be resolved by removing human pilots altogether, Hart notes that the airline industry has been trying unsuccessfully to do this for decades. Further, he comments that "When automation encounters unanticipated circumstances, recovery usually depends entirely upon the pilots.... Situations involving automation with substantial human operator involvement have demonstrated two extremes. On the one hand, the human is the most unreliable part of the system. On the other hand, if the system encounters unanticipated circumstances, a highly-trained proficient human operator can save the day by being the most adaptive part of the system." So, while it is important to design automated task models to compensate for the weaknesses of human pilots, it is also critical to recognise human strengths in this environment and design automation which takes full advantage of human abilities.

3.5 Managing aggregate model risk

In the previous sections we have covered some key types of model risk arising at each stage of our overarching model of algorithmic models and model risk: Data, Design, Implementation and Performance. The management of aggregate model risk is different, in that it attempts to take an overview of the activities and risks of each of the other stages.

Aggregate model risk management consists in managing the risks associated with each individual model, and simultaneously in managing the aggregated risks arising from the organisation's portfolio of models. These aggregated risks will be affected by the total number of models being managed, the proportion of these models which are connected to each other or correlated with each other and therefore not independent, the strength and variety of the interactions between them, and the possible presence of superordinate goals.

Managing model risk is typically achieved in financial services through the Three Lines of Defence approach (Cohn, 2019; IIA, 2013), introduced in the early 2000s, and since updated, by the Institute of Internal Auditors (IIA). Line 1 consists in the model owners and developers,

who work in accordance with established procedures and checks in a model design, inspection and validation pipeline. Line 2 is the organisation's risk function, whose role is to advise on, oversee, validate and check the work of Line 1 staff, again in line with established procedures. Line 3 is the organisation's internal audit function, whose job is to oversee the work of both the Line 1 staff and the Line 2 staff, to ensure that model risk policies and procedures as they apply to both Lines 1 and 2 are fit for purpose and are being adhered to competently and in good faith, and to report any concerns upwards to the Board, and where necessary, to external regulators.

The Three Lines of Defence model has evolved since its introduction, partly in recognition of the importance of recognising business opportunities as well as threats under the risk umbrella; and partly to address what was originally a rather rigid separation between the three Lines, often resulting in an unnecessarily confrontational relationship between them (Cohn, 2019). The updated version aims to allow closer cooperation between the three Lines without compromising the independence of Line 2 and Line 3.

Figure 2 shows a somewhat simplified process for model development and approval at a typical financial institution. The main actors in this diagram are Line 1 staff, i.e., the people working in Line of Business (LOB) functions and the members of the Innovation Group who lead model development practice. Together they develop potential product development portfolios for prioritisation and evaluation, each of which the Innovation Group checks and decides whether it qualifies for development. The project scope and size are then determined, and a team assembled, which identifies acceptable models to achieve project objectives. The relevant LOB staff then choose a model to be developed and documented by the model development team.

The next stage involves Line 2 staff, who develop challenger models and evaluate the proposed model against the performance of the challengers. If the model passes the Line 2 evaluation, risk monitoring and control processes are then put in place and the model goes "live" into the business. Once it is active, it will continue to be monitored and will be recalibrated periodically as necessary or retired and replaced at some point with a better model.

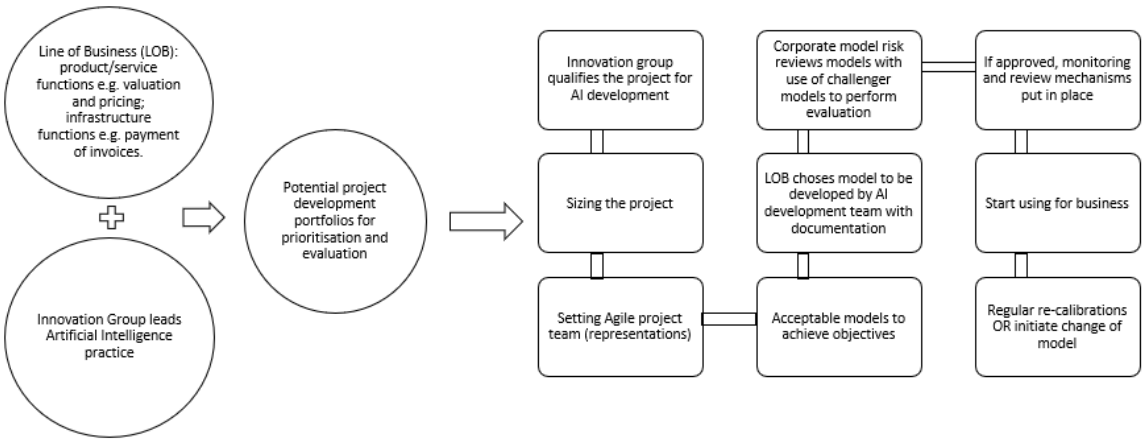

Figure 6. A simplified model development and approval pipeline typical of a financial institution

The entire model development process, from start through challenge and approval to production, re-calibration and retirement, is audited by the firm's internal audit function (Line 3). As we argued in the Design section, this type of model risk management process could usefully be adapted for a variety of industries and organisations outside financial services.

In addition to this process for developing and approving individual models, there also needs to be a process for assessing and managing the risk arising from the entire model portfolio. In banks, models are typically assigned to a risk tier, with higher risk models subjected to greater scrutiny at the review stage, both internally and by the regulator (Shi et al., 2016). This raises the question of whether, if the bank has a collection of similar, small models, these should each be treated separately as lower-tier models (tier 2 and 3) or whether they should be grouped together and treated as one larger tier 1 model. According to Shi, Young et al., this decision is often a matter of regulator preference, but it has serious consequences, since tier 1 models are typically much more heavily scrutinised than lower-tiered models. In addition, they point out that models are often interdependent, because some models' outputs are used as inputs for other models, and because models which share key inputs will produce correlated results. To deal with this, it is possible to trace direct output-input relationships between models, and to estimate the degree of correlation between models with shared inputs, to calculate approximately the risk arising from the whole network of models in addition to the individual model risk. Like the process for revalidating individual models, this process needs to be repeated at regular intervals to take account of changes in the model portfolio. It is therefore important that responsibility for these calculations and decisions is allocated between

Lines 2 and 3, and that there is a clear route for escalation of concerns as far as Board level when required.

The financial services industry has so far found ways to deal with unpredicted errors and unusual emergent behaviour of ML algorithmic models through a combination of alert systems to warn employees if unusual patterns are detected, automatic suspensions of activity if certain thresholds are breached, "kill switches" to enable employees to stop ML behaviour which looks worrying or dangerous, and "humans in the loop", who are required to approve proposed ML model actions before they go ahead. These kinds of mechanisms are integrated into each institution's model risk management practice, backed up by good information governance practices and the work of the risk management function within the Three Lines of Defence framework (Cohn, 2019; IIA, 2013), and the oversight of the relevant regulators (BankofEngland, 2019; FSB, 2017).

SR 11-7 recommends that boards and their committees adopt a structured approach to oversight of model risk within the broader risk management framework of the organisation. The board's goal is to reduce model risk to a level consistent with the institution's overall risk appetite, and they are led in this by management's approach to identifying, managing and mitigating model risk. The board's role is to challenge assumptions made by management, critically assess results and mandate changes where necessary; this can be difficult given the very large numbers of models in use, their complexity and the range of different types of output, and the fact that board members are not necessarily modellers themselves (Yoost, 2013). We place these challenges in their broader organisational context in the following section.

4. Technology and Enterprise Risk Management

Algorithmic models carry broader and long-term implications for an organisation, including:

- Reputation risks: The use of algorithmic models can significantly increase an organisation's exposure to reputation risks. This is particularly true if the various stakeholders believe that the workings of the model are not aligned to the ethics and values of the organisation, or if the models are designed to covertly manipulate consumers, regulators, or employees.
- Financial risks: Errors or vulnerabilities in models, especially those used for financial and strategic decision making, can result in significant revenue loss for organizations and negatively impact the integrity of their financial reporting.

- Operational risks: As models are used to automate supply chain and other operational areas, errors can result in significant operational disruptions.
- Regulatory risks: Algorithmic models that make decisions in violation of the law, circumvent existing rules and regulations, or discriminate against certain groups of people can expose organisations to regulatory and legal actions.
- Engineering and cyber-security risks: The wide-scale use of advanced algorithmic models can create new points of vulnerability for IT infrastructure.
- Strategic risks: With algorithmic models being used increasingly as sources for strategic decision making, errors or vulnerabilities within them can put an organization at a competitive disadvantage.

To handle these risks, organisations use Enterprise Risk Management. ERM is a coordinated set of activities to direct and control an organisation in relation to risk (RIMS, 2011). ERM is a strategic business discipline that supports the achievement of an organisation's objectives by addressing the full spectrum of its risks and managing the combined impact of those risks as an interrelated risk portfolio. Beyond model risk management, we have identified four main complementary approaches to enterprise risk management of a digital organisation: organisational culture; governance and regulation; organisational simulation; and investigation and testing.

The literature on organisational culture and risk stresses the importance of a culture in which risk is seen as the responsibility of all members (Agarwal et al., 2020; Kaplan & Mikes, 2012a). Hardy and colleagues (Hardy et al., 2020) highlight the literature on safety cultures, which assumes that a culture can be engineered to reduce accidents and contain risks and identifies features of a safety culture, such as low thresholds for reporting incidents, extensive communication about the importance of safety, training and protections for whistle-blowers (Flin, Mearns, O'Connor, & Bryden, 2000; Reason, 1997, 1998, 2000; Silbey, 2009). However, they also point out that this approach has been criticised for lack of conceptual rigour and failure to demonstrate causal relationships between safety culture and safety performance.

In the organisational literature more broadly, books and articles aimed at practitioners offer various recommendations for setting cultural expectations, including: the importance of a mission statement, regularly reviewed and clearly communicated, including the organisation’s core values, encompassing the perspectives of all stakeholders (Schwartz, Hagel, Wooll, & Monahan, 2019); the need to communicate clear ethical boundaries for

acceptable behaviour throughout the organisation (Simons, 1999); the need to ensure that business leaders' and employees' personal incentives are fully aligned with the interests of the organisation in relation to any risks to which they may expose the organisation (Pernell et al., 2017; Stanley & Wdowin, 2018; Stulz, 2009); the need to recognise, understand and value the contribution of their workforce, including casual workers, contractors, remote workers, overseas divisions and outsourced parts of the operation (Gray & Suri, 2019); and the importance of promoting workforce diversity actively in all functions and at every level of the organisation (Nunn, 2018).

There is empirical evidence to support some of these prescriptions: for example, the work of Pernell and colleagues on the relationship between CEO incentives and financial institutions' performance in the financial crisis (Pernell et al., 2017), and ethnographic work looking at the roles of remote and casual workers in digital organisations (Gray & Suri, 2019; Shestakofsky, 2017). And while there is as yet little work examining the business outcomes of high employee ethnic diversity, there is certainly evidence of the damage resulting from lack of diversity (Buolamwini & Gebru, 2018).

In the area of governance and regulation, the literature again offers a mixture of advice for practitioners and empirical studies of organisations. At the prescriptive end, organisations are recommended to: have clear procedures and governance structures to ensure compliance with relevant regulations, including for model risk and for other types of technology risk nested within enterprise risk (Cohn, 2019; IIA, 2013; Kaplan & Mikes, 2012a; Kosoff, 2016); to develop a constructive and honest relationship with relevant regulators, enabling full disclosure of relevant information and cooperation in the public interest (NAO, 2017; Stanley & Wdowin, 2018); and to promote understanding of and honest communication about management of all types of risk (preventable, strategic, external, systemic) throughout the organisation, including at board level (Agarwal et al., 2020; De Smet, Lund, Weiss, & Nimocks, 2021; Kaplan & Mikes, 2012a; Yoost, 2013). The empirical literature mainly highlights organisations in which, because of poor governance and/or lack of appropriate regulation, the rights, needs and concerns of individual customers or service users have been compromised (Dieterich et al., 2016; Foxglove, 2020; Haines, 2020; Hern, 2020; Larson et al., 2016; Martin, 2019; Molnar & Gill, 2018), with several of these authors calling for improvement.

A third approach to managing risk can be summarised as organisational simulation. Simulation covers a very wide range of methods and techniques for modelling real or

hypothetical situations and trying out alternative scenarios. Much of our discussion has been concerned with the models which organisations use to pursue their activities, and any such model is a simulation. By organisational simulation, we refer specifically to attempts to model the organisation itself, or a substantial subset of its operations, with the goal of improving the organisation's self-understanding and the quality and relevance of interventions to bring about organisational change. One influential and longstanding method for organisational simulation, championed by John Sterman and evidenced by many case studies, is System Dynamics, which encourages and enables teams to apply understanding from the behaviour of physical systems (such as stocks and flows and feedback loops) to model the behaviour of their own organisation (Sterman, 2000, 2002, 2018; Sterman, Oliva, Linderman, & Bendoly, 2015). Another simulation tradition covers the closely related practices of scenario planning and war-gaming, which are more qualitative approaches for which it is difficult to measure results, but which nonetheless have sustained their popularity with practitioners over a long period (Augier, Dew, Knudsen, & Stieglitz, 2018). All these methods could be seen as a response to the insight that organisations are open systems, deeply interconnected with the world around them, including the behaviour of suppliers, customers and competitors, relevant technological developments, social and political trends and pending regulatory change (Rousseau, 1979), and that therefore organisations need to understand and monitor these relationships continuously to survive and thrive.

Finally, the literature highlights the important role of investigation and testing in any risk management or risk organising regime. Hardy and colleagues (Hardy et al., 2020) identify three modes of risk organising: prospective, i.e., during the design and planning phase of an operation; real-time, i.e., while an operation is in progress; and retrospective, i.e., after a risk has been realised and an incident has occurred. Investigation and testing are important in all three modes: initially, to establish the safety and suitability of the proposed operation, perhaps partly using prototypes and simulations; then during operation to ensure the operation is proceeding as planned, and is robust to possible changes in the environment, with the testing regime itself kept under continuous review through governance processes and structures (Kaplan & Mikes, 2012a; Sterman, 2000). Finally, when critical incidents do occur, practitioner articles stress the need to review losses, failures and mistakes in a timely, consistent and effective fashion, as mandated by appropriate governance processes (Agarwal et al., 2020; Kaplan & Mikes, 2012a; Kosoff, 2016). In practice, the empirical literature suggests that at least in the case of large formal proceedings such as public inquiries, retrospective

investigations are often hampered in their quest for understanding and future incident prevention by a focus on who is to blame, which makes employees less willing to report mistakes (McArdle, Burns, & Ireland, 2003; Waring, 2005) and encourages managers to limit the scope of investigations (Elliott & McGuinness, 2002; Hutter, 1992) – emphasising that while investigations and tests are important, they rely for their effectiveness on the culture in which they are embedded.

For ERM to be effective therefore, an organisation should have an organisation-wide approach to manage technical and cultural risks: This approach should include principles, policies, and standards; roles and responsibilities; control processes and procedures; and appropriate personnel selection and training. It also should provide transparency and processes to handle enquiries, which can help organisations use algorithmic models responsibly. There should be clear processes and approaches for the development, deployment and use of algorithmic models; and the processes and approaches should be aligned with the governance structure to address the model life cycle, including data selection, algorithm design, integration and implementation and live use in production. Further, processes for assessing and overseeing algorithmic data inputs, workings and outputs should be established, leveraging state-of-the-art tools as they become available. Internal and external parties can provide objective reviews of algorithms and related issues (Deloitte, 2020).

4.1 Enterprise risk and wider regulation

Regulation covers a huge range of different types of activity – for a useful summary see the National Audit Office's recent overview (NAO, 2017). The regulatory goals mentioned by the NAO largely fall into two broad categories. Firstly, there is a need to protect the public from a wide variety of harms, which Stanley and Wdowin (2018) summarise as: those arising from information asymmetries between companies and customers; the protection of specific vulnerable customer groups such as children, the elderly, the poor, the disabled; and protecting all customers from unethical and exploitative behaviour by companies. Secondly, regulators are tasked with making sure that social systems such as markets, industries and public services work efficiently and fairly in the public interest, largely through improving the competitiveness of markets or controlling prices where competition is impossible.

A third focus of regulatory concern (already mentioned above) is systemic risk, i.e., the risk that failure of a single node in a network, for instance a financial institution, could cause

the collapse of an entire system, market or industry. Systemic risk has become a central concern for the financial services industry and its regulators since the global financial crisis of 2008 (Arner et al., 2019; FSB, 2017; Holopainen & Sarlin, 2017; Kou, Chao, Peng, Alsaadi, & Herrera-Viedma, 2019), and is now being recognised and studied in other domains as well (Goldin & Mariathasan, 2014; Renn, Lucas, Haas, & Jaeger, 2019). Therefore, in addition to oiling and rebalancing the wheels of the economy and protecting individuals from hazards and violations of their rights, regulators need to consider sources of risk which could lead to systemic crisis or collapse. This is of course more difficult when they are tackling globally interconnected, non-linear risks requiring multilateral cooperation between governments.

In principle, regulators try to take a "God's-eye view" of an industry or market and then intervene from the outside to adjust the way it operates. However, regulators are human and like everyone else, subject to the typical range of human cognitive biases and errors, on which there is a substantial literature in psychology and organisational behaviour (Goswami & Pandey, 2019; L. A. Jackson, Sullivan, & Hodge, 1993; Janis, 1972; Jones, Davis, & Berkowitz, 1965; Kahneman, 2011; Lee, Hallahan, & Herzog, 1996; Pinto, 2014; Vaughan, 1997). Furthermore, they become part of the system they are regulating, because of the necessary relationships between themselves and the industry, both corporately and individually. This means there is considerable potential for what Sterman (2002) calls "policy resistance", i.e. interventions having unrecognised feedback effects which in the longer-term are counterproductive – for instance, we summarised above the work of Pernell, Jung et al. (2017), showing how the US Sarbanes-Oxley Act, aimed at reducing the risk carried by banks, led to the appointment of Chief Risk Officers in financial institutions, which paradoxically led to those institutions taking on more risk.

Stanley and Wdowin (2018) note that, even in the best case, regulators depend on the expertise and cooperation of those in the industries they regulate, which inevitably limits their ability to act independently; in addition, those working for regulators have often previously worked in the industry they regulate, and may expect to do so in the future, which further reduces their independence and willingness to criticise or penalise industry (a form of Principal-Agent problem); they also point out that regulators tend to be underfunded, limiting the resources available for regulatory activity and the salaries offered compared to those available in the industry, which reduces the regulators' effectiveness and increases their reliance on industry cooperation and goodwill – a point echoed by the Centre for Data Ethics and Innovation in the specific context of AI (CDEI, 2020). These problems together can add

up to “regulatory capture” in which the regulator is too weak to restrain risky industry behaviour, and instead becomes the industry’s advocate and servant.

There is also a structural problem, in that any given regulatory regime is static, and designed to manage industry behaviour under certain assumptions, which may not hold true for very long, necessitating continuous revision and updating of regulations. For example, as mentioned above, GDPR is already failing to cover IoT devices adequately (Sullivan, 2018).

In recent years, there have been some attempts to use algorithms in an explicitly regulatory role, an approach which has some obvious advantages. Firstly, algorithms can respond to actual behaviour as it happens, to monitor and punish abuses much more comprehensively and systematically than a rule book backed up only with human resources would normally allow. Stanley and Wdowin (2018) give the example of traffic speed cameras with number plate and face recognition capabilities, which can be deployed much more broadly and effectively than occasional police speed traps. Secondly, in the case of social media moderation, algorithms relieve humans of the trauma of viewing very violent or pornographic content in the course of their work. On the other hand, serious problems have emerged with algorithmic moderation of social media: current algorithmic models are incapable of distinguishing between abuse and legitimate discussion of that abuse, with the result that victims of racist, sexist or homophobic abuse on social media are then frequently banned from social media for discussing their mistreatment (Guynn, 2019; Watanabe, 2019).

It is important to see these limitations of regulatory regimes in context, recognising that modern approaches to regulating and managing risk have been hugely successful in areas like reducing fatal accidents at work, reducing chronic illness and fatal diseases caused by exposure to toxins in the workplace or environmental pollution, improving fire safety, transport safety and so on (Brione, 2017; Renn et al., 2019). Unfortunately, as society has successfully tackled these traditional types of risk, it has in the process created new ones, which are often more challenging to address, because they are non-linear and globally connected – for instance, climate change (IPCC, 2014), the global financial system (Arner et al., 2019), increasing inequality between rich and poor (Atkinson, 2015; WEF, 2020) and of course global pandemics (Fan, Jamison, & Summers, 2018).

4.2 Case example

In financial services, the role of risk professionals and regulators has evolved significantly since the 1980s when risk management first emerged as a specialism. Much recent discussion

has understandably focused on regulation and risk management within banks and other financial institutions in a post-financial-crisis world, particularly with reference to AI/ML technology (Arner et al., 2019; BankofEngland, 2019; FSB, 2017; Kou et al., 2019). However, there are also lessons to learn from the role of risk management in the run-up to the financial crisis which show some of the reasons that risk management is intrinsically difficult. According to Pernell, Jung and Dobbin (2017), the profile of risk professionals was raised in the early 2000s after the passing in the US of the Sarbanes-Oxley Act in 2002, which laid out new areas of responsibility for banks but did not specify detailed compliance standards. To meet these new responsibilities, banks began appointing Chief Risk Officers (CROs). Pernell et al present evidence that the appointment of CROs was to some extent counter-productive, leading banks with a CRO to take on more risk, for two reasons: "organisational licensing", i.e. the knowledge that there was someone responsible for risk management led others in the organisation to be less concerned and careful about risk; and the preference of the CROs themselves for new forms of derivatives such as credit-default swaps, which they saw as tools for enabling fast and precise adjustments to risk exposures in the service of maximising risk-adjusted returns. Thus, Pernell et al argue that the rise of CROs since 2002, prompted by legislation to limit bank exposure to risk, was in fact a contributory cause of the financial crisis in 2008.

For these reasons, while there are principles which can guide organisations towards managing risk effectively, it is not a function which can ever be fully delegated to specialised risk managers, however skilled and conscientious, either by the head of the organisation or by the business function managers within it (Kaplan & Mikes, 2012a). The same therefore also applies to any technology or risk management system brought in to support risk management, at any level, including the model risk management approaches we have described – however good it is, responsibility for protecting the organisational systems remains with humans, throughout the organisation but especially at the top. This responsibility is not merely formal or legal, while the actual work is done by machinery and people elsewhere in the organisation; rather it is a real ongoing requirement that top management and business function managers stay informed, open to discussion and challenge and aware of potential emerging risks across the organisation, particularly where these may interact. And despite the limitations of poor regulator funding and regulatory capture, in many industries there is also an important risk management role for external regulators (CDEI, 2020; NAO, 2017; Stanley & Wdowin, 2018).

5. Conclusion

Modern risk management has become very effective at dealing with traditional health and safety risks (Brione, 2017; Renn et al., 2019). This may be one reason for the common tendency noted by Kaplan and Mikes (2012a) to focus risk management efforts almost entirely on compliance. As the risk environment becomes more challenging, the need grows for a broader and more proactive approach to risk, less purely focused on risk measurement and more concerned with *organising* risk. To develop this successfully, we will need to find ways of reliably recognising and routinely overcoming human cognitive biases and systematic errors within organisational practice (Kaplan & Mikes, 2012a; Stulz, 2009).

Some theorists have suggested that in the face of high levels of ambiguity, or "unforeseeable uncertainty", organisations should not attempt to apply traditional risk management methods at all. Instead, they should operate on a basis of continuous learning and adaptation as changing situations unfold (Sommer & Loch, 2004). These researchers argue that, although traditional project risk management methods work well in contexts where the project team can reasonably foresee and understand potential threats, in situations where it is impossible to fully understand all relevant variables and interactions, the traditional methods breakdown. In these circumstances they recommend an approach of constant environmental scanning to recognize an unforeseen event when it arises, combined with problem-solving and a willingness to modify policies to develop an appropriate response quickly.

Whatever approach is used, it is unarguable that risk management is hard, and while there are guidelines and principles which can help organisations to protect themselves more effectively, the effort to do so is always to some extent fighting against human nature. The technical and human elements of this challenge cannot be dealt with in isolation from each other. As Sterman (2002: 4) comments, "For many of the most important problems, there are no purely technical solutions. Indeed, there are no purely technical problems.". Rather, each set of problems needs to be considered in terms of a model which recognises our interdependence with machines within complex dynamic systems. Within such systems, there is potential for humans and machines either to correct each other's faults and amplify each other's strengths, or conversely to negate each other's strengths and amplify each other's faults.

This relates to the concept of “risk work”, which is defined as “situated human effort, in combination with material infrastructure, through which risk management and governance practices come to be constructed” (Power, 2016: 3). This perspective examines how risk is organized through the interactions of embedded, embodied agents in particular contexts as they engage in day-to-day activities (e.g. Palermo, 2016). It assumes that organizational encounters with risk are “a routine and systematic part of daily organizational life”(Vaughan, 2005: 33). This perspective thus adopts a finely grained, bottom-up focus that directs attention to “the actions and routines through which organizational actors make sense of risk, of themselves and their roles, and collectively try to enact institutional scripts” (Power, 2016: 8).

In a slightly different sense, the processes of model risk management and in the Three Lines of Defence model as they have been developed and enacted so far in the financial services industry have also been largely bottom-up, i.e. they have started with the work that individuals are engaged in and have attempted to create processes and structures for and around that work to manage the risks arising from it (Cohn, 2019; IIA, 2013; Kosoff, 2016; USFederalReserve, 2011). However, we also see a developing critique in the financial services industry of “traditional” MRM processes; current practice has been criticised for being overly compliance-based and prescriptive, rigid and siloed, and lacking top-down, strategic and reflective dimensions (Bridgers et al., 2020; Hill, 2020; Shi et al., 2016). Further, there is an emerging awareness that while the Three Lines of Defence model creates pathways for issues to be escalated to board level, boards may not always be equipped to resolve such issues and will themselves need appropriate experience, guidance and support (De Smet et al., 2021; Yoost, 2013). These critiques highlight the inevitable connections and overlaps between successful model risk management and enterprise risk management, and the need to situate and attend to model risk, both within its immediate human context and the wider context of enterprise risk. There is at the same time an ongoing evolution of MRM practice within financial services from pure model validation to a risk management practice which adds value to the institution (Crespo et al., 2017; Garro, 2020). Outside the organisation, regulators also have a critically important if necessarily imperfect role in recognising emerging risk and supporting and challenging regulated organisations; so, enterprise risk is further nested within industry or sector-wide and even societal systemic risk (Arner et al., 2019; BankofEngland, 2019; CDEI, 2020; Ehrmann & Schure, 2019; FSB,

2017; Goldin & Mariathasan, 2014; Kou et al., 2019; NAO, 2017; Pernell et al., 2017; Renn et al., 2019; Stanley & Wdowin, 2018; WEF, 2020).

We have argued that model risk is a concept which can usefully be applied in many different industries and sectors beyond the financial services industry. Garro makes the point that various unfortunate historical incidents which led to huge losses can reasonably be described as the result of poor management of model risk, even though they might not have been so conceptualised at the time; he refers to the collapse of the Tacoma Narrows Bridge in the United States in 1940, which occurred because engineers had modelled the design of the bridge without considering the vertical movements of the wind (Garro, 2020). Given our definition of a model as *any attempt to predict or forecast what will happen given a specific set of variables and a relevant set of data*, it is evident that such models are everywhere in industry and indeed in everyday life.

While model risk is not a widely used term outside financial services, it is increasingly relevant to the activities of companies and organisations using algorithmic models, often including machine learning functionality, to process large quantities of data for a rapidly expanding range of purposes. As MRM practices mature in financial services, they face new challenges because of the vast scale and variety of model-related activity in that industry. Meantime, there are lessons from financial services for other organisations about how to begin to standardise and scale up the process of prototyping, validating and testing individual models, how to establish a model validation and governance framework, how to deal with aggregate model risk, and how to nest this successfully within the wider structures and processes of enterprise risk management.

5.1 Directions for future research

As more organisations start to grapple with the risk implications of the models they use, there are opportunities for observational or action research case studies outside financial services with organisations establishing a model risk management function and embedding MRM practices within a wider enterprise risk management approach.

A theme running through this article has been the importance of the roles of the human users/operators of models, and the need for such roles to be designed to dovetail with the models themselves so that the resulting machine-human systems can work harmoniously and effectively. Wilson and Daugherty have outlined the (ideally) mutually supportive roles of humans and AI/ML systems in achieving organisational goals (Daugherty & Wilson, 2018;

Wilson & Daugherty, 2018); yet we cited several studies describing high-friction interactions between algorithmic models and respectively, nurses (Russell, 2012), radiologists (Lebovitz et al., 2019), telephone salespeople (Bader & Kaiser, 2019) and airline pilots (Hart, 2017). More studies are certainly needed to look at how human-machine dovetailing can be intentionally designed to work smoothly and efficiently. More encouragingly, we also note the existence of ML models which train or support individuals or groups of human workers to achieve superior results (Brynjolfsson & McAfee, 2017; Weinstein, 2019); this seems like a worthwhile area for further work.

Regulation of AI/ML itself is likely to expand, but ML applications are also likely to be used increasingly to support regulatory compliance in many different fields. We have touched on the strengths and weaknesses of algorithms as regulatory tools, but there is certainly scope for further research. The problem of social media moderation algorithms penalising abuse victims is one which merits further study, for example.

The work of Pernell et al. (2017) provides a warning of the potential for regulation to be counter-productive, and of the ways business priorities and personal incentives for leaders can subvert good risk management. Given this stark illustration of Sterman's "policy resistance" (Sterman, 2002), it would be interesting to look at risk management practices in an organisation or industry through a dynamic systems lens, with the goal of helping risk professionals, business leaders and regulators to proactively identify possible blind spots in their approach to both model risk and enterprise risk. There could also be an opportunity to use ML to model an organisation's risk management behaviour.

Government departments and regulators wanting to rely on algorithmic models for efficiency and accuracy will need to make them transparently fair and capable of providing explanations of decisions, while at the same time protecting sensitive information – an area where greater understanding is certainly required. While existing work has exposed how difficult it can be to avoid algorithmic bias, and the problems caused by lack of explainability, the trade-offs between ensuring a necessary degree of transparency and the need in some sensitive contexts for confidentiality (for example the 2020 A level algorithm crisis) have perhaps not yet been fully explored.

Algorithmic bias and lack of algorithmic explainability are both critical aspects of model risk which every organisation considering using ML technologies needs to consider and take steps to address in advance of any problems arising. While there is already

considerable research in these areas, they will continue to evolve and research will need to keep pace to contribute to the establishment of organisational best practice, and to inform changes to regulation and the law. Beyond model risk, other technology risks including cybersecurity, technology-driven “digital divides”, big data management and issues around trust are extremely important and require further study.

Considered as systems, organisations use model risk management and enterprise risk management techniques to protect themselves from internal failures, external attacks and environmental shifts; but they can only do this effectively by recognising that they are embedded in larger dynamic systems within society, with which they constantly communicate – supply chains, communities, industries, markets, ecosystems, nations, trading blocs and so on. The management of model risk, including risks specific to ML, is just one complex area within a huge range of opportunities and challenges that organisations and societies face together in an increasingly interconnected and interdependent world.

REFERENCES

Agarwal, H., Argarwal, R., Kayyali, B., & Stephens, D. (2020). Four ways to improve technology service resiliency. *McKinsey Digital*, 1-5. Retrieved from https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/four-ways-to-improve-technology-service-resiliency

Allahverdyan, A. E., & Galstyan, A. (2014). Opinion dynamics with confirmation bias. *PLoS One, 9*(7), e99557. doi:10.1371/journal.pone.0099557

Anderson, S., Hartswood, M., Proctor, R., Rouncefield, M., Slack, R., Soutter, J., & Voss, A. (2003). *Making Autonomic Computing Systems Accountable: The Problem of Human-Computer*. Paper presented at the 14th International Workshop on Database and Expert Systems Applications (DEXA'03), Prague, Czech Republic. https://www.researchgate.net/publication/221646888_Making_Autonomic_Computing_Systems_Accountable_The_Problem_of_Human-Computer

Arner, D. W., Avgouleas, E., Busch, D., & Schwarcz, S. L. (Eds.). (2019). *Systemic Risk in the Financial Sector: Ten Years after the Great Crash*: CIGI Press.

Athey, S., & Stern, S. (2002). The Impact of Information Technology on Emergency Health Care Outcomes. *The RAND Journal of Economics, 33*(3), 399-432. doi:10.2307/3087465

Atkinson, A. B. (2015). *Inequality: What can be done?* Cambridge, Massachusetts; London, England: Harvard University Press.

Augier, M., Dew, N., Knudsen, T., & Stieglitz, N. (2018). Organizational persistence in the use of war gaming and scenario planning. *Long Range Planning, 51*, 511-525.

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration*. *The Quarterly Journal of Economics, 118*(4), 1279-1333. doi:10.1162/003355303322552801

Azulay, D. (2019, 22 January 2019). Data Collection and Enhancement Strategies for AI Initiatives in Business. Retrieved from https://emerj.com/partner-content/data-collection-and-enhancement-strategies-for-ai-initiatives-in-business/

Bader, V., & Kaiser, S. (2019). Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization, 26*(5), 655-672. doi:10.1177/135058419855714

BankofEngland. (2019). *Machine learning in UK Financial Services*. Retrieved from

Bayer, S., Sandy, S., Schrader, U., & Spiegl, M. (2021). Leveraging digital and analytics in biopharma operations: Six principles. *Operations Practice*. Retrieved from https://www.mckinsey.com/business-functions/operations/our-insights/leveraging-digital-and-analytics-in-biopharma-operations-six-principles

Bevan, O., Ganguly, S., Kaminski, P., & Rezek, C. (2016). 'The ghost in the machine': Managing technology risk. *McKinsey Quarterly online*. Retrieved from https://www.mckinsey.com/business-functions/risk-and-resilience/our-insights/the-ghost-in-the-machine-managing-technology-risk

Brands, K. (2020). CREATING CYBERSECURITY AWARENESS. *Strategic Finance*, 60-61. Retrieved from https://ezp.lib.cam.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bsu&AN=141036038&site=bsi-live&scope=site

Bridgers, A., Lee, H., & Kosoff, J. (2020). Adopting a Top-Down Approach to Model Risk Governance to Optimize Digital Transformation. *The RMA Journal, 103*(1), 28-31.

Brione, P. (2017). *Minds over Machines: New Technology and Employment Relations*. Retrieved from London:

Bromiley, P., McShane, M., Nair, A., & Rustambekov, E. (2015). Enterprise Risk Management: Review, Critique and Research Directions. *Long Range Planning, 48*, 265-276.

Brotcke, L. (2020). Modifying model risk management practice in the era of AI/ML. *Journal of Risk Managment in Financial Institutions, 13*, 255-265.

Brynjolfsson, E., & McAfee, A. (2017). The Business of Artificial Intelligence: What it can - and cannot - do for your organization. *Harvard Business Review, The Big Idea*. Retrieved from https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81*(2018 Conference on Fairness, Accountability and Transparency), 1-15.

Butcher, B. (2013). Hollowing out and the future of the labour market - the myth. *VOX CEPR Policy Portal: Column*. Retrieved from https://voxeu.org/article/hollowing-out-labour-market-new-evidence

Cadwalladr, C., & Graham-Harrison, E. (2018, 17 March 2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. Retrieved from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Campa, R. (2019). Three Scenarios of the Future of Work: Technological Unemployment, Compensation, Hollowing Out. *Sociology and Technoscience (Sociologia y tecnociencia), 9*(2), 140.154. doi:10.24197/st.2.2019.140-154

Carini, S. (2019, 19 February 2019). Scholar explores impact of bias in facial-recognition software. *Emory News Center.* Retrieved from http://news.emory.edu/stories/2019/02/er_provost_lecture_buolamwini/campus.html

CDEI. (2020). *AI Barometer Report*. Retrieved from London, UK: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf

Chowdhury, R. (2018). Auditing Algorithms for Bias. *Harvard Business Review*. Retrieved from https://hbr.org/2018/10/auditing-algorithms-for-bias

Cobbe, J., & Morison, J. (2019). Understanding the Smart City: Framing the challenges for law and good governance. In J. Auby, É. Chevalier, & E. Slautsky (Eds.), *Le Futur de Droit Administratif/The Future of Administrative Law*. Paris: LexisNexis.

Cohn, M. (2019). A fresh look at risk management: The IIA is re-evaluating its long-time model. *Accounting Today, 33*(2).

Conforti, R., La Rosa, M., Fortino, G., ter Hofstede, A. H. M., Recker, J., & Adams, M. (2013). Real-time risk monitoring in business processes: A sensor-based approach. *Journal of Systems and Software, 86*(11), 2939-2965. doi:https://doi.org/10.1016/j.jss.2013.07.024

Cresci, S. (2020). A Decade of Social Bot Detection. *Communications of the ACM, 63*(10). doi:10.1145/3409116

Crespo, I., Kumar, P., Notebook, P., & Taymans, M. (2017). The evolution of model risk management. *Risk*. Retrieved from https://www.mckinsey.com/business-functions/risk/our-insights/the-evolution-of-model-risk-management

Daugherty, P. R., & Wilson, H. J. (2018, 17 July 2018). What Are The New Jobs In A Human + Machine World? *Forbes*. Retrieved from https://www.forbes.com/sites/insights-intelai/2018/07/17/what-are-the-new-jobs-in-a-human--machine-world/#71e9a56663e3

(2021, 8 April 2021). *Boards and decision making* [Retrieved from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/boards-and-decision-making

Deloitte. (2020). AI and risk management: Innovating with confidence. Retrieved from https://www2.deloitte.com/uk/en/pages/financial-services/articles/ai-and-risk-management.html

Dharasathy, A., Jain, S., & Khan, N. (2020). When governments turn to AI: Algorithms, trade-offs and trust. *McKinsey Public Sector Practice*. Retrieved from

https://www.mckinsey.com/industries/public-and-social-sector/our-insights/when-governments-turn-to-ai-algorithms-trade-offs-and-trust

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity performance of the COMPAS risk scales in Broward County.* Retrieved from http://go.volarisgroup.com/rs/430-MBZ-989/images/ProPublica_Commentary_Final_070616.pdf

Doshi-Velez, F., & Kortz, M. A. (2017). Accountability of AI under the Law: the Role of Explanation. *Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet and Society working paper*. Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

Downes, L. (2018). GDPR and the End of the Internet's Grand Bargain. *Harvard Business Review*. Retrieved from https://hbr.org/2018/04/gdpr-and-the-end-of-the-internets-grand-bargain

Duhigg, C. (2012, 16 February 2012). How Companies Learn Your Secrets. *The New York Times Magazine*. Retrieved from https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Dzinkowski, R. (2019). Cyber Risk: Time to Elevate the Agenda. *Strategic Finance*. Retrieved from https://sfmagazine.com/post-entry/october-2019-cyber-risk-time-to-elevate-the-agenda/

Ehrmann, M., & Schure, P. (2019). The European Systemic Risk Board - governance and early experience. *Journal of Economic Policy Reform, 23*(3), 290-308. Retrieved from https:doi.org/10.1080/17487870.2019.1683011

Elliott, D., & McGuinness, M. (2002). Public inquiry: Panacea or placebo? *Journal of Contingencies and Crisis Management, 10*(1), 14-25.

Falkner, R., & Jaspers, N. (2012). Regulating Nanotechnologies: Risk, Uncertainty and the Global Governance Gap. *Global Environmental Politics, 12*(1), 30-55. doi:10.1162/GLEP_a_00096

Fan, V. Y., Jamison, D. T., & Summers, L. H. (2018). Pandemic risk: how large are the expected losses? *Bulletin of the World Health Organization, 96*(2), 129-134.

Fingas, J. (2019). New York investigates claims of sexism in Apple Card credit limits (updated). Retrieved from https://www.engadget.com/amp/2019/11/09/new-york-investigates-apple-card-credit-limit-sexism/?guccounter=1&guce_referrer=aHR0cHM6Ly90LmNvL1FQMkkyUHNMNE8_YW1wPTE&guce_referrer_sig=AQAAAJ9qnu9j21bn8pIE1_rf5_9cnaIHXYGeRQfwP2VKTyPndmN3Qwom890JSPz3TWuz1xDif7fQKL6W0oQro0P5IjFv6AyMVQu3-dX2EwdThNf7opbX48MhTOXx4rkKArNHPgeymHKJf45LP5EB-ssZC8TgYhmxNKWF8hLIs06H0OT2&__twitter_impression=true

Flin, R., Mearns, K., O'Connor, P., & Bryden, R. (2000). Measuring safety climate: Identifying the common features. *Safety Science, 34*(1-3), 177-192.

Foxglove. (2020, 4 August 2020). How we got the government to scrap the visa streaming algorithm - some key legal documents. Retrieved from https://www.foxglove.org.uk/news/c6tv7i7om2jze5pxs409k3oo3dyel0

FSB. (2017). *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications*. Retrieved from

Garro, M. (2020). The evolution of model risk management processes. *Journal of Risk Management in Financial Institutions, 13*(1), 16-23.

Goldin, I., & Mariathasan, M. (2014). *The Butterfly Defect: How Globalization Creates Systemic Risks, and What to Do about It*. Princeton, NJ: Princeton University Press.

Goos, M., & Manning, A. (2007). Lousy and Lovely Jobs: The Rising Polarization of Work in Britain. *The Review of Economics and Statistics, 89*(1), 118-133. Retrieved from https://ideas.repec.org/a/tpr/restat/v89y2007i1p118-133.html

Goswami, A., & Pandey, J. (2019). Credit attribution bias and its impact on employee morale and retention. *Strategic HR Review, 18*(2), 80-83. doi:10.1108/SHR-04-2019-159

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*: Houghton Mifflin Harcourt USA.

Gressin, S. (2017). The Equifax Data Breach: What to Do. Retrieved from https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do

Grover, V., Chiang, R. H. L., Liang, T.-P., & Zhang, D. (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems, 35*(2), 388-423. doi:10.1080/07421222.2018.1451951

Guynn, J. (2019). Facebook while black: Users call it getting 'Zucked', say talking about racism is censored as hate speech. *USA Today*. Retrieved from https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

Haag, H., Peláez-Pier, F., Cohen, R., Greene, E., Guynn, R., Peder Hammarskiöl, . . . Wong, L. (2010). *The International Bar Association's Task Force on the Financial Crisis: A survey of current regulatory trends*. Retrieved from https://www.davispolk.com/files/uploads/FIG/Financial_Crisis_Report_IBA.pdf

Hackett, R. (2015). Experian data breach affects 15 million people including T-Mobile customers. *Fortune*. Retrieved from https://fortune.com/2015/10/01/experian-data-breach-tmobile/

Haines, T. S. (2020). A-Levels: The Model is not the Student. Retrieved from http://thaines.com/

Hall, J. V., Horton, J. J., & Knoepfle, D. T. (2021). Pricing in Designed Markets: The Case of Ride-sharing. 1-61. Retrieved from http://john-joseph-horton.com/papers/uber_price.pdf

Hannif, Z., Cox, A., & Almeida, S. (2014). The impact of ICT, workplace relationshups and management styles on the quality of work life: insights from the call centre front line. *Labour and Industry: a journal of the social and economic relations of work, 24*(1), 69-83. Retrieved from https://doi.org/10.1080/10301763.2013.877120

Hardy, C., Maguire, S., Power, M., & Tsoukas, H. (2020). Organizing Risk: Organization and Management Theory for the Risk Society. *Academy of Management Annals, 14*(2), 1032-1066. doi:10.5465/annals.2018.0110

Hart, C. (2017). What Can Self-Driving Cars Learn from Aviation? Retrieved from https://www.thedrive.com/tech/13903/what-can-self-driving-cars-learn-from-aviation

He, C. Z., Frost, T., & Pinsker, R. E. (2020). The Impact of Reported Cybersecurity Breaches on Firm Innovation. *Journal of Information Systems, 34*(2), 187-209. doi:10.2308/isys-18-053

Hern, A. (2020, Friday 21 August 2020). Ofqual's A-level algorithm: why did it fail to make the grade? *The Guardian*. Retrieved from https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels

Heudecker, N., & Hare, J. (2016). *Survey analysis: Big data investments begin tapering in 2016.* . Retrieved from https://www.gartner.com/doc/3446724/survey-analysis-big-data-investments

Hill, J. R. (2020). A smarter model risk management discipline will follow from building smarter models. *Journal of Risk Management in Financial Institutions, 13*(1), 24-34.

Holopainen, M., & Sarlin, P. (2017). Toward robust early-warning models: a horse race, ensembles and model uncertainty. *Quantitative Finance, 17*(12), 1933-1963.

HouseofCommons, & DCMSCommittee. (2019). *Disinformation and 'fake news': Final Report. Eighth Report of Session 2017-2019*. London, UK: House of Commons Retrieved from Available at: www.parliament.uk/dcmscom

Hutter, B. (1992). Public accident inquiries: The case of the railway inspectorate. *Public Administration, 70*(3), 177-192.

ICO. (2021a). Guide to Data Protection. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/

ICO. (2021b). Guide to Privacy and Electronic Communications Regulations. Retrieved from https://ico.org.uk/for-organisations/guide-to-pecr/

IIA. (2013). *The Three Lines of Defense in Effective Risk Management and Control*. Retrieved from https://na.theiia.org/standards-guidance/Public%20Documents/PP%20The%20Three%20Lines%20of%20Defense%20in%20Effective%20Risk%20Management%20and%20Control.pdf#:~:text=The%20Three%20Lines%20of%20Defense%20model%20distinguishes%20among,defense,%20operational%20managers%20own%20and%20manage%20risks.%20They

IPCC. (2014). *Climate Change 2014: Synthesis Report*. Retrieved from Geneva:

Jackson, J. R. (2018). Algorithmic Bias. *Journal of Leadership, Accountability and Ethics, 15*(4), 55-65.

Jackson, L. A., Sullivan, L. A., & Hodge, C. N. (1993). Stereotype effects of attributions, predictions, and evaluations: No two social judgments are quite alike. *Journal of Personality and Social Psychology, 65*(1), 69-84. doi:10.1037/0022-3514.65.1.69

Janis, I. L. (1972). *Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes*. Boston: Boston, Houghton, Mifflin.

Johnson-Laird, P. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge UK: Cambridge University Press.

Jones, E. E., Davis, K. E., & Berkowitz, L. (1965). *From Acts to Dispositions: The Attribution Process in Person Perception.*: Academic.

Jørgensen, L., & Jordan, S. (2016). Risk mapping: Day-to-day riskwork in inter-organizational project management. In M. Power (Ed.), *Riskwork: Essays on the organizational life of risk management* (pp. 50-71). Oxford, UK: Oxford University Press.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farra, Strauss and Giroux.

Kaplan, R. S., & Mikes, A. (2012a). Managing Risks: A New Framework. *Harvard Business Review*(June).

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: the new contested terrain of control. *Academy of Management Annals, 14*(1), 366-410.

Kochan, T. A. (2019). Shaping the Future of Work: Challenges and Opportunities for US Labor Management Relations and Workplace Dispute Resolution. *Dispute Resolution Journal, 74*(1), 11-31.

Kosoff, J. (2016). Best Practices in Model Risk Audit. *The RMA Journal, 98*(6), 36-41.

Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., & Herrera-Viedma, E. (2019). MACHINE LEARNING METHODS FOR SYSTEMIC RISK ANALYSIS IN FINANCIAL SECTORS. *Technological & Economic Development of Economy, 25*(5), 716-742. doi:10.3846/tede.2019.8740

Lambrecht, A., & Tucker, C. (2019). Algorithmic Bias? An Empirical Study of Apparent Gender-based Discrimination in the Display of STEM Career Ads. *Management Science, 65*(7), 2966-2981. Retrieved from https://doi.org.10.1287/mnsc.2018.3093

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lebovitz, S., Lifschitz-Assaf, H., & Levina, N. (2019). To Incorporate or Not Incorporate AI for Critical Judgments: The Importance of Ambiguity in Professionals' Judgment Process. *Available at SSRN 3480593*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3480593

LeCun, Y. (2018). The Power and Limits of Deep Learning. *Research Technology Management, 61*(6), 22-27. doi:10.1080/08956308.2018.1516928

Lee, F., Hallahan, M., & Herzog, T. (1996). Explaining Real-Life Events: How Culture and Domain Shape Attributions. *Personality and Social Psychology Bulletin, 22*(7), 732-741. doi:10.1177/0146167296227007

Lippke, J., Mongillo, J., Cullen, T., Waller, C., Harasewych, K., Muhammad, Z., & Bennett, J. (2020). Assessing Data Integrity Risks in an R&D Environment: A data-integrity risk assessment tool has been developed for use with standalone R&D dataacquisition and processing software. *Pharmaceutical Technology, 44*(8), 51-53. Retrieved from

https://ezp.lib.cam.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bsu&AN=145102023&site=bsi-live&scope=site

Liu, J., Zhou, H., Wan, M., & Liu, L. (2019). How Does Overconfidence Affect Decision Making of the Green Product Manufacturer? *Mathematical problems in engineering, 2019*, 1-14. doi:10.1155/2019/5936940

Liu, M., Brynjolffson, E., & Dowlatabadi, J. (2018a). *Do digital platforms reduce moral hazard? The case of Uber and taxis.* Cambridge MA

Marcus, G. (2018a). Deep Learning: A Critical Appraisal. *arXiv*, 1-27.

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*: Random House.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics, 160*, 835-850.

McArdle, D., Burns, N., & Ireland, A. (2003). Attitudes and beliefs of doctors towards medication error reporting. *International Journal of Health Care Quality Assurance, 16*(6), 326-333.

McCollum, T. (2020). AUDIT PLANS IGNORE KEY RISKS: Cybersecurity and third parties are among omissions, Pulse says. *Internal Auditor, 77*(2), 10-11. Retrieved from https://ezp.lib.cam.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bsu&AN=142569511&site=bsi-live&scope=site

McIntosh, S. (2013). *Hollowing out and the future of the labour market*. Retrieved from London: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/250206/bis-13-1213-hollowing-out-and-future-of-the-labour-market.pdf

Mills, J. (2020, Tuesday 25 August 2020). Head of exams regulator Ofqual resigns after results scandal. *Metro*. Retrieved from https://metro.co.uk/2020/08/25/head-exams-regulator-ofqual-resigns-results-scandal-13176774/?ico=more_text_links

Molnar, P., & Gill, L. (2018). *Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System*. Retrieved from Toronto, Canada: https://ihrp.law.utoronto.ca/sites/default/files/media/IHRP-Automated-Systems-Report-Web.pdf

Murray, J. (2020, Monday 24 August 2020). Royal Statistical Society hits back at Ofqual in exams algorithm row. *The Guardian*. Retrieved from https://www.theguardian.com/education/2020/aug/24/royal-statistical-society-hits-back-at-ofqual-in-exams-algorithm-row

Myers West, S. (2019). Data Capitalism: Redefining the Logics of Surveillance and Privacy. *Business and Society, 58*((1)), 20-41. doi:10.1177/0007650317718185

NAO. (2017). *A Short Guide to Regulation*. National Audit Office Retrieved from https://www.nao.org.uk/report/a-short-guide-to-regulation-2/

Nunn, R. (2018). Workforce diversity can help banks mitigate AI bias. *American Banker, 183(104),* 1-1.

O'Neill, C., & Mann, G. (2016). Hiring Algorithms Are Not Neutral. *Harvard Business Review*. Retrieved from https://hbr.org/2016/12/hiring-algorithms-are-not-neutral

Ofqual. (2020). *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf

OxfordAnalytica. (2018, 3 September 2018). INTERNATIONAL: Facebook seeks new users and markets as troubles mount. Retrieved from https://dailybrief.oxan.com/Analysis/DB238178/Facebook-seeks-new-users-and-markets-as-troubles-mount

Palermo, T. (2016). Technoculture: Risk reporting and analysis at a large airline. In M. Power (Ed.), *Riskwork: Essays on the Organizational Life of Risk Management*. Oxford, UK: Oxford University Press.

Palermo, T., Power, M., & Ashby, S. (2017). Navigating Institutional Complexity: The Production of Risk Culture in the Financial Sector. *Journal of Management Studies (John Wiley & Sons, Inc.), 54*(2), 154-181. doi:10.1111/joms.12241

Pernell, K., Jung, J., & Dobbin, F. (2017). The Hazards of Expert Control: Chief Risk Officers and Risky Derivatives. *American Sociological Review, 82*(3), 511-541.

Perrow, C. (1984). *Normal Accidents: Living with High Risk Systems*. New York: Basic Books.

Pinto, J. K. (2014). Project management, governance, and the normalization of deviance. *International journal of project management, 32*(3), 376-387. doi:10.1016/j.ijproman.2013.06.004

Power, M. (2004). The nature of risk: The risk management of everything. *Balance Sheet, 12*(5), 19-28. Retrieved from https://ezp.lib.cam.ac.uk/login?url=https://www.proquest.com/scholarly-journals/nature-risk-management-everything/docview/204699978/se-2?accountid=9851

https://libkey.io/libraries/603/openurl?genre=article&au=Power%2C+Michael&aulast=Power&issn=09657967&isbn=&title=The+nature+of+risk%3A+The+risk+management+of+everything&jtitle=Balance+Sheet&pubname=Balance+Sheet&btitle=&atitle=The+nature+of+risk%3A+The+risk+management+of+everything&volume=12&issue=5&spage=19&date=2004&doi=&sid=ProQuest

Power, M. (2007). *Organized uncertainty: Designing a world of risk management*. Oxford: Oxford University Press.

Power, M. (2009). The risk management of nothing. *Accounting, Organizations and Society, 34*, 849-855.

Power, M. (2016). Introduction. In M. Power (Ed.), *Riskwork: Essays on the Organizational Life of Risk Management*. Oxford, UK: Oxford University Press.

Preimesberger, C. (2018, 2018/11/07/). *Gartner Lists Top 10 Strategic IoT Technologies, Trends Through 2023*.

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, 48*, 137-141.

Reason, J. T. (1997). *Managing the risks of organizational accidents*. Aldershot, UK: Ashgate.

Reason, J. T. (1998). Achieving a safe culture: Theory and practice. *Work and Stress, 12*(3), 293-306.

Reason, J. T. (2000). Safety paradoxes and safety culture. *Injury Control and Safety Promotion, 7*(1), 3-14.

Renn, O. (2008). *White paper on risk governance: Towards an Integrative Framework*. Retrieved from europepmc.org: https://europepmc.org/article/PMC/PMC7120941

Renn, O. (2011). The social amplification/attenuation of risk framework: application to climate change. *WIREs Climate Change, 2*(2), 154-169. doi:https://doi.org/10.1002/wcc.99

Renn, O., Lucas, K., Haas, A., & Jaeger, C. (2019). Things are different today: the challenge of global systemic risks. *Journal of Risk Research, 22*(4), 401-415.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv*.

RIMS. (2011). *An Overview of Widely Used Risk Management Standards and Guidelines: A Joint Report of RIMS Standards and Practices Committee and RIMS ERM Committee*. Retrieved from https://www.scribd.com/document/325373240/An-Overview-of-Widely-Used-Risk-Management-Standards-and-Guidelines-1-pdf

Rousseau, D. M. (1979). Assessment of Technology in Organisations: Closed versus Open Systems Approaches. *Academy of Management Review, 4*(4), 531-542.

Russell, B. (2012). Professional call centres, professional workers and the paradox of the algorithm: the case of telenursing. *Work, employment and society, 26*(2), 195-210. doi:10.1177/0950017011433155

Samek, W., Wiegand, T., & Muller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv*.

Schwartz, J., Hagel, J., Wooll, M., & Monahan, K. (2019). Reframing the Future of Work: Future of work initiatives promise lots of noise and lots of activity, but to what end? *MIT Sloan Management Review*, 12-18. Retrieved from https://ezp.lib.cam.ac.uk/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bsu&AN=135258136&site=ehost-live&scope=site

Shestakofsky, B. (2017). Working Algorithms: Software Automation and the Future of Work. *Work and Occupations, 44*(4), 376-423.

Shi, Y., Young, H. W., Lantsman, Y., & Wei, C. (2016). Aggregate Model Risk Management Theory and Practice. *The RMA Journal, 99*(3), 44-49.

Silbey, S. S. (2009). Taming Prometheus: Talk of safety and culture. *Annual Review of Sociology, 35*, 341-369.

Simons, R. (1999). How Risky Is Your Company? *Harvard Business Review,* (May-June 1999). Retrieved from https://hbr.org/1999/05/how-risky-is-your-company

Smith, A. (2018, 30 August 2018). Franken-algorithms: the deadly consequences of unpredictable code. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger

Snow, J. (2018). Q+A: Timnit Gebru. In (Vol. 121, pp. 28-31): MIT Technology Review.

Sommer, S. C., & Loch, C. H. (2004). Selectionism and learning in projects with complexity and unforeseeable uncertainty. *Management Science, 50*, 1334-1347.

SRA. (Ed.) (2019).

Stanley, M., & Wdowin, J. (2018). *Getting Regulation Right*. Retrieved from Cambridge UK: https://www.bennettinstitute.cam.ac.uk/media/uploads/files/Getting_regulation_right_jMvOiGb.pdf

Sterman, J. (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Boston: Irwin McGraw-Hill.

Sterman, J. (2002). System Dynamics: Systems Thinking and Modeling for a Complex World. *MIT Engineering Systems Division Working Paper Series*, 1-29.

Sterman, J. (2018). System dynamics at sixty: the path forward. *System Dynamics Review, 34*(1/2), 5-47.

Sterman, J., Oliva, R., Linderman, K., & Bendoly, E. (2015). System dynamics perspectives and modeling opportunities for research in operations management. *Journal of Operations Management, 39*, 1-5. doi:10.1016/j.jom.2015.07.001

Stulz, R. M. (2009). 6 Ways Companies Mismanage Risk. *Harvard Business Review*(March), 86-94.

Sullivan, C. (2018). GDPR Regulation of AI and Deep Learning in the Context of IoT Data Processing - A Risky Strategy. *Journal of Internet Law, 22*(6), 7-23.

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM, 56*(6), 44-54.

Sweney, M. (2019, 8 July 2019). BA faces £183m fine over passenger data breach. *The Guardian*.

Tarafdar, M., Tu, Q., Ragu-Nathan, T. S., & Ragu-Nathan, B. S. (2011). Crossing to the Dark Side: Examining Creators, Outcomes, and Inhibitors of Technostress. *Communications of the ACM, 54*(9), 113-120. doi:10.1145/1995376.1995403

Travis, G. (2019). How the Boeing 737 Max Disaster Looks to a Software Developer. Retrieved from https://spectrum.ieee.org/aerospace/aviation/how-the-boeing-737-max-disaster-looks-to-a-software-developer

USFederalReserve. (2011). *SR Letter 11-7 Attachment: Supervisory Guidance on Model Risk Management*. Retrieved from https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf

Vaughan, D. (1997). *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: Chicago: University of Chicago Press.

Vaughan, D. (2005). Organizational rituals of risk and error. In B. Hutter & M. Power (Eds.), *Organizational Encounters with Risk* (pp. 33-66). Cambridge: Cambridge University Press.

Vial, G., Jiang, J., Giannelia, T., & Cameron, A.-F. (2021). The Data Problem Stalling AI. *MIT Sloan Management Review, 62*(2), 47-53. Retrieved from https://ezp.lib.cam.ac.uk/login?url=https://www.proquest.com/scholarly-journals/data-problem-stalling-ai/docview/2479113427/se-2?accountid=9851

https://libkey.io/libraries/603/openurl?genre=article&au=Vial%2C+Gregory%3BJiang%2C+Jinglu%3BGiannelia%2C+Tanya%3BCameron%2C+Ann-Frances&aulast=Vial&issn=15329194&isbn=&title=The+Data+Problem+Stalling+AI&jtitle=MIT+Sloan+Management+Review&pubname=MIT+Sloan+Management+Review&btitle=&atitle=The+Data+Problem+Stalling+AI&volume=62&issue=2&spage=47&date=Winter+2021&doi=&sid=ProQuest

Waring, J. (2005). Beyond blame: Cultural barriers to medical incident reporting. *Social Science and Medicine, 60*(9), 1927-1935.

Watanabe, M. (2019). Bad Revenue: When Creators Harass Queer People, YouTube Profits.

WEF. (2020). *The Global Risks Report 2020*. Retrieved from http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf

Weinstein. (2019, September/October 2019). AI or Just Sci-Fi? *Training,* 18-22.

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative Intelligence: Humans And AI Are Joining Forces. *Harvard Business Review*.

Yoost, D. A. (2013). Board Oversight of Model Risk Is a Challenging Imperative. *The RMA Journal, 96*(3), 20-26,13.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology, 30*, 75-89.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for A Human Future At The New Frontier of Power*. London: Profile Books.