European Business Organization Law Review (2021) 22:593–625
https://doi.org/10.1007/s40804-021-00224-0

ARTICLE

Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective

Eleanore Hickman[1] · Martin Petrin[2,3]

Accepted: 24 August 2021 / Published online: 6 October 2021

Abstract

AI will change many aspects of the world we live in, including the way corporations are governed. Many efficiencies and improvements are likely, but there are also potential dangers, including the threat of harmful impacts on third parties, discriminatory practices, data and privacy breaches, fraudulent practices and even 'rogue AI'. To address these dangers, the EU published 'The Expert Group's Policy and Investment Recommendations for Trustworthy AI' (the Guidelines). The Guidelines produce seven principles from its four foundational pillars of respect for human autonomy, prevention of harm, fairness, and explicability. If implemented by business, the impact on corporate governance will be substantial. Fundamental questions at the intersection of ethics and law are considered, but because the Guidelines only address the former without (much) reference to the latter, their practical application is challenging for business. Further, while they promote many positive corporate governance principles—including a stakeholder-oriented ('human-centric') corporate purpose and diversity, non-discrimination, and fairness—it is clear that their general nature leaves many questions and concerns unanswered. In this paper we examine the potential significance and impact of the Guidelines on selected corporate law and governance issues. We conclude that more specificity is needed in relation to how the principles therein will harmonise with company law rules and governance principles. However, despite their imperfections, until harder legislative instruments emerge, the Guidelines provide a useful starting point for directing businesses towards establishing trustworthy AI.

Keywords Corporate governance · Artificial intelligence · Company law · Ethics · European Union

✉ Eleanore Hickman
eleanore.hickman@bristol.ac.uk

Extended author information available on the last page of the article

1 Introduction

Artificial intelligence (AI) is becoming increasingly important for businesses. Most visible are the various AI-driven products and services—from self-driving cars to robotic trading of securities—that are either already in use or expected to emerge in the near future. AI is also increasingly common under the corporate surface.[1] AI applications are set to transform many aspects of companies' daily operations and business practices. Broadly speaking, AI can support three business needs: automating business processes, gaining insight through data analysis, and engaging with customers and employees.[2] Now, a fourth frontier is also beginning to emerge: AI as a tool to govern companies themselves through assisting or even replacing humans in corporate leadership.[3] Indeed, there are already examples of companies that claim to have appointed AI machines to managerial or board positions.[4]

Despite these advances, the use of AI today operates almost in a vacuum, with no or very limited regulation in place. Soft law-type guidance is also just emerging. Conscious of this gap, regulatory attention towards AI has globally been gaining momentum. As with many national jurisdictions and policy-making organisations, the European Union has begun to contribute to this discourse.[5] In 2017, the European Parliament recommended that the European Commission propose general principles on robotics and AI;[6] the Commission emphasized AI's strategic importance in a communication on Digital Single Market Strategy;[7] and the European Council invited the Commission to put forward a European AI approach.[8]

In April and December 2018, respectively, the European Commission released two documents outlining a broader AI strategy: A communication on 'Artificial Intelligence for Europe'[9] and a 'Coordinated Plan on Artificial Intelligence'.[10] Most recently, the Commission also released a white paper on AI[11] as well as a report on the safety and liability implications of AI and other technologies.[12]

The 2018 communication outlined three aims of a European initiative on AI, namely (1) boosting the EU's technological and industrial capacity and AI uptake across the economy; (2) preparing for socio-economic change brought about by AI;

[1] Kolbjørnsrud et al. (2016), p 17.

[2] Davenport and Ronanki (2018). See also Nilsson (2010), pp 510–511.

[3] See below, Sect. 3.3.

[4] See, for example, Burridge (2017); Tieto Press Release (2016).

[5] As one commentator has noted, there is now an abundance of AI and ethics related guidance, which makes it difficult for the relevant parties to know what applies to them. Smuha (2019), p 99. See also Veale (2020), who notes the large number of AI policy-making bodies in the UK.

[6] European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

[7] European Commission (2017), pp 21–22.

[8] European Council, Conclusions (EUCO 14/17) (19 October 2017), p 7.

[9] European Commission (2018b).

[10] European Commission (2018c). Also in 2018, 25 EU Member States agreed to cooperate on AI. See European Commission (2018a).

[11] European Commission (2020a).

[12] European Commission (2020b).

and (3) ensuring an appropriate ethical and legal framework.[13] It also emphasized the need for a European framework to support an AI approach 'that benefits people and society as a whole'[14] and is 'human-centric, inclusive'[15] in nature. Building on this, the Coordinated Plan on Artificial Intelligence set out the details of a broad range of intended measures 'to maximise the impact of investments at EU and national levels, encourage synergies and cooperation across the EU, exchange best practices and collectively define the way forward to ensure that the EU as a whole can compete globally.'[16]

In order to 'anchor […] more firmly in the development and use of AI' the principles of a human-centric and ethics-by-design approach, the Commission in 2018 appointed an independent AI high-level expert group and tasked it with developing draft AI ethics guidelines.[17] This group, the High-Level Expert Group on Artificial Intelligence (hereinafter the 'Expert Group'), was given the mandate to draft two deliverables: (1) ethics guidelines on AI, and (2) policy and investment recommendations.[18] It is the former, the Expert Group's Ethics Guidelines for Trustworthy Artificial Intelligence (hereinafter the 'Guidelines'),[19] published in April 2019, that this article will focus on. Although non-binding, the Guidelines will likely influence future legislation on AI, which the EU plans to develop in the coming years.[20]

While the Guidelines are applicable to AI systems in general and a wide range of 'AI practitioners',[21] this article will examine them from a company law and corporate governance perspective. Indeed, the European Commission has recently indicated that this is an area it wishes to explore further. In 2019, the Directorate-General Justice and Consumers launched a tender concerning the relevance and impact of artificial intelligence for company law and corporate governance,[22] and the resulting project is ongoing. Clearly, 'AI and company law' is an emerging area that is set to rapidly gain in importance, and ethical issues are certain to arise.

The article begins with an overview of the Guidelines' four ethical pillars and seven key principles that they promulgate. In the subsequent sections, the article then moves to examine the potential significance and impact of the key principles on

[13] European Commission (2018b), p 3.

[14] Ibid., p 2.

[15] Ibid., p 12.

[16] European Commission (2018c), p 2.

[17] Ibid., p 8. For further background, see Smuha (2019), pp 97–106. See also the criticisms in Molavi Vasse'i (2019) and Metzinger (2019).

[18] High-Level Expert Group on Artificial Intelligence (2019), p 4. The Expert Group's Policy and Investment Recommendations for Trustworthy AI, dated 26 June 2019, are available at https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence. For a critique of this document, see Veale (2020).

[19] See also European Commission (2019a).

[20] See European Commission (2019b); von der Leyen (2019), p 13.

[21] See High-Level Expert Group on Artificial Intelligence (2019), pp 5 and 36. AI practitioners are defined as 'all individuals or organisations that develop (including research, design or provide data for) deploy (including implement) or use AI systems, excluding those that use AI systems in the capacity of enduser or consumer' (p 36).

[22] Request for service JUST/2018/MARK/FW/CIVI/0190 (2019/06) (on file with authors).

selected corporate law and governance issues. First, in Sect. 3 the discussion focuses on corporate leadership and intra-corporate oversight of AI systems. Among others, this touches upon the crucial question of the extent to which AI should be subject to human intervention in light of the Guidelines' apparent rejection of fully autonomous AI systems. In Sect. 4 the article examines diversity, non-discrimination and fairness, with a special focus on diversity in corporate recruitment and on boards of directors. Section 5 explores the Guidelines' potential impact on the corporate purpose, noting that their focus on what in corporate governance parlance would be 'stakeholderism' is currently difficult to reconcile with legal requirements. In Sect. 6 we discuss the principles of technical robustness and safety, focusing on the issues of illicit uses of AI and 'rogue AI'. Section 7 looks at the principle of privacy and data governance, including their apparent tension with the Guidelines' emphasis on algorithmic transparency. The final section concludes.

2 The Guidelines

The heart of the Guidelines is set out in its 'Framework for Trustworthy AI' section.[23] The Framework is based around four ethical pillars[24] from which seven key principles emerge.[25] The Guidelines also contain methods for implementing the requirements and operationalising trustworthy AI in practice as well as a pilot version of a list to assess operationalization of different types of AI applications.[26]

The foundation to the Guidelines' four ethical principles consists of three overlapping components, which should be met throughout an AI system's lifecycle. According to these, AI should be: (1) *lawful*, complying with all applicable laws and regulations; (2) *ethical*, ensuring adherence to ethical principles; and (3) *robust*, both from a technical and social perspective, in order to avoid unintended adverse impacts.[27] While the Guidelines do not elaborate on lawfulness and instead focus on the ethics and robustness components, they note that 'while the two latter are to a certain extent often already reflected in existing laws, their full realisation may go beyond existing legal obligations'.[28]

[23] In terms of geographical scope, the Guidelines primarily apply to AI systems that are developed, deployed and used in EU Member States as well as systems developed or produced elsewhere but deployed and used in the EU. However, the Guidelines also aspire to be relevant outside the EU. High-Level Expert Group on Artificial Intelligence (2019), p 3, fn. 4.

[24] Ibid., p 2.

[25] Ibid., p 14.

[26] Ibid., p 24. For the purposes of the Guidelines, AI is defined as 'software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal'. High-Level Expert Group on Artificial Intelligence (2019), p 36 (footnote omitted).

[27] Ibid., pp 5–7.

[28] Ibid., p 6.

The Guidelines' approach to 'ethics' is grounded in fundamental rights, in particular: dignity, freedoms, equality and solidarity, citizens' rights and justice.[29] In the context of AI, the Guidelines extrapolate from these foundational pillars the four ethical principles of (1) respect for human autonomy; (2) prevention of harm; (3) fairness; and (4) explicability.[30] From these four pillars the Guidelines formulate a non-exhaustive list of seven key principles that AI systems have to meet to be deemed trustworthy. In short, these principles are:[31]

- *Human agency and oversight:* AI systems should support human autonomy and decision-making, with human oversight and intervention as integral elements.
- *Technical robustness and safety:* AI systems have to be resilient, reliable and secure, developed with a focus on preventing and minimizing unintended harm.
- *Privacy and data governance:* AI systems should provide adequate governance in terms of privacy and data protection, and quality, integrity, and access to data.
- *Transparency:* the data, system, and business models of AI systems should be transparent and explainable for stakeholders. Humans need to be informed when they interact with an AI system, and apprised of its capabilities and limitations.
- *Diversity, non-discrimination and fairness:* AI systems need to avoid unfair bias and provide for accessibility and universal design. Stakeholders who may be affected by an AI system should be considered and involved.
- *Societal and environmental well-being:* AI systems should be sustainable, environmentally friendly, considering broader society and other sentient beings. The impact on institutions and democracy also needs to be taken into account.
- *Accountability:* Mechanisms for ensuring responsibility, accountability, and potential redress for AI systems and their outcomes should be put into place. Negative impacts should be identified, assessed, documented, and minimized.

From the outset, the Guidelines also refer to their human-centric underpinning.[32] These have to be taken into account in further interpreting and applying the key principles. As the Expert Group writes in the introduction to the Guidelines:

> AI is not an end in itself, but rather a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. In particular, AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals […]. To do this, AI systems need to be human-centric, resting on

[29] Ibid., pp 9–11. As used in the Guidelines, fundamental rights are those that 'lie at the foundation of both international and EU human rights law and underpin the legally enforceable rights guaranteed by the EU Treaties and the EU Charter.' Ibid., p 7, fn. 12.

[30] Ibid., pp 12–13.

[31] Ibid., pp 14–20.

[32] The principle of human-centrism also forms the basis for the OECD's and the G20's AI Principles. See https://www.oecd.org/going-digital/ai/principles.

> a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom.[33]

Explaining its focus on the 'trustworthiness' of AI, the Expert Group notes:

> In a context of rapid technological change, we believe it is essential that trust remains the bedrock of societies, communities, economies, and sustainable development. We therefore identify Trustworthy AI as our foundational ambition, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place.[34]

To be sure, the Expert Group was aware that trust by itself is insufficient, or, to put it differently, has to be backed up by accountability mechanisms. Thus, the Guidelines require 'that mechanisms be put in place to ensure responsibility and accountability for AI systems'[35] and emphasize the importance of redress 'when things go wrong'.[36] This 'trust but verify' approach is welcome although, as we will discuss in more detail below, will in large part be based on legal, as opposed to ethical, principles. Legal aspects have consciously been removed and it was the drafters' intention for them to be beyond the (direct) scope of the Guidelines.[37] Nevertheless, in practice, there will almost inevitably be certain interactions between legal and ethical principles, and it is difficult to completely separate the two. In this vein, the following will also consider legal frameworks in evaluating the Guidelines and their applicability to corporate law and governance.

3 Corporate Leadership and Oversight

The Guidelines refer specifically, albeit briefly, to the role of corporate leadership in the context of recommended approaches to assessing and operationalizing trustworthy AI and the allocation of governance tasks within companies.

The Guidelines note that 'management attention at the highest level is essential to achieve change'[38] and suggest that management and the board should discuss and evaluate the development, deployment or procurement of AI systems and 'serve[] as an escalation board for evaluating all AI innovations and uses, when critical concerns are detected'.[39] Another recommended task for management and the board is to 'involve[] those impacted by the possible introduction of AI systems (e.g.

[33] High-Level Expert Group on Artificial Intelligence (2019), p 4.

[34] Ibid., p 4.

[35] Ibid., p 19.

[36] Ibid., p 20.

[37] Smuha (2019).

[38] High-Level Expert Group on Artificial Intelligence (2019), p 25.

[39] Ibid.

workers) and their representatives throughout the process via information, consultation and participation procedures'.[40]

3.1 Intra-Corporate Oversight of AI

The suggestion that AI issues deserve the attention of high-level management, especially when there are 'critical concerns', seems uncontroversial and is consistent with directors' (and other senior managers') existing oversight duties posited by corporate laws. These typically require that adequate risk management and internal control systems are in place and that the company's activities are actively monitored.[41] Yet, the Guidelines lack any details on the degree and manner of management and boards' involvement in overseeing AI systems and AI uses.

Business will face a number of questions in this regard. For example, what should be the division of responsibilities between directors and managers in this area? Should there be specific new roles, such as a 'Chief AI Officer', or a dedicated board committee that focuses on AI? Further, and importantly, to what extent are managers and boards equipped and able to carry out the functions that the Guidelines suggest they assume with regards to AI? At least initially, and until a business has obtained sufficient in-house expertise, it seems that this would necessitate either extensive training and/or reliance on third party expert advisors.

3.2 Involvement of Stakeholders

The involvement of stakeholders is an important consideration in the Guidelines, as it is reiterated in the principle relating to diversity, non-discrimination and bias, as well as societal and environmental well-being. In the corporate context, this would in part also fall upon the board or senior management to implement. However, it is not clear what form the involvement of 'those impacted by the possible introduction of AI systems' should take. In systems with employee representation on boards, such as in Germany, and companies that have established channels through which they consult with employees or their representatives, these could be utilized for the purpose of 'AI consultations' and similar involvement. Jurisdictions without such representation would need to find other forums for giving employees a voice on AI matters. In any event, with regards to other stakeholders, such as consumers or the public at large, new ways for adequate discourse may have to be established.

In addition to the procedural issues of *how* to involve affected parties, the more difficult question is what the substance and consequences of involving workers, consumers, etc. about the use of AI should be. 'Softer' forms of involvement, in the shape of information and disclosure, would seem relatively easy to implement, although these measures could already necessitate more than voluntary guidelines

[40] Ibid.

[41] For the UK, see generally Moore and Petrin (2017), pp 195–214. For a detailed exploration of managerial liability and AI, see Möslein (2018) and Petrin (2019).

or soft law to ensure their uptake. Conversely, consultation and participation that would provide stakeholders with the option of advisory or even binding input into the use of AI will likely be met with resistance from business.

For instance, if relevant stakeholders turn out to oppose the use of AI for specific tasks, to what extent should or must businesses take these preferences into account, and how should they balance efficiency-enhancing effects of new technologies against legitimate concerns by affected parties? The Guidelines' 'human-centric' approach and its ethical principles do not provide a clear answer to this question. In part, the question of stakeholder involvement relates to the fundamental issue of defining the overarching corporate purpose. We will return to this further below.

3.3 Autonomous AI Management vs. Human Agency and Oversight

Further issues emerge if we fast forward to future models of AI-governed businesses. Because the Guidelines appear to be solely focused on corporate leadership that *oversees* AI systems, they do not seem to be equipped for potential scenarios in which AI itself will lead and manage businesses.

Broadly speaking, there are three types of roles that AI technology can assume in corporate management: assisted AI; advisory or 'augmented' AI; and autonomous AI.[42] Within each of these roles the degree of the AI system's autonomy and proactivity differs significantly.

Assisted AI is at one end of the spectrum, with low or no autonomy and capabilities that are narrow and limited. Augmented AI, on the other hand, has a heightened level of autonomy and can provide support in solving more complex problems, but decision-making rights either remain with humans or are at most shared between them and the AI system. The 'augmentation' here refers to a combination of artificial and human intelligence, in which AI does not replace human intelligence, but leverages or improves it by, for example, giving information and advice that would otherwise be unavailable or difficult to obtain. Finally, autonomous AI can proactively evaluate options and make decisions by itself without human input, with decision rights fully allocated to the machine.

Today, a plausible case can be made that AI systems will in the future not only increasingly assume managerial tasks but that they might eventually take over corporate leadership altogether, with humans only playing a marginal or drastically limited role in this regard.[43] It is even possible that 'Algorithmic Entities' will emerge, i.e. legal entities that pursue business ventures fully autonomously and without any ongoing human input whatsoever.[44]

From this perspective, the Guidelines' emphasis on maintaining human involvement in AI seems incompatible with one of the possible future forms of corporate

[42] Kolbjørnsrud et al. (2016).

[43] See Petrin (2019), pp 971–976; Armour and Eidenmüller (2020), pp 105–109. But compare Enriques and Zetsche (2020), p 90 (arguing that new technologies are unlikely to replace existing corporate governance mechanisms).

[44] LoPucki (2018).

governance, namely leadership by autonomous AI systems. While the Guidelines are aligned with assisted and augmented AI models, they do not appear to allow fully autonomous AI. Generally, the Guidelines suggest that to be deemed trustworthy, AI systems necessarily need to work with and under the supervision of humans. In other words, fully autonomous AI systems without meaningful human input cannot be trustworthy. As the Guidelines explain:

> Humans interacting with AI systems must be able to keep full and effective selfdetermination over themselves […]. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight […].[45]

The principle of respect for human autonomy is further detailed in the Guidelines' requirement of *human agency and oversight*. It prescribes that AI systems should support human autonomy and decision-making, with human oversight and intervention as integral elements.[46] The element of 'human agency' aims to ensure that users are put in a position to make informed decisions. They should be provided with 'the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system'. Indeed, '[t]he overall principle of user autonomy must be central to the system's functionality' with the key point being 'the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them'.[47]

In the corporate context, an example of automated, AI-powered decisions that could lead to such impacts are AI applications in employee-related decisions. If AI is used to make 'hiring and firing' decisions, then this might well entail the legal or similarly significant effects to which the Guidelines refer. According to the Guidelines, completely automated processes are not allowed in these circumstances and affected parties would have to be given adequate information and rights including those related to 'challenging the system'.[48] Other examples of processes where human involvement in decision-making might be required, due to each areas' potential for having an impact on individuals' legal rights or similar other significant effects, include dealing with product-related customer complaints, interactions with investors, and handling of intra-firm whistleblowing procedures.

In turn, the element of 'human oversight' aims to ensure that an AI system does not undermine human autonomy or causes other adverse effects.[49] The Guidelines

[45] High-Level Expert Group on Artificial Intelligence (2019), p 12.

[46] Ibid., p 16.

[47] Ibid.

[48] Ibid.

[49] Ibid.

explain that oversight may be achieved through various approaches that differ in the level and mode of human involvement. Possible approaches mentioned in the Guidelines are (1) human-in-the-loop, which allows for human intervention in every decision cycle of the system; (2) human-on-the-loop, which entails possible human intervention during the design cycle and monitoring of a system; and (3) human-in-command approaches. The human-in-command approach

> Refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk.[50]

In the context of AI in corporate management, human-in-the-loop approaches are unlikely to make sense if there should be efficiency gains; indeed, as the Guidelines note, AI with 'the capability for human intervention in every decision cycle of the system, [is] in many cases […] neither possible nor desirable'.[51] Conversely, both human-on-the-loop and human-in-command (HIC) appear to be feasible when used as part of an AI-based corporate management system. However, regulators tackling the issue of human oversight in the future will need to develop clear criteria that balance the potentially competing interests of human-centrism versus utility and efficiency. Indeed, as one commentator noted, the 'robot-friendliness' of the law in different jurisdictions may lead to a new regulatory race to the bottom.[52]

In this context, it is important to note that the principle of mandating human oversight, which is firmly embedded in the Guidelines and the EU's approach to AI, is not without its own problems. Indeed, humans tend to perform poorly when tasked with exercising supervision of machines and automated processes or systems. It is thus not yet clear whether even systems with limited or minimal options for human intervention are the most desirable choices going forward.

As Riikka Koulu notes, '[f]unction allocation research as well as later work on teleoperations, human-machine interaction and cognitive engineering have demonstrated the inherent shortcomings of human oversight'.[53] The reasons for this can be varied, ranging from 'psychological reasons such as boredom at routine monitoring to automation bias and alert fatigue'.[54] However, 'the importance of keeping humans

[50] Ibid. See also the Guidelines' Assessment List, which seems to require 'a stop button or procedure to safely abort an operation where needed'. Ibid., p 27.

[51] Ibid., p 16.

[52] Eidenmüller (2017), pp 14–15.

[53] Koulu (2020).

[54] Ibid., p 9.

in control of automation is widely agreed upon in legal scholarship', although this seems to be based either on conventions and practical reasons including liability, or on 'a more fundamental argument' that is ultimately related to democratic legitimacy, rather than direct benefits of human oversight.[55]

Thinking about how human oversight could work in practice, questions that immediately arise are who, within a business entity, should be in charge of overseeing the AI systems? Furthermore, to what extent will these individuals be able to exercise adequate oversight and what should be the criteria for human intervention? The difficulty for business will be to avoid creating a new layer of corporate leaders with 'ultimate' decision-making power. This body, whatever shape it would take, will be prone to suffer from the same limitations and governance challenges that AI-management systems would seek to minimize.

Although the Guidelines' value-based emphasis on human agency and oversight is understandable and appears to be in line with human perceptions,[56] concerns about the wisdom of mandating them are also justified and regulators should keep an open mind while monitoring the latest technological developments and research on human-machine interactions. Another way of approaching this issue would be to allow for flexibility and ensure that proper hard law liability rules are in place, forcing business to choose the approach that minimizes negative externalities.[57]

4 Diversity, Non-Discrimination and Fairness

Fairness is one of the four main ethical principles of the Guidelines, upon which its recommendations are based. Although an obviously important concept, fairness can be problematic because of its lack of universal meaning. The Guidelines acknowledge this difficulty and clarify that the recommendations relate to substantive fairness, i.e. freedom from discrimination, equality of opportunity, absence of deception and a balance between competing interests.[58] Fairness in this sense is also described as having a procedural dimension, enabling individuals to challenge decisions with which they disagree.

Speaking directly to the substantive element of the fairness principle, the Guidelines suggest that in order to build trustworthy AI, it must be designed to ensure equal access and equal treatment. This principle is elaborated with three general recommendations. Firstly, the use of AI should avoid unfair bias, counteracting the threat that '[d]ata sets used by AI systems […] may suffer from the inclusion of

[55] Ibid.

[56] See Lee (2018) (suggesting that we perceive algorithmic decisions as less fair and trustworthy for tasks that are thought to require 'uniquely human skills', such as hiring work evaluation).

[57] On this, see Casey (2016), p 1350: '[P]rofit-maximizing firms will design their robots to behave not as good moral philosophers, but as Holmesian bad men—concerned less with "ethical rule[s]" than with the legal rules that dictate whether they will be "made to pay money"'. Compare however Smuha (2019), p 101, who suggests that the ethics and law of AI need to work in tandem.

[58] High-Level Expert Group on Artificial Intelligence (2019), p 12.

inadvertent historic bias, incompleteness and bad governance models'.[59] Secondly, AI should be accessible, 'enabl[ing] equitable access and active participation of all people'.[60] Thirdly, stakeholder participation should be enabled. This may mean allowing for consultation with those potentially affected by the system, throughout its lifecycle.

The following will consider these substantive principles of fairness and the possibility of achieving them when making decisions based on the processing of historical data.[61] In light of the ongoing debate about diversity within corporations, the focus will be on the potential impact of AI on the specific diversity and discrimination issues relating to employee recruitment and board composition, and how these issues may be influenced by the Guidelines.

4.1 Diversity

For some time, diversity has been a hot topic in corporate governance.[62] Government commissioned reports suggest that different leadership experiences and variation in gender, ethnicity, race, nationality and socio-economic backgrounds can provide effective means to tackle complacency, generate new ideas, and result in better risk management.[63] The implication is that there is not one right or best answer to board deliberations. Human decision-making is limited by imperfect information and bounded rationality.[64] Therefore, including a variety of perspectives in decision-making tends to improve the quality of decisions.

Much of the board diversity rhetoric focuses on the economic benefits that diversity of background and perspective can bring. This is known as 'the business case'.[65] According to this view, board decisions improve when boards can draw on a mix of directors with a depth of background and perspectives. Thus, increased board diversity should have a positive impact on firm value or similar metrics.[66]

There are now policies targeting a relatively broad conception of diversity, which go beyond the initial focus on female directors. For instance, the UK Corporate Governance Code provides that board appointments and succession plans should, among other considerations, promote diversity of gender, social and ethnic backgrounds.[67]

One of the many considerations in relation to the role of AI on the board is whether AI can emulate the broader level of perspective that is presently sought

[59] Ibid., p 19.

[60] Ibid.

[61] Ibid., p 13.

[62] After the boardroom gender diversity debate had been developing for some time, in 2006 Norway became the first country to impose a gender target on their corporate boards, with many jurisdictions following suite or imposing a variety of other measures, to varying levels of success. For more on this see Machold et al. (2013); Thomas (2020).

[63] Vinnicombe, Atewologun and Battista (2019); Sir John Parker (2016).

[64] Williamson (1985).

[65] Francoeur, Labelle and Sinclair-Desgagne (2008).

[66] Adams and Funk (2012), p 219.

[67] UK Corporate Governance Code Principle J.

through diversity, by introducing into the decision-making process a previously unavailable ever growing and developing bank of information.[68]

If AI cannot account for and include a variety of perspectives into its output, one might question what role it should have on the board at all. If it can, the business case for diversity on the board would no longer be present. Although there are other reasons for maintaining diversity, including equality and social justice, these have not dominated the boardroom diversity debate. Consequently, it is arguable that the addition of AI into the boardroom makes it all the more necessary to consider diversity as a value in itself.[69] This is true not just for board appointments, but also for the earlier stages of AI design and implementation. From a board composition perspective, valuing diversity without reference to the business case will become particularly critical should algorithms begin to reintroduce non-diverse boards, for example because 'merit', for the purposes of directorial appointments, is defined according to all white male board norms.[70]

4.2 Non-Discrimination

Another diversity related consideration for the AI board is the effect of historic bias in the data it processes. The Guidelines seek for historic bias to be counteracted and it is easy to see why. AI is used to process data and provide objective answers to the questions it has been tasked to answer. Although evidence based, decisions made on this information may produce unfair effects on particular groups of people.[71] As Mittelstadt states, 'for affected parties, data driven discriminatory treatment is unlikely to be more palatable than discrimination fueled by prejudices or anecdotal evidence'.[72]

There are already a number of real world examples of this, such as: an algorithm used to assess the potential for criminals to reoffend which produced ethnically biased risk assessments;[73] Google's targeted advertisements for highly paid CEO jobs that were more likely to be seen by men than by women;[74] and facial recognition technology that has been shown to have issues recognizing faces of people of colour.[75]

In addition to the impact on affected parties and broader ethical and social ramifications, corporations that rely on AI decision-making that treats certain groups of individuals differently from others will face corporate reputation and liability issues. It remains to be seen whether boards will be able to deflect liability by referencing or 'blaming' unintended consequences of AI applications. Even if that is possible,

[68] Brown and Duguid (2000).

[69] For more on diversity benefits unrelated to performance, see Hickman (2014), pp 389–391.

[70] For a discussion of the meaning of merit and its impact on diversity see Hickman (2020).

[71] Mittelstadt et al. (2016), p 4.

[72] Ibid., p 8.

[73] Yong (2018).

[74] Datta et al. (2018).

[75] Buolamwini and Gebru (2018).

business will see decisions affecting certain groups more than others as undesirable from a reputational perspective. Arguably, it is difficult to avoid this on the basis that 'if two groups of people are measurably different, then any rule about how they are treated—be it an algorithm or human judgement—will end up looking unfair, if not by one measure then another'.[76]

Attempts to address data bias in AI usage are beginning to emerge. For example, consultancy firm Accenture has developed a tool which 'measures disparate impact and corrects for predictive parity to achieve equal opportunity'.[77] Such a tool could make compliance with the Guidelines much simpler in this respect. But, as long as such tools are not widespread and reliable, it will be important for corporations to find other ways to mitigate and redress the potential imbalances that AI may create. This may not be possible at an individual corporate level as biases and discriminatory treatment are often only visible from a multi corporate level.[78] Addressing this sort of discrimination may require some collective action, for which government intervention may be necessary. To the extent corporations wish to reduce the remit of regulation on their business, they have an economic incentive, on top of their moral one, to try to minimize the unfairness that may be created by the use of AI.

Although it is necessary to be wary of the biases that may result from the use of biased data in AI, it is also important not to ignore AI's potential to overcome biases. One of the major flaws in human decision making is its susceptibility to conscious and unconscious biases. There are many ways in which AI, if properly designed and implemented, can help to mitigate these issues notwithstanding any biases within the data it processes. As stated by Abbott, '[a] way to manage human bias is to cede some agency to AI, which can be explicitly programmed to never consider race or even proxy variables; doing so might be the best chance for society to avoid discrimination in a racially stratified world'.[79]

4.3 Equal Access and Treatment

It has been said that the most important aspect of management is getting the right people involved.[80] The Guidelines specifically state that 'hiring from diverse backgrounds, cultures and disciplines […] should be encouraged' and this is included in the context of helping to ensure AI systems do not suffer from unfair bias.[81]

As mentioned above, one of the more publicised areas of AI involvement in business appears to be recruitment. This is one area where AI could, theoretically, prove helpful in overcoming the unconscious biases of human decisions. However, experience to date suggests that this is more complicated than at first imagined. Amazon

[76] Harford (2019).

[77] Chowdhury and Mulani (2018).

[78] Such as was the case in the research that found job applicants were being discriminated on the basis of the ethnic associations of their name. See Bertrand and Mullainathan (2003).

[79] Abbott (2020) p 253.

[80] Collins (2001).

[81] High-Level Expert Group on Artificial Intelligence (2019), p 18.

started using AI in recruitment in 2014, but the results were highly dominated by gender—with the system favouring men—because of the imbalanced input.[82] Attempts have moved on and human resources professionals now use AI to identify suitable job candidates through analysis of their facial expressions, tone and language during interviews.[83]

Given the antecedent levels of diversity on the board, the use of AI in recruitment decisions could be a cause for concern if it relies on historic data and the status quo to draw conclusions about the 'right' type of board member. David Cox, of the Massachusetts Institute of Technology, has said AI 'is guaranteed to take the past and give it back to you'.[84] If that is true, ethnic minorities and women may be justifiably concerned. The Guidelines suggest this issue may be mitigated through diversity in the AI design teams, describing it as 'critical' that 'the teams that design, develop, test and maintain, deploy and procure these systems reflect the diversity of users and of society in general'.[85]

Historically, the issue of the glass ceiling has been a significant problem for women and ethnic minorities in corporations.[86] Diversity of senior roles has not reflected the diversity of the workforce, and the issue becomes more pronounced the further up the hierarchy one travels. This has improved, but very slowly and far from sufficiently.[87] Future corporations, in which AI takes a central role in the recruitment of board members, may tell a different story. It is also possible that the effect of AI recruitment may only be to shift the attention from diversity on the board to diversity in the teams that are responsible for AI design, usage, etc. This will certainly be the case should fully autonomous AI boards materialize.

The closer we get to AI boards, the smaller the human element of the board will become and the more challenging the imposition of diversity requirements. For instance, research from the UK suggests that only 17% of technology or information and communications technology professionals are female, and only 1 in every 10 students taking a computer science A-level is female.[88] Given this dominance of men in technology fields, AI design team diversity may prove an even more challenging problem to solve than boardroom diversity. However, AI itself is providing sources of hope here. For example, IBM has introduced a 'Tech Re-Entry Program' driven by AI, which seeks to help women and long-term unemployed return to work through tailoring courses based on the skills they have.[89]

[82] Hill (2019).

[83] Hymas (2019).

[84] Ball (2019).

[85] High-Level Expert Group on Artificial Intelligence (2019), p 23.

[86] Yap and Konrad (2009).

[87] Vinnicombe, Atewologun and Battista (2019).

[88] Tech Talent Charter (2019).

[89] Mulholland (2020).

5 Corporate Purpose and Stakeholders

The Guidelines' principle entitled 'societal and environmental wellbeing' states that 'broader society, other sentient beings and the environment should be considered as stakeholders throughout the AI system's life cycle'.[90] This requirement operates on a number of levels. Firstly, the principle requires the consideration of sustainability and ecology in the use of AI. This relates to the ongoing debate concerning the corporate purpose, as discussed below. Given the current and ever-increasing focus on sustainability it seems a natural addition to the Guidelines and its 'human-centric' nature. It is also something boards are already likely to have high on their agenda. A second aspect of the principle requires consideration of the practical consequences of the use of AI on society, noting that AI systems can both enhance and deteriorate social skills.[91] This requirement may be more relevant for some companies than others, depending on the sector or industry, but boards will need to take care to identify potential problems before they arise.[92]

What is also relevant from a corporate governance perspective, and will be further considered below, is the requirement to 'take into account institutions, democracy and society at large'.[93] For business, such a broad requirement could conflict with the need for AI to have clear and specific goals, particularly within the context of defining the corporate purpose.

Narrowing down and clearly articulating the corporate mission is only the first part of the problem, the second is ensuring that, once a corporate purpose has been defined, AI accounts for and properly reconciles any relevant interests that may intersect or overlap. If this is done incorrectly then AI will not produce the desired outcomes a business is aiming for. Given the enormous complexity of the task, it may produce unintended and undesirable consequences.

According to Russell, the main problem for AI is not conflicting interests but, rather, the problem of accounting for the full scope of interests at stake. As he notes,

> Satisfying conflicting goals is not the problem—that's something that's been built into the standard model from the early days of decision theory. The problem is that the conflicting goals of which the machine is aware, do not constitute the entirety of human concerns.[94]

The Guidelines' requirement to consider institutions, democracy and society will be considered here from two corporate governance perspectives. Firstly, the broader issue of how the use of AI within the corporation is governed. The issues arising therefrom are described as 'moral trade-offs', a concept further discussed below. Secondly, the requirement will be considered from the perspective of AI as a future

[90] High-Level Expert Group on Artificial Intelligence (2019), p 19.

[91] Ibid., p 19; Lindebaum, Vesa and den Hond (2019), p 24.

[92] The related consideration of how AI on boards may negatively affect human knowledge is considered in Sect. 6.1.

[93] High-Level Expert Group on Artificial Intelligence (2019), p 19.

[94] Russell (2019), p 168.

'member' of the management team or board, or as a tool upon which corporate leaders rely. This will require it to take account of stakeholders directly. This is described as the 'stakeholder trade-offs' and is discussed in Sect. 5.2.

5.1 Moral Trade-Offs

Long before the potential introduction of AI in business, the corporate objective has been considered by scholars, policy-makers, and managers as part of the question of corporate purpose. But with the new capabilities and uses of AI, as well as its potential limitations, this question appears in new contexts. At some stage boards will be faced with a choice as to the basis upon which the AI it deploys tackles trade-offs and conflicts of interest. That is the question considered here.

AI provides an exciting opportunity to process information and provide answers founded upon much deeper resources than humans have access to, and without their fallibilities. However, in order to provide answers, AI requires an objective. Decisions are often not straight-forward. Conflicts of interest will need to be weighed and trade-offs made. The basis upon which these trade-offs are to be made will be a consequence of the decision's objective. Determining the appropriate objective necessitates certain ethical considerations. This goes beyond any material objective of the corporations, such as profit maximization. The question is, on what basis should AI be programmed to make trade-offs which may impact third parties and/or stakeholders in different ways. Should the basis of the trade-off be the greatest good to the greatest number? The Guidelines imply a more deontological approach, in their use of the word fairness. But they are not clear in this regard.

Taking a somewhat pessimistic perspective, some commentators argue that AI designed by corporations 'will not maximise morality but minimize liability'.[95] Casey describes a self-driving car having to decide whether to potentially kill 5 pedestrians who have run into the road, or the occupant of the car. He contends that an AI system will decide based on the potential cost to the car manufacturer, meaning the outcome may differ depending on whether the car is driving in a jurisdiction of strict liability, or one of contributory negligence.[96] This problem is a version of the well-known 'trolley problem'.[97] Even without reference to this thought experiment, it is easy to see how a self-driving car designed to minimize legal liability would be problematic. For example, such a car might be more likely to injure or kill those on lower incomes (with less resources available to initiate legal action and lower expected damages for loss of earnings).[98]

[95] Casey (2016), p 1350.

[96] Ibid.

[97] The trolley problem, as originally devised by Phillipa Foot (1967), involves a trolley going down a track which has five people working on the track ahead. There is no way to stop the trolley or for the people to get out of the way, but there is a fork in the track ahead. If the driver pulls a lever to divert the trolley the five people on the original track would be saved, but there is one person working on the diverted track, who would be killed instead. The problem is the moral question of what the driver should do.

[98] Davis (2018).

If boards or management are responsible for providing the parameters for AI decision-making, they will need to take these trade-offs into account. In considering how AI can be programmed to deal with moral questions, Davis suggests that there are three options.[99] Firstly, there is the top-down approach, by which AI is loaded up with principles from the start and simply has to apply those principles. Here, the Guidelines could offer a useful starting point and basis for a common approach, although until they or other regulatory instruments provide more detailed instructions, businesses will in large part be left to their own devices. Perhaps an entity's choice of AI decision-making priorities will become a new basis upon which customers determine which companies they will voluntarily engage with. The second option is a bottom-up approach where AI develops its principles based on its own experiences. The third option involves AI predicting moral decisions on a context basis, with human oversight. As already discussed, human involvement may eliminate some of the benefits of deploying AI in the first place.

Ultimately, how AI tackles value judgements is a question that needs to be considered at the highest level within a corporation. It is possible, even likely, that this will be covered in the near future, to an extent, by regulation. But insofar as it is not, boards will need to engage with these complex and controversial debates.

5.2 Stakeholder Trade-Offs

For corporations governed by humans alone, taking consideration of a wide range of stakeholders has long been challenging. If AI is involved at board level, this important question must be considered afresh.

The Guidelines aim to ensure that business' use of AI increases 'individual and societal well-being and the common good', and also makes reference to AI's potential for enhancement of production processes and improvement of welfare.[100] Boards and managers are therefore left with the difficult task of having to solve situations where tensions between these goals arise, such as in the case of decisions on whether to use AI systems that would increase profitability but also lead to job losses. Although the Guidelines clearly speak to these situations, boards also, and primarily, have to adhere to applicable legally binding rules, including those that govern the corporate purpose and directors' and officers' fiduciary duties. In many jurisdictions, and arguably in contradiction to the Guidelines' general thrust, this would currently mean that shareholder wealth needs to be prioritized. In some jurisdictions—notably including the UK—the board has a duty to 'have regard' to the interests of other stakeholders, although this duty is rather weak and does not meaningfully protect non-shareholders.[101]

AI may prove highly beneficial here. If the stakeholders for which AI is to 'have regard' need to be identified ex ante and programmed in, it follows that in a world with AI in the boardroom, corporations should necessarily shift to a broader

[99] Ibid., p 186.

[100] High-Level Expert Group on Artificial Intelligence (2019), p 4.

[101] Choudhury and Petrin (2019), pp 47–49.

purpose. In conjunction with the Guidelines' main principles, this should also lead to increased transparency and accountability.[102] For example, if a law firm claims to value employee health and work life balance, this should be programmed into AI decision making. A likely consequence would be lower target billable hours, more staff being hired and lower profits overall. This need for clarity could help to cut through the 'happy talk'[103] so that employees, customers and suppliers can identify what a company's values really are and make their choices accordingly.

There will be few decisions in the life of corporate management in which the interests of two or more stakeholders will not be in conflict. A common example is the conflict between shareholders and employees when a company would be more profitable with fewer employees. Another scenario pertains to shareholders and customers, where customers would be better off if the company did not market their product using unfounded claims concerning health benefits. There are many other possible examples. However, as mentioned above, today these decisions should be made on the basis of the still dominant shareholder wealth maximization ('SWM') principle. Shareholders are said to be in this privileged position because of the business capital they provide that is at risk. Contrasting theories argue that businesses have responsibilities to a variety of stakeholders such as employees, communities, and governments. This is often referred to as the stakeholder model.[104] There are some indications that the discussion is moving away from SWM towards the stakeholder model,[105] but reality still shows little sign of genuine change.

Ensuring that AI is stakeholder orientated within a SWM framework seems a conflict which may be difficult to resolve. It is therefore possible that the increased use of AI will move the dial of the corporate purpose discussion towards a more stakeholder focused approach. However, without legal reform, potential problems may arise since the law currently requires directors to prioritize the interests of shareholders.[106] To the extent the use of AI, in accordance with the Guidelines, alters the legislative choice, the Guidelines' legitimacy may be called into question. From a practical perspective, the use of AI in board decision-making should at least make consideration of goals that go beyond shareholder wealth more feasible. This would be a positive step forward, but it is a step that needs to be taken with caution.

A further cause for caution is the potential for self-serving behavior in which directors prioritize interests other than those of the company and its stakeholders before the AI is even switched on. Decisions on the way in which AI is designed and programmed can be unduly influenced or manipulated. According to Enriques and Zetzsche, '[i]f management is in control of that decision, we expect it to choose coders and technology designs catering to its own interests, which may not be perfectly aligned with the interests of the shareholders'.[107] Similarly, management may not

[102] Accountability has been discussed above at Sect. 3.

[103] Bell and Hartmann (2007), p 911.

[104] On this see Keay (2010).

[105] Business Roundtable (2019).

[106] Companies Act 2006, section 172.

[107] Enriques and Zetzsche (2020), p 90.

ensure an appropriate balancing of impacts on those affected by corporate activities. Preventing this and other unwanted external motivations from being programmed into AI, may require a high level of expert technology oversight.

One of the dominant arguments in support of SWM is that of accountability. It is thought by some that management cannot effectively serve two or more masters because one can be played off against another.[108] Others disagree on the basis that making judgements using multiple subjective criteria is commonplace.[109] When the issue was considered at the drafting stage of the UK Companies Act 2006, the Steering Committee felt that retaining focus on shareholder profit eliminated the distraction and cost of having to reconcile conflicting interests.[110] Yet, as has been previously argued, AI may be able to 'pursue multiple objectives simultaneously […] and optimize the outcomes of several objectives at once'.[111] Despite the desirability of this, there remains a risk that if the objective of the optimization of outcomes is insufficiently clear, AI decision-making may be suboptimal. As one commentator already observed in 1960, 'if we use, to achieve our purposes, a mechanical agency whose operation we cannot efficiently interfere […], we had better be quite sure that the purpose put into the machine is the purpose we really desire'.[112]

Relatedly, where AI does take account of and properly balances the needs of a variety of stakeholders, the same approach may still lead to completely different outcomes depending on the jurisdiction. Bruner notes the difference between the UK and US corporate governance approaches and attributes it to the extent to which stakeholders are taken care of by the state.[113] He describes the US system, whereby healthcare is provided for by corporations, as having 'inextricably bound US corporate governance to the achievement of a range of social goals that lie well beyond what Britons expect their own corporations to accomplish'.[114]

If AI systems optimize stakeholder welfare, the outcome will look different in the UK and the US, whose welfare states are very different. Where there is less variation in the decision-making process because, for example, all the FTSE 100 and Fortune 500 enterprises are using the (fictitious) 'CorporateBoard3000' AI management software, context may have more impact than it currently does, unless of course the software can account for these differences. In an increasingly globalized world this may impact decisions regarding where to conduct business and may even have policy implications. Furthermore, this suggests that when it comes to AI and corporate management, one size does not fit all.

[108] Bainbridge (1993), p 1427.

[109] Stout (2012).

[110] The Company Law Review Steering Group (1999), p 44.

[111] Petrin (2019), p 970.

[112] Wiener (1960), p 1358.

[113] Bruner (2009), p 646.

[114] Ibid., p 649.

5.3 Impact on Society

As is the case in other jurisdictions, English law has long moved away from directly defining corporate purpose. Previously, the objects of the company and the concept of *ultra vires* used to restrict boards in their decision making so as to ensure they remained within what had been envisaged as the raison d'etre of the company.[115] Allowing room for business judgement was at the heart of this change. However, while discretion may be beneficial, giving AI discretion is challenging and raises a number of concerns.[116] Decision making with the use of discretion can be viewed as the operation of 'substantive rationality'. Lindebaum et al. describe substantive rationality as finding the correct solution, bearing in mind '"what is", "what can" and "what ought to be" in empirical, moral and aesthetic terms'.[117] In the context of the board, this allows their decisions to take account of normative concerns.

It is impossible to say for sure what the limitations of AI are going to be and whether or not AI will be able to consider normative concerns is an area for debate. AI, as it is currently conceived, operates on the basis of 'formal rationality', meaning it seeks the 'logically or mathematically correct solution to a dataset' given the relevant constraints, boundaries and conditions.[118] Arguably, more extensive use of AI in corporate management going forward will mean a shift from substantive rationality towards formalized rationality.

It seems there is a balance that needs to be considered. On the one hand, humans face difficulties when exercising discretion and making decisions because of conscious and unconscious biases as well as self-interest.[119] An AI system that can take into account a vast amount of data may indeed be better at this process. On the other hand, should we accept a future in which normative concerns are only formally considered? Humans will make more mistakes than AI, but mistakes are not always a bad thing. Had Alexander Fleming not accidentally left his petri dish full of streptococcus bacteria unwashed while he went on holiday, penicillin may not have been discovered.[120] Clearly, the potential positives arising from mistakes is an argument that is less persuasive in the context of the boardroom than in the context of research and development.

In the extreme scenario, where corporate boards are, one day, entirely AI based, this could mean an absence of substantive rationality. The consequences of this could be severe but the prospect remains distant.[121] A more likely scenario in the

[115] For an understanding of the *ultra vires* doctrine (although as applied in the US), see Greenfield (2001).

[116] Armour and Eidenmüller (2020).

[117] Lindebaum, Vesa and den Hond (2019), p 6.

[118] Ibid., p 5.

[119] Williamson (1985).

[120] Science History Institute (2016). The force of this argument has weakened further since AI led to the recent identification of a powerful new antibiotic. Trafton (2020).

[121] Armour and Eidenmueller (2020), p 108, consider it likely that competition will push companies towards entirely AI based boards and raise the issue of giving them the same level of discretion as humans.

nearer term is one in which AI is used to recommend particular courses of action.[122] This is apparently already happening on the board of software company Salesforce, where a robot named 'Einstein' is consulted to appraise corporate plans in order to aid decision making.[123] In this 'human-in-command' scenario, directors will make the final decision and at that point can contribute substantive rationality. Difficulty will arise when their substantive rationality departs from what the super-computer informs them is the objectively best course of action within the parameters of the decision. It is not hard to imagine shareholder action against the board for deviating from what AI suggests. This raises a number of issues in relation to accountability, already discussed above.

Beyond accountability concerns, it is arguable that even to the extent humans have the opportunity to input substantive rationality, there will be a tendency not to, because of a potential lack of willingness to contradict the machine. People tend to 'overtrust' decisions made by machines.[124] With reference to decision-making, Lawrence argues that while AI may be able to produce a consistent decision, consistency is not the same as truth.[125] Reliance upon consistent AI decisions could have a significant impact over time. According to Lindebaum, it 'represents an orientation to decision making that is inhospitable to the diversities, complexities, spontaneity, vicissitudes and richness of life, and may generate a kind of human life that many would reject'.[126] Such decisions, taken over and over again will begin to 'reshape the way the world is conceptualised and eventually, how it is socially and politically organised'.[127]

This is an especially important concern from a corporate governance perspective given the impact corporations have on the lives of the individuals within the societies they operate in. If AI is only capable of employing formal rationality, thought needs to be given to methods of mitigating this narrowness. Russell believes that it is possible to recalibrate the current course of AI and steer it only towards systems that 'defer to humans and gradually align themselves to user preferences and intentions'.[128] This would ensure human input but would discourage human intervention in the ways already discussed and thereby may not solve the problem. This conception of deferential AI goes beyond that which is considered by the Guidelines,[129] which again prove to be restrictive in this regard.

[122] Libert, Beck and Bonchek (2017).

[123] Paquette (2018).

[124] Thornhill (2020).

[125] Lawrence (2017).

[126] Lindebaum, Vesa and den Hond (2019), p 16.

[127] Mittelstadt et al. (2016), p 4.

[128] Russell (2019), p 179.

[129] High-Level Expert Group on Artificial Intelligence (2019), p 12. See Sect. 3.

6 Technical Robustness and Safety

One of the four ethical foundations of the Guidelines is the prevention of harm. The principle on technical robustness and safety relates directly to this foundation. The Guidelines acknowledge that AI has the potential to cause harm and directs the usage of AI towards consideration, prevention and minimization of risk.[130] Specifically, the Guidelines suggest that AI should be resilient to attacks (such as hacking), accurate (particularly when the system may affect human lives), reliable and reproducible, and with safeguards in place to enable a fallback plan in case of problems such as adversarial attacks or other unexpected situations.

Theories abound about the nature of the risks AI poses and these can be applied in degrees to AI and its involvement in corporate governance and leadership. At one extreme are apocalyptical theories that paint pictures of deviant machines working against humankind,[131] and at the other, there is the view that AI is a force for good and fears are unfounded.[132] The problem is that, to a large extent, what is being discussed is speculative. At this stage, the form and role that AI is going to take within the corporation is unknown.

AI may, as already discussed, be in the assisted, advisory or autonomous form.[133] Many of the more dramatic concerns have fully autonomous AI at their heart. To the extent this will be possible at all, it is the most distant AI future in corporate governance terms. It also seems incompatible with the human oversight emphasis of the Guidelines. Nevertheless, as argued above, it is a (potentially beneficial) possibility, and as such it is necessary to carefully consider, at the planning and development phase, the potential for creating and causing unintentional harm or allowing for the development of intentional problems.

6.1 Unintentional Problems

There is a perception about the infallibility of AI and yet there are various examples in which AI goes wrong for unknown reasons. Section 4 has already discussed failures relating to, or stemming from bias, but there are many other errors where the origin is much harder to discern. That the error may have originated from the design of AI may prove little comfort to victims of AI-controlled corporate activities or to boards found responsible for deploying them. It is not difficult to understand why errors may arise. There is considerable pressure on companies, and even countries, to be at the forefront of AI development, and therefore the probability of corners being cut may be quite high. This is of such broad concern that unusual agreements are emerging, such as the 'Rome Call for AI Ethics' which saw two of the leading

[130] Ibid., p 16.

[131] See for example Russell (2019).

[132] Floridi (2016). Other examples include 'IBM's Open Letter to Congress on Artificial Intelligence', THINKPolicy, 27 June 2017, https://www.ibm.com/blogs/policy/kenny-artificial-intelligence-letter (accessed 23 January 2020).

[133] See Sect. 3.

AI developers (IBM and Microsoft) pledge to the Vatican that AI would be developed in a way that will 'safeguard the rights of humankind'.[134] Given so much rests on trust in the pledges of large corporations, it is arguable that the Guidelines don't go far enough. As Russell puts it vividly:

> Whereas we are accustomed to the idea that pharmaceutical companies have to show safety and (beneficial) efficacy through clinical trials before they can release a product to the general public, the software industry operates by a different set of rules—namely the empty set. A 'bunch of dudes chugging Red Bull' at a software company can unleash a product or an upgrade that affects literally billions of people with literally no third-party oversight whatsoever.[135]

Our increasing dependence and reliance on AI is almost certain to bring with it economic and other gains, but the more we rely on it, the greater the impact of failure will be. Russell demonstrates this with the example of airlines in which the level of AI involvement in operations has escalated quickly and dramatically. In 2018, this level of reliance meant that a single computer error 'caused 15,000 flights in Europe to be significantly delayed or cancelled'.[136] Similarly, trading algorithms went mysteriously wrong in 2010, 'wiping out $1 trillion in a few minutes' and leaving as the only solution 'to shut down the exchange'.[137] There have also been fatal consequences to Tesla's 'autopilot system limitations' upon which a 'driver' was overrelying and consequently crashed in California in 2018.[138]

From a corporate governance perspective, the importance of these examples are in the understanding that AI can go wrong with dramatic consequences. At board level, if directors start to rely on AI to produce answers, they may have little understanding of the causes of potential failures and expend less cognitive energy trying to foresee areas of potential failure. Some already see that the use of technology is making the pace of business change so fast and complex that boards are struggling to keep up.[139] It is possible that, the more reliance is placed on machines to produce answers or to replace humans in key responsibilities, the more disengaged from the decision making and operation process humans become, resulting in a 'learned helplessness'.[140]

During the early stages of AI development, managers and employees will have experience of the tasks being delegated to AI. But if these tasks are delegated consistently, that experience will no longer be gained and over time human knowledge may be lost. Describing this issue on a larger scale, Russell calls it the 'tragedy of the commons', 'for any individual human it may seem pointless to engage in years

[134] Espinoza (2020).

[135] Russell (2019), p 252.

[136] Ibid., p 130.

[137] Ibid., p 130.

[138] 'Tesla Crash Investigation Yields 9 NTSB Safety Recommendations', https://www.ntsb.gov/news/press-releases/Pages/NR20200225.aspx (accessed 8 April 2020).

[139] Libert, Beck and Bonchek (2017).

[140] Lindebaum, Vesa and den Hond (2019), p 24.

of arduous learning to acquire knowledge and skills that machines already have, but if everyone thinks that way, the human race will, collectively, lose its autonomy'.[141]

On a smaller scale this can be applied to boards. Before long it may only be AI that fully understands the issues at hand, the decisions, and their processes. In those circumstances, it will be hard for humans to fix problems when they arise. This is a corporate governance concern because, in order to manage risk (one of the board's key responsibilities) companies will need to have the foresight to ensure humans remain appropriately knowledgeable to be able to step in should problems occur. Yet, at the same time this may significantly reduce the efficiency gains AI can bring.

Aside from the problems we are aware of but cannot solve, there may be more examples of situations in which AI has provided an incorrect answer that is never discovered. It is known that algorithms produce knowledge which is highly probable, yet it remains uncertain to a degree.[142] The probability of AI inaccuracy may be so low, or at least perceived as such, that a party who feels aggrieved by an outcome may be unlikely to pursue any course of action due to the seemingly insurmountable task of disproving the AI. This potentially contravenes the Guidelines' principles on diversity and fairness. Boards will need to consider the potential for this and find ways which enable people to challenge AI outcomes but without facilitating vexatious challenges. That may be a fine balance to achieve.

6.2 Illegal Acts and 'Rogue AI'

A question which may, at present, seem better placed in a science fiction novel is 'what happens if AI goes rogue?' According to Mulgan

> The twin threats of digital population explosion and morally unreliable digital beings exacerbate one another. If digital reproduction is constrained only by internalised moral norms, then a single morally unreliable digital being could very quickly dominate a law-abiding population.[143]

Mulgans's paper presents a worrying picture in which 'morally unreliable digital beings merge and start populating the world with corporate groups whose priorities are entirely divorced from any human concerns'.[144] Whilst this corporate future seems quite far-fetched, it is not such a stretch to imagine humans using AI for unethical or illegal means. Such was the case when Volkswagen used algorithms to evade emissions legislation.[145] In that case six executives were charged with fraud. Although fraud may be nothing new, this highlights the new and potentially very difficult to detect ways in which AI can be used to perpetrate illegal acts.

AI will only be able to process the data it is given and so, if management were to provide an AI board with only certain, specifically selected data, in order to

[141] Russell (2019), p 255.

[142] Mittelstadt et al. (2016), p 4.

[143] Mulgan (2019), p 912.

[144] Ibid., p 914.

[145] Miller (2020).

manipulate the outcome of its processing, AI may be unable to detect this. In contrast, human boards are trained to identify such attempts and could potentially detect it based on experience and instinct.[146] In this way, AI may provide a new opportunity for self-interested management to manipulate the company. This again highlights the need for AI expertise at the level of management oversight.

Further potential problems arise if fraud and other criminal acts are perpetuated by fully autonomous algorithmic entities, without ongoing human input, as mentioned previously in this article.[147] If this scenario materializes, we will have reached what Mulgan describes as ‘morally unreliable digital corporations’.

The latter scenarios should never come to fruition if organisations were to follow the Guidelines and their emphasis on human oversight. They could thus be said to be beyond the Guidelines’ scope and more of a civil and criminal law matter. In the case of unintentional problems, as described above, the question will be what measures businesses should implement to minimize them and whether human oversight, flawed as it may be,[148] is a sufficiently ‘reliable and reproducible safeguard’ as required by the Guidelines. Safeguards could be automated but, under the Guidelines, would in the last instance have to be subject to human involvement.

7 Privacy and Data Governance

The use of data is an essential part of AI and this places increased risk on privacy of personal data. Not only does privacy need to be protected, but the usage of data also needs careful thought and access to individuals’ data needs to comply with strict protocols.[149] The Guidelines call for ‘adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy’.[150]

Much of the devil of this issue will be in the detail, and in this respect, beyond the broader considerations of corporate governance. However, boards will need to ensure the appropriate systems and processes are in place. The legal and practical landscape of data collection and usage is undoubtedly going to evolve as more and more data is captured and processed. At present, compliance with the EU’s General Data Protection Regulation is necessary but this will no doubt develop, expand and be replaced.[151] According to Armour et al., the use of AI will necessitate new oversight challenges in data governance that will ultimately fall to be reviewed by the

[146] Enriques and Zetzsche (2020), p 82.

[147] See Sect. 3.

[148] See Sect. 3.3.

[149] High-Level Expert Group on Artificial Intelligence (2019), p 17.

[150] Ibid., p 17.

[151] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [2016] OJ L 119/1 (GDPR).

board.[152] But as most boards are majority independent, they are unlikely to have the time or expertise necessary at present and, as such, a new board Technology Committee is a logical next step.[153]

The requirement for privacy is an aspect of the Guidelines that demonstrates the inherent conflicts in what we want AI to do. On the one hand we expect AI to increase transparency enormously. On the other hand, transparency can only be increased if we have access to the data on which the decisions are being made.[154] This is another example of a necessary balance or trade off when we use and regulate AI. To the extent this issue is not decided by regulation, such a high-level decision should be something considered by management and the board.

Concerns have been raised about the use of individuals' personal data to make decisions about them, and their ability to question that. Some have suggested the GDPR allow individuals a right to an explanation where a decision using their data goes against them. This appears to be corroborated in the Guidelines' principle on diversity and fairness, which requires stakeholder participation. Conversely, Wachter et al. doubt that EU data protection rules go as far, although they do not rule out future changes along these lines as a result of further legislation or judicial interpretation.[155]

Given the possibility of a right to an explanation, companies may need to grapple with how they would explain their AI-rendered decisions. Part of this problem will be understanding the algorithms that produce those decisions. Davis notes that this is not necessarily a certainty as 'the way a computer thinks' may not be susceptible to human understanding, and that a computer may organize information in a manner that humans cannot understand.[156] Providing explanations is likely beyond the remit of corporate boards but they will need to understand the requirement and be prepared to respond to relevant requests. This reinforces the need for new specific AI based roles within companies or the involvement of external expertise.[157]

8 Conclusion

As is true of many other aspects of society, AI's permeation of the corporate sphere seems inevitable. The impact of AI-influenced corporate governance has the potential to be very positive. AI's ability to quickly process large amounts of data, far beyond the capability of humans may bring many efficiencies and improvements. This can enhance the quality of a broad array of business decisions from strategy to human resources. However, the transformation is potentially radical. There are many unknowns about the impact of the change, including various

[152] Armour and Eidenmueller (2020), p 115.

[153] Ibid., p 21.

[154] Abbott (2020), p 250

[155] Wachter, Mittelstadt and Floridi (2017).

[156] Davis (2018), p 183.

[157] See Sect. 2

threats ranging from harmful impacts on third parties, discriminatory practices, data and privacy breaches, to fraudulent practices and 'rogue AI'.

The EU's Guidelines for Trustworthy AI seek to tackle some of these unknowns and mitigate potential dangers. If implemented by business, they will have substantial impacts on corporate governance. This will be amplified should the Guidelines (or some aspects therein) become integrated in legislative instruments. Fundamental questions at the intersection of ethics and law are considered by the Guidelines, but the choice has been made to only address the former without (much) reference to the latter. This makes their practical application challenging for business. Further, while the Guidelines promote many positive corporate governance principles—including a stakeholder-oriented ('human-centric') corporate purpose and diversity, non-discrimination, and fairness—it is clear that their general nature leaves many questions and concerns unanswered.

Some of these concerns are procedural, relating to the ways in which issues with AI can be dealt with, who should be in control of AI within a business entity (and how), and who is to be held accountable for its actions. Insofar as the Guidelines address these questions, they suggest that fundamental aspects of AI are a matter for the board and senior management. The Guidelines also propagate the principle of human oversight and intervention and are disinclined to allow fully autonomous AI. However, human involvement will require considerable expertise in AI and technology in order to be effective, and it is not clear to what extent it will be feasible or even desirable. In any event, AI is set to bring considerable changes to the boardroom and management. For human managers, business experience may in the future become less relevant than technical expertise, and judgment and discretion may become more valuable than hard knowledge and professional skills.

Concerns relating to ensuring trustworthiness in the corporate use of AI necessitate consideration of complex questions that are not sufficiently dealt with in the Guidelines. These include: on what principle should AI make value judgements; what does fairness mean in practice; and, in a world where tasks are delegated to machines, how will individuals still learn those tasks and thereby maintain an ability to control AI, as required by the Guidelines. These issues essentially relate to the incremental impact of humans becoming more removed from decision-making processes as well as the threat of perpetuating biases that are already prevalent in society.

There are also concerns about the effectiveness of the Guidelines in preventing or mitigating situations where AI produces unintended results, fails or malfunctions, or is used for improper means. How this can and should be dealt with from a corporate governance perspective is not clear. In many respects these concerns are the hardest to account for as they are likely to be the least predictable.

The Guidelines apply to boards, to developers, and to users in many disparate ways. The approach is principles-based and with few hard lines. Left to their own devices, corporations may use AI to further the interests of their shareholders above those of the wider stakeholders, despite what is promulgated in the Guidelines. The Guidelines' soft character and (at least partial) incompatibility with applicable corporate laws, is a consequence of dealing in complex unknowns. To improve the effectiveness and applicability of the Guidelines, a bottom-up approach, as advised by Mittelstadt et al., may be preferable to the top-down approach currently taken.[158]

Finally, questions remain as to how the Guidelines will translate into practice. Mittelstadt notes that 'norms and requirements can rarely be logically deduced directly from principles without accounting for specific elements of the technology, application, context of use, or relevant local norms'.[159] It is debatable whether the question of 'how will this requirement work in practice' will even have an acceptable answer given the complexity and open-endedness of the subject matter. It is a heavy burden to lay at the door of corporations and their boards to work this out. Understanding and debating what we want AI to be is just part of the process, the next steps need to focus on how to achieve this.[160]

Overall, we suggest that business will benefit from being able to rely on ethical guidelines for AI, but more specificity is needed relating to and harmonization with company law rules and governance principles. It also remains to be seen what, if any, 'harder' legislative instruments on AI will emerge in the near future, and whether they will provide more granular guidance for corporations based on the Guidelines' ethical framework. Until then, the Guidelines are at least a useful, albeit an incomplete and imperfect, starting point for directing businesses and their leaders towards establishing trustworthy AI.

References

Abbott R (2020) The reasonable robot. Cambridge University Press, Cambridge

Adams RB, Funk P (2012) Beyond the glass ceiling: does gender matter? Manag Sci 58(2):219–235

Armour J, Eidenmueller H (2020) Self-driving corporations? Harv Bus Law Rev 10:88–116

Bainbridge SM (1993) In defense of the shareholder wealth maximization norm: a reply to Professor Green New Directions in Corporate Law. Wash Lee Law Rev 50(4):1423–1448

Ball P (2019) How do machines think? New Statesman. https://www.newstatesman.com/science-tech/technology/2019/12/how-do-machines-think. Accessed 10 Apr 2020

Bell JM, Hartmann D (2007) Diversity in everyday discourse: the cultural ambiguities and consequences of 'happy talk'. Am Soc Rev 72(6):895–914

[158] Mittelstadt et al. (2016).

[159] Mittelstadt (2019), p 6.

[160] See also Smuha (2020), arguing that we should move from a human rights-based approach to governance to start taking measures for its concrete implementation.

Bertrand M, Mullainathan S (2003) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. National Bureau of Economic Research

Brown JS, Duguid P (2000) The social life of information—chapter one: limits to information. First Monday. https://doi.org/10.5210/fm.v5i4.738. Accessed 21 Nov 2019

Bruner CM (2009) Power and purpose in the Anglo–American coporation. Va J Int Law 50:579–654

Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. Proc Mach Learn Res 81:77–91

Burridge N (2017) Artificial intelligence gets a seat in the boardroom. Nikkei Asian Review, 10 May 2017. https://asia.nikkei.com/Business/Companies/Artificial-intelligence-gets-a-seat-in-the-boardroom. Accessed 22 Aug 2021

Business Roundtable (2019) Statement on the Purpose of a Corporation. August 2019. https://opportunity.businessroundtable.org/ourcommitment/. Accessed 7 Jan 2020

Casey B (2016) Amoral machines, or: how roboticists can learn to stop worrying and love the law online essay. Northwest Univ Law Rev 111(5):1347–1366

Choudhury B, Petrin M (2019) Corporate duties to the public. Cambridge University Press, Cambridge

Chowdhury R, Mulani N (2018) Auditing algorithms for bias. Harvard Business Review, 24 October 2018. https://hbr.org/2018/10/auditing-algorithms-for-bias. Accessed 28 Feb 2020

Collins J (2001) Good to great. https://www.jimcollins.com/article_topics/articles/good-to-great.html. Accessed 23 Nov 2019

Datta A, Datta A, Makagon J, Mulligan DK, Tschantz MC (2018) Discrimination in online advertising a multidisciplinary inquiry. Proc Mach Learn Res 81:20–34

Davenport TH, Ronanki R (2018) Artificial intelligence for the real world. Harvard Business Review January-February 2018:108-116. https://hbr.org/2018/01/artificial-intelligence-for-the-real-world. Accessed 22 Aug 2021

Davis JP (2018) Law without mind: AI, ethics, and jurisprudence. Calif Western Law Rev 55(1):165–220

Eidenmüller HGM (2017) The rise of robots and the law of humans. Oxford Legal Studies Research Paper No 27/201. https://ssrn.com/abstract=2941001. Accessed 21 Nov 2019

Enriques L, Zetzsche DA (2020) Corporate technologies and the tech nirvana fallacy. Hastings Law J 72:55–98

Espinoza J (2020) IBM and Microsoft sign Vatican pledge for ethical AI. Financial Times, 27 February 2020. https://www.ft.com/content/5dc6edcc-5981-11ea-a528-dd0f971febbc. Accessed 8 Apr 2020

European Commission (2017) Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions on the mid-term review on the implementation of the digital single market strategy—a connected digital single market for all. COM (2017) 228 final (10 May 2017)

European Commission (2018a) EU Member States sign up to cooperate on artificial intelligence (10 April 2018). https://digital-strategy.ec.europa.eu/en/news/eu-member-states-sign-cooperate-artificial-intelligence. Accessed 22 Aug 2021

European Commission (2018b) Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, Artificial intelligence for Europe. COM (2018) 237 final (25 April 2018)

European Commission (2018c) Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions—Coordinated plan on artificial intelligence. COM (2018) 795 final (7 December 2018)

European Commission (2019a) Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee and the Committee of the Regions, Building trust in human-centric artificial intelligence. COM(2019) 168 final (8 April 2019)

European Commission (2019b) Europe in May 2019: Preparing for a more united, stronger and more democratic Union in an increasingly uncertain world (Contribution to the informal EU27 leaders' meeting in Sibiu (Romania) on 9 May 2019 (9 May 2019)

European Commission (2020a) White paper on artifical intelligence—a European approach to excellence and trust. COM (2020) 65 final (19 February 2020)

European Commission (2020b) Report on the safety and liability implications of artificial intelligence, the internet of things and robotics. COM (2020) 64 final (19 February 2020)

Floridi L (2016) Should we be afraid of AI? Aeon essays. Aeon. https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible. Accessed 23 Jan 2020

Foot P (1967) The problem of abortion and the doctrine of double effect. Oxf Rev 5:5–15

Francoeur C, Labelle R, Sinclair-Desgagne B (2008) Gender diversity in corporate governance and top management. J Bus Ethics 81(1):83–95

Greenfield K (2001) Ultra vires lives—a stakeholder analysis of corporate illegality (with notes on how corporate law could reinforce international law norms). Va Law Rev 87(7):1279–1380

Harford T (2019) Algorithms judge us so know their rules. Financial Times, 22 November 2019. https://www.ft.com/content/e155f91a-0b86-11ea-b2d6-9bf4d1957a67. Accessed 23 Nov 2019

Hickman E (2014) Boardroom gender diversity: a behavioural economics analysis. J Corp Law Stud 14(2):385–418

Hickman E (2020) Diversity, merit and power in the FTSE100 C-suite. UCL doctoral thesis

High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed 18 Oct 2019

Hill A (2019) Amazon offers cautionary tale of AI-assisted hiring. Financial Times, 24 January 2019. https://www.ft.com/content/5039715c-14f9-11e9-a168-d45595ad076d. Accessed 23 Nov 2019

Hymas C (2019) AI used for first time in job interviews in UK to find best applicants. https://www.telegraph.co.uk/news/2019/09/27/ai-facial-recognition-used-first-time-job-interviews-uk-find/. Accessed 23 Nov 2019

Keay A (2010) Stakeholder theory in corporate law: has it got what it takes. Richmond J Global Law Bus 9:249–300

Kolbjørnsrud V, Amico R, Thomas RJ (2016) The promise of artificial intelligence: redefining management in the workforce of the future. https://www.researchgate.net/publication/306039533_The_promise_of_artificial_intelligence_Redefining_management_in_the_workforce_of_the_future. Accessed 22 Aug 2021

Koulu R (2020) Human control over automation: EU policy and AI ethics. Eur J Legal Stud 2020(1):9–46

Lawrence N (2017) Decision making and diversity. Inverseprobability.Com (blog). https://inverseprobability.com/2017/11/15/decision-making#fn9. Accessed 8 Apr 2020

Lee MK (2018) Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. Big Data Soc 2018:1–16

Libert B, Beck M, Bonchek M (2017) AI in the boardroom: the next realm of corporate governance. MIT Sloan Management Review. https://sloanreview.mit.edu/article/ai-in-the-boardroom-the-next-realm-of-corporate-governance/. Accessed 16 Aug 2021

Lindebaum D, Vesa M, den Hond F (2019) Insights from 'The Machine Stops' to better understand rational assumptions in algorithmic decision-making and its implications for organizations. Acad Manag Rev. https://doi.org/10.5465/amr.2018.0181. Accessed 19 Sept 2019

LoPucki LM (2018) Algorithmic entities. Wash Univ Law Rev 95:887–953

Machold S, Huse M, Hansen K, Brogi M (2013) Getting women on to corporate boards: a snowball starting in Norway. Edward Elgar Publishing, Cheltenham

Metzinger T (2019) Ethics washing made in Europe. Tagesspiegel, 4 April 2019. https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html. Accessed 22 Aug 2021

Miller J (2020) Germany charges Six Volkswagen executives over Dieselgate. Financial Times, 14 January 2020. https://www.ft.com/content/7493759c-36c3-11ea-a6d3-9a26f8c3cba4. Accessed 23 Jan 2020

Mittelstadt B (2019) AI ethics—too principled to fail? [Cs], June. http://arxiv.org/abs/1906.06668. Accessed 23 Jan 2020

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):2053951716679679. https://doi.org/10.1177/2053951716679679.Accessed16July2019

Molavi Vasse'i R (2019) The Ethical Guidelines for Trustworthy AI—a procrastination of effective law enforcement. Comp Law Rev Int 5(2019):129–136

Moore M, Petrin M (2017) Corporate governance: law, regulation and theory. Palgrave McMillan, London

Möslein F (2018) Robots in the boardroom: artificial intelligence and corporate law. In: Barfield W, Pagallo U (eds) Research handbook on the law of artificial intelligence. Edward Elgar, Cheltenham, pp 649–670

Mulgan T (2019) Corporate agency and possible futures. J Bus Ethics 154(4):901–916. https://doi.org/10.1007/s10551-018-3887-1

Mulholland P (2020) Tech retraining employed to narrow stem gender gap. Financial Times, 10 March 2020. https://www.ft.com/content/dbb917ec-3e0a-11ea-b84f-a62c46f39bc2. Accessed 8 Apr 2020

Nilsson NJ (2010) The quest for artificial intelligence. Cambridge University Press, Cambridge

Paquette D (2018) In boardroom, robot gets a seat at the table. Washington Post.Com, 26 January 2018, sec. wonkblog. http://global.factiva.com/redir/default.aspx?P=sa&an=WPCOM00020180126ee1q005br&cat=a&ep=ASE. Accessed 21 Nov 2019

Parker J, The Parker Review Committee (2016) A report into the ethnic diversity of UK boards. https://assets.ey.com/content/dam/ey-sites/ey-com/en_uk/news/2020/02/ey-parker-review-2017-report-final.pdf. Accessed 17 Aug 2021

Petrin M (2019) Corporate management in the age of AI. Columbia Bus Law Rev 2019(3):965–1030

Russell S (2019) Human compatible. Penguin Random House, London

Science History Institute (2016) Alexander Fleming. https://www.sciencehistory.org/historical-profile/alexander-fleming. Accessed 15 Feb 2020

Smuha NA (2019) The EU approach to ethics guidelines for trustworthy artificial intelligence. Comp Law Rev Int 20:97–106

Smuha NA (2020) Beyond a human rights-based approach to AI governance: promise, pitfalls, plea. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3543112. Accessed 22 Aug 2021

Stout LA (2012) The shareholder value myth: how putting shareholders first harms investors, corporations, and the public. Berrett-Koehler Publishers, San Francisco

Tech Talent Charter (2019) What, why and who is the Tech Talent Charter? https://www.techtalentcharter.co.uk/about-the-tech-talent-charter. Accessed 15 Feb 2020

The Company Law Review Steering Group (1999) Modern company law for a competitive economy. The Strategic Framework. https://webarchive.nationalarchives.gov.uk/20121101191957/http://www.bis.gov.uk/files/file23279.pdf. Accessed 17 Aug 2021

Thomas D (2020) EY diversity head: UK still has ‘some way to go’. Financial Times, 29 January 2020. https://www.ft.com/content/274a8e34-339d-11ea-a329-0bcf87a328f2. Accessed 28 Feb 2020

Thornhill J (2020) Trusting AI too much can turn out to be fatal. Financial Times, 2 March 2020. https://www.ft.com/content/0e086832-5c5c-11ea-8033-fa40a0d65a98. Accessed 8 Apr 2020

Tieto Press Release (2016) Tieto the first Nordic company to appoint artificial intelligence to the leadership team of the new data-driven businesses unit. https://www.bloomberg.com/press-releases/2016-10-17/tieto-the-first-nordic-company-to-appoint-artificial-intelligence-to-the-leadership-team-of-the-new-data-driven-businesses-unit. Accessed 22 Aug 2021

Trafton A (2020) Artificial intelligence yields new antibiotic. MIT News. http://news.mit.edu/2020/artificial-intelligence-identifies-new-antibiotic-0220. Accessed 7 Apr 2020

Veale M (2020) A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence. Eur J Risk Regul 2020:1–10

Vinnicombe S, Atewologun D, Battista V (2019) The female FTSE board report: moving beyond the numbers. Cranfield University

von der Leyen U (2019) A Union that strives for more: my agenda for Europe. Political Guidelines for the Next European Commission 2019-2024. https://ec.europa.eu/info/sites/default/files/political-guidelines-next-commission_en_0.pdf. Accessed 26 Jul 2021

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Priv Law 7(2):76–99

Wiener N (1960) Some moral and technical consequences of automation. Science 131(3410):1355–1358

Williamson OE (1985) The economic institutions of capitalism: firms, markets, relational contracting. Collier Macmillan, London

Yap M, Konrad AM (2009) Gender and racial differentials in promotions: is there a sticky floor, a mid-level bottleneck, or a glass ceiling? J Ind Relat 64:593–619

Yong E (2018) A popular algorithm is no better at predicting crimes than random people. The Atlantic, 17 January 2018. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/. Accessed 23 Nov 2019

Authors and Affiliations

Eleanore Hickman[1] · Martin Petrin[2,3]

Martin Petrin
mpetrin@uwo.ca

[1] University of Bristol, Bristol, UK

[2] Western University, London, ON, Canada

[3] University College London, London, UK