Chapter

Human Factor on Artificial Intelligence: The Way to Ethical and Responsible Economic Growth

Helena García-Mieres, Ignacio Pedrosa and Jimena Pascual

Abstract

Artificial intelligence (AI) is substantially changing the world of business. The growth of AI and its impact on business and society are explored in this chapter, including dilemmas and emerging challenges, highlighting the existing gap in the adoption of ethical principles. The elements of human-centered AI (HCAI), such as privacy, explainability, equity, and fairness, are presented, analyzing its relevance in business. and how they need to be addressed to guarantee profitable investments in AI technologies. The aim of this book chapter is to present the essential knowledge needed by business about how to behave more ethically in AI development and deployment. In this regard, we discuss how to increase user confidence and usage of AI devices by presenting a best-practice guide to underscore biases and ensure fairness in AI-based products, exemplified in the financial and healthcare sector. Improving people's understanding of how AI models operate and having a clear HCAI strategy that evaluates negative potential biases of AI systems will increase user trust, spread, and usage of AI devices, thus ensuring the full acceptance of AI in society, thus promoting human, economic, and institutional growth.

Keywords: AI ethics, trustworthy AI, AI biases, human-centered AI, innovation, artificial intelligence, business

1. Introduction

The era of artificial intelligence (AI) has already entered our daily lives and is changing how most businesses, industries, societies, and humanity will operate. AI enables the creation of new products and services, the improvement of existing ones, and the optimization of processes and operations. AI may also have the potential to enhance the productivity and efficiency of human workers by augmenting their skills and capabilities or automating repetitive and mundane tasks.

AI is a branch of computer science, which focuses on the development of algorithms and systems that can perform tasks that typically require human intelligence, including cognitive processes such as learning, perception, reasoning, and decision-making. Technologies based on AI are designed to interpret large and complex

volumes of data to learn from them using mathematical algorithms and use them to perform predictive analyses based on real-world experience. These predictive analyses, thus, can generate new insights and discoveries, leading to innovation and competitive advantage for business sectors.

As AI systems are moving from theoretical mathematics and advanced hardware to everyday aspects of life, it becomes of interest and value to the modern economy and technology. In fact, AI technology actually underlies much of our daily routines. Once we switch on our devices, we immediately connect to AI functionalities such as face ID, online banking, digital voice assistants, or driving aids.

This is also the case on the organizational side, transforming the way they operate with business applications that are familiar to everyone [1], such as:

- Customer service: AI can help businesses provide faster and more personalized responses to customer queries, complaints, and feedback using chatbots, voice assistants, and sentiment analysis.

- Personalized marketing: AI can help businesses optimize their marketing campaigns, target their audience, generate leads, and personalize their content and offers through data analytics, natural language generation, and recommendation systems.

- Operations: AI can help businesses improve their efficiency, productivity, and quality of their processes and products using automation, robotics, computer vision, and predictive maintenance.

- Financial services: AI can improve business financial performance, helping with risk management and fraud detection through machine learning, natural language processing, and anomaly detection.

- Human resources: AI can assist companies in attracting, retaining, and developing their talent-by-talent acquisition, employee engagement, performance evaluation, learning, and development.

- Personalized preventive care: Factors in social determinants of health in AI models can contribute to delivering personalized and preventative healthcare.

AI is becoming a key driver of sustainable economic growth in technology and ensures the competitive advantage of any business sector. According to the most recent report by McKinsey [2], AI could potentially add 16% by 2030 to current global economic output. This report also shows that AI adoption has increased to approximately 50% of the companies surveyed, compared to 20% in 2017. The level of investment of companies in AI has increased along with its growing adoption. Importantly, 63% of McKinsey's respondents mentioned that they expect the investment of their organizations to increase over the next few years. This is not just a trend, but its benefits are already being felt as organizations that have adopted AI report realizing meaningful cost decreases and revenue increases [3].

The current top use cases for AI are optimization of service operations, marketing and sales, product and service development, and strategy and corporate finance. The biggest reported earning effects for business are found in marketing and sales, product and service development, and strategy and corporate finance, while the highest

cost benefits of AI are in supply chain management [2]. This proves that AI can be a real catalyst for the transformation of the financial sector [1].

2. The gap between the technical and ethical adoption of AI in business

Despite these advances, most companies are still in the initial stages of digitalization and adoption of AI, thus having the potential to create enormous value for consumers, business, and society, but this position also implies many profound challenges and risks [4].

Important AI-related risks are concerns in cybersecurity, regulatory compliance, personal privacy, explainability of AI models, organizational reputation, or equity and fairness [2, 5]. An area of consistent concern is the extent to which organizations are actively involved in risk mitigation to enhance digital trust. While AI adoption and investment have increased, there have been no substantial increases in reported mitigation of any AI-related risks compared to 2019 (**Figure 1**), the first year of McKinsey's survey [2]. This situation is worrying as research investment in AI technologies is also substantial. For instance, the EU invested €10 billion into AI through its framework programs between 2014 and 2020, representing 13.4% of all available funding. However, only 30.3% of funding calls related to AI mention trustworthiness, privacy, or ethics [6].

This flattening in risk mitigation strategies by companies contrasts with the increasing and severe incidence of ethical misuses of AI tools. This is compounded by the growing trend of AI ethics and security researcher layoffs within the tech giants [7].

According to the most recent report of Stanford University's AI index [3], the number of AI-related controversies has increased up to 26 times in 10 years. Similarly,

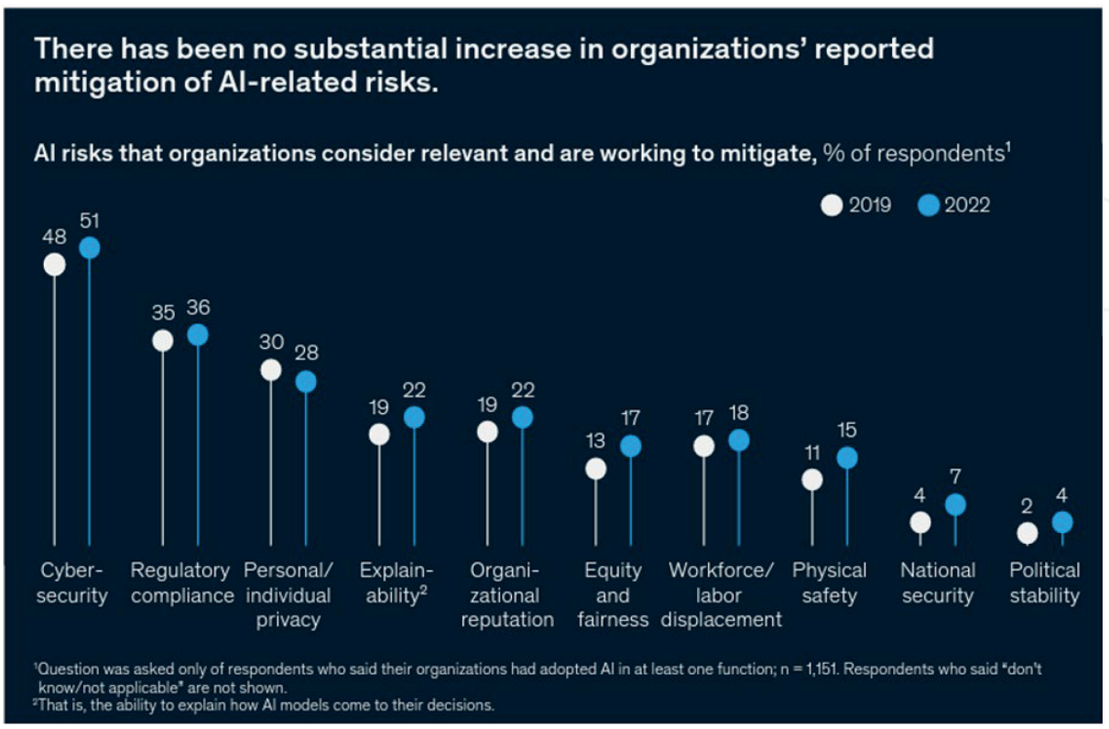

Figure 1.
Survey of risk mitigation strategies in AI technologies in United States companies. Source. McKinsey report [2].

academic developments and scientific publications in ethical aspects of AI tools and methods to mitigate these risks have had an exponential increase since 2012. The most researched topics have been strategies for better management of privacy concerns, explainability, equity, and regulatory processes [3, 8]. This evidences a gap between the scientific development of ethics in AI and the practical adoption of these developments by companies.

These issues are exacerbated by the lack of up-to-date regulation despite recent progress at the EU level. Several initiatives are being developed to enhance the adoption of the ethical aspects of AI. In the European Union, the European Commission's High-Level Expert Group on AI (HLEG). In 2019, the ethics guidelines for trustworthy artificial intelligence were defined [9], putting forward seven key requirements that AI systems should meet to be deemed trustworthy.

Additionally, the rapid expansion of AI is already outpacing the development and deployment of legal and regulatory frameworks. In this sense, in May 2021, the EU Commission became the first worldwide governmental body to present the so-called "first legal framework on AI," aimed at regulating the use of AI (AI Act, European Commission) [10]. Since data is the essential foundation of AI, other regulations have progressively joined in such as the Data Act, the data governance Act, or the Digital Services Act. On top of this, there are recent national initiatives as proposals to regulate AI (e.g., UK pro-innovation approach to AI regulation) [11].

The challenge now is for the industry to harness that power to face current challenges and create sustainable and efficient solutions. As companies are adopting and deploying AI tools and technologies more routinely, the complicated ethical challenges mentioned are expected to continue to rise and negatively impact companies and consumers.

The research community suggests that technology companies, admissions officers, hiring managers, banking executives, and other decision-makers adopt a human-centered approach to AI products rather than a purely technological one. In fact, several firms that famously adopted purely technological processes have found it necessary to reintroduce humans to provide control in AI products [5]. This implies that 1) more agile strategies for translating scientific development into operational lines of business are needed to offer more ethical and sustainable AI-based products and 2) As AI becomes more prevalent in productive processes and across labor market demands, countries will need to make extra efforts to provide effective training opportunities for individuals. This will enable them to benefit from the advantages that this innovative technology can offer [12].

This book chapter, thus, presents the essential knowledge needed by business about how to behave more ethically in AI development and deployment. We will begin by exposing the concept and principles of human-centered AI (HCAI). We then continue by examining the possible challenges of AI for economic growth if HCAI principles are not enough considered in the business strategy, illustrated by two case studies in the financial and healthcare sectors. Third, we will discuss how to increase user trust and usage of AI devices by proposing a good practice guide around the principles of HCAI to address bias.

3. Human-Centered AI: concept and principles

The widespread adoption of AI by industry and many aspects of human life is impacting individuals and generating consequences in society in ways that are not

yet well understood. This includes both potential benefits and risks. In this context, is where HCAI was born. HCAI is an emerging discipline that aims to design and develop AI systems that are aligned with human values, needs, and preferences. HCAI seeks to preserve human control in a way that ensures AI meets our needs while also operating transparently, delivering equitable outcomes, and respecting privacy [5]. The overall purpose of HCAI is to create AI solutions that are trustworthy, ethical, and beneficial to people, businesses, and society.

To achieve these purposes, HCAI systems should be built following several principles [9]:

a. *Ensure responsible design and development of AI systems.* Responsible AI is an umbrella term for several efforts to develop legal, ethical, and moral perspectives into the design and usage of AI products. Responsible AI is a concept often used to show how an organization is managing potential negative consequences of their AI products. Therefore, responsible AI is the general framework for designing AI products that engender trust by employees, business, customers, and society. The subcomponents of responsible AI establish that explainability, transparency, fairness, accountability, and ethics should be considered in the design phase of any AI tool.

b. *Preserve human privacy*. Most AI applications are based on human data and their interactions; therefore, the very nature of AI implies a threat to privacy throughout its lifecycle such as initial data collection, the metadata extracted, the inferences drawn by AI models, or the need to safeguard stored data. Therefore, AI systems should consider privacy (and data protection) by design and by default principle, including strong cybersecurity measures.

c. *Ensure transparency and explainability*. As AI becomes more complex and advanced, it becomes difficult for humans to understand how an algorithm provides an outcome. The calculation process is often referred to as a "black box,", which is impossible to interpret. To open this "black box," AI models should be built under the principles of transparency and explainability. Transparency in AI systems refers to the ability to understand how an AI model arrived at a specific decision or output, thus making the decision-making process of an AI model visible and comprehensible to final users.

   Similarly, explainable artificial intelligence (XAI) is the set of processes and methods that allow to reach this transparency in AI models. XAI enables human users (both AI experts and nontechnical experts) to understand and trust the output produced by machine learning algorithms. XAI, thus, becomes essential for organizations to build trust and confidence in AI. As AI becomes more complex, XAI helps developers ensure that the system works as intended, meets regulatory standards, and enables stakeholders to challenge or alter AI decisions, thus boosting the embracement of the tools.

d. *Comply with regulatory requirements*. AI regulation is still in its infancy worldwide. For example, the AI Act mentioned previously is a proposed European law on AI. Its main provisions include a risk-based approach for AI systems, requirements for transparency and accountability, limitations on certain uses of AI, and mandatory data and recordkeeping obligations for high-risk AI systems. The regulation also proposes significant fines for noncompliance and establishes a

European Artificial Intelligence Board to oversee its implementation. The specific risks of AI are categorized into four different levels: unacceptable risk, high risk, limited risk, and minimal risk [10]. Although this draught regulation needs further refinement and adoption by countries, it stands as an initial legal instrument to guide the compliance of AI products with legislation in which rights in the digital age should also be considered [12].

e. *Ensure equity and fairness.* These principles aim to ensure that AI systems do not perpetuate bias or discrimination toward certain groups of people. Equity in AI means that individuals should be treated based on their unique circumstances and needs. This means that AI systems should consider factors such as race, gender, age, or socioeconomic status and adjust their outputs accordingly. On the other hand, fairness in AI means that AI systems should not systematically favor or discriminate against particular groups of people or against some of their characteristics such as their age or their gender. AI systems should be designed to minimize the impact of bias and discrimination, thus ensuring all individuals are treated fairly and equally.

f. *Include the participation and feedback of relevant stakeholders.* AI systems should be co-design, developed, and assessed in collaboration with pertinent stakeholders incorporating their input and feedback. It is crucial to engage a diverse range of stakeholders during the creation and implementation of AI systems to ensure that a broad spectrum of perspectives and requirements are considered. It is, therefore, necessary for public administrations that guarantee fundamental rights to be involved in the tenders for AI-based solutions and to set up social impact assessment procedures.

Moreover, preliminary research should be conducted to investigate the needs, feasibility, and acceptability of the intended end users. This research can provide insights into how individuals interact with and trust AI systems, enabling the design of a final AI product that aligns with user expectations.

Businesses should include elements of HCAI in their AI agenda, thus contributing to enhancing customer satisfaction, employee engagement, social responsibility, and competitive advantage. By adopting HCAI in their practice, businesses can foster trust, collaboration, and innovation with their stakeholders. Similarly, integrating HCAI elements since the designing of the AI products may also avoid unexpected costs after market entry, or even prevent millionaire investments in AI products that may not align with human needs and/or preferences.

4. Possible consequences that resulted from the failure to include HCAI principles: The case of smart speakers

By the end of 2022, many journal articles have been published about the challenges that the big company Amazon has faced with its Alexa voice assistant in the business market. According to several articles published in specialized media such as Business Insider, Ars Technica, or The Guardian [13–15], Amazon has reportedly lost around $10 billion in its efforts to gain a foothold in the enterprise market, which has been largely dominated by competitors such as Microsoft and Google. Amazon has faced challenges in convincing businesses and expected home users to adopt Alexa

for the workplace and online shopping tasks and has struggled with reliability and security issues.

The first controversy faced by Amazon and Google and their voice assistants has been their previous evidenced violation of privacy policies. A study carried out by researchers from the University of Clemson [16] showed that Amazon and Google voice assistants had high-level privacy issues such as broken and incorrect privacy policy URLs, duplicate privacy policy links, lack of privacy policies in skills where they were needed, or inconsistencies and errors in the content of the policies. The results of this study reached the community and resulted in discussions about the security of smart speakers, spreading a lack of confidence in the product among end users, and negatively impacting the reputation of the companies.

However, one of the greatest challenges faced by Amazon in relation to its voice assistant is related to the expected profitability of the product. The primary objective of this US-based multinational corporation was not only merely to generate revenue from the sale of the devices but also to capitalize on their usage by customers such as through shopping on Amazon. The aim was to establish these devices as a novel interface for consumers, comparable to the adoption of smartphones for online purchases. However, the actual usage of Echo speakers and the Alexa assistant has not conformed to this profile in most cases. Although smart speakers have exceeded projections made a few years ago and achieved widespread adoption in United States households, the primary use of these devices by most of their final users has been for routine activities such as information retrieval and music playback [13–15]. Although these services are valuable, they hinder the company's original objective of monetization.

In sum, Amazon developed an AI product aimed at creating a new human and potentially profitable need (i.e., do online shopping using smart voice assistants) with a specific strategy for its scaling up (for instance, selling the Echo speakers nearly at their cost of production) and reaching a big success selling the Echo speakers, but the end users are not using Alexa for the initial purpose expected by the company. The overall financial losses incurred by the company in its efforts to break into the business market have been significant.

It is not expected to oversimplify a complex case such as Amazon with its voice assistant as the market scenario involves a multifaceted interplay of factors beyond the scope of this chapter. However, from a perspective rooted in HCAI, we posit that a more conscientious integration of HCAI principles in the design of voice assistants could lead to greater market success for companies interested in producing smart speakers.

One possibility for dealing with privacy concerns considering the principles of HCAI, as proposed by Liao et al. [16], might be that companies could add a solution to inform users about the data collection capabilities of a voice app. They propose a built-in intent that scans for data collection capabilities and notifies users about it. The intent could be invoked when the app is enabled and provide a brief privacy notice. Additionally, the intent could advise users to look at a detailed policy provided by the developers. The authors also proposed to extend this approach to automatically generate privacy policies for voice-apps in the future.

Regarding the profitability of smart speakers, one plausible scenario is that companies could formulate an effective strategy if the involvement and feedback of relevant stakeholders, including prospective end users in domestic settings, are involved. By conducting thorough and extensive research on the requirements, feasibility, and acceptability of smart speakers among the target audience, companies could achieve a properly aligned product with user expectations and make a more beneficial investment.

This case exemplifies that a safeguard economic growth in AI-based products inherently needs to include the principles of HCAI in the design of the devices. AI devices should be designed and used with ethics, transparency, and trust in their pipeline to ensure that end expected users adopt them. This idea is also supported by several studies showing that the adoption of AI services is positively associated with the ability of customers to understand the product, its perceived usefulness, knowledge or awareness of AI technology, positive attitude, and trust in AI [17, 18].

The case of smart speakers serves as an illustration of the potential negative consequences that can arise from a failure to apply the principles of HCAI in companies' AI products. Such consequences may include loss of profitability, reputation damage, and negative social impact. It is important to note that these risks are not limited to large companies or specific sectors but are increasingly relevant in other domains where the adoption of AI is rapidly expanding such as finance or healthcare technology. Therefore, it is essential to prioritize the principles of HCAI to avoid such negative consequences.

With regards to general ethical principles and the application of HCAI in the fields of finance and health, there exist some interesting guidelines. Regarding the first one, the "Code of Conduct for the Ethical Use of AI in Canadian Financial Services" is a valued soft-law source in Canada. The document is a set of principles, developed in consultation with various Canadian financial service organizations. The objective of this document is to promote the ethical use of AI in financial institutions by offering practical guidance to prevent ethical implications in the daily usage of AI. This code represents a milestone toward practical and industry-specific ethical principles. For the case of healthcare companies, an interesting starting point could be the World Health Organization [19] guide: "*Ethics and governance of artificial intelligence for health*." The report identifies the ethical challenges and risks associated with the use of artificial intelligence in healthcare. It presents six consensus principles that should be followed to ensure that AI works for the public's worldwide benefit.

Next, with the aim to provide a more operational and concise insight in HCAI, we focus on the principle of ensuring equity and fairness in AI systems because of the potential impact that both issues can make in terms of social justice. To this end, the next section addresses the challenge of detecting biases in AI-based financial and healthcare services, as well as providing a set of best practices aimed at promoting and achieving equity and fairness in AI.

5. The challenge of bias and fairness in AI-based services: the cases of financial and health sectors

As big data involves vast amounts of data reflecting society, AI-driven models could just perpetuate biases that already exist in society and are reflected in such databases. Bias can lead to unfair outcomes for certain groups of people such as women and minorities.

To be effective and avoid ethical pitfalls, companies need to ensure that AI is not programmed with biases that could lead to ethically charged decision-making or cause AI to malfunction in some way. In a report of NTT data that surveyed eight sectors of business in the United States [20], about one-fifth of respondents who used AI models in their companies say that they offered them suggestions that reflected bias against a particular vulnerable group. Organizations cannot risk wasting on technology investments gone wrong, therefore they must pivot their organizations to focus on ethics and other pressing issues.

The independent high-level expert group on artificial intelligence [9] defines bias as "*an inclination of prejudice toward or against a person, object, or position.*" Bias drives the value of most risk prediction models (wanted bias), but it can also be detrimental to it. In certain cases, bias can result in unwanted discriminatory and/or unfair outcomes, labeled in this document as unfair bias.

Bias in AI models can arise from all the steps of the machine learning algorithm pipeline. These include bias in training data, algorithmic bias, bias in logic-based AI, bias arising from self-learning and adaptation, or bias arising from personalization. Bias can be caused by several factors such as underrepresented populations, erroneous data, outlier data, and biased human decision-making in data collection or labeling. Bias can also arise from limited contexts in which a system is used, resulting in a lack of opportunity to generalize it to other contexts.

Addressing bias in AI services should result in fairness in their implementation. This means ensuring that the effect a model has on individuals and groups is free of unfair bias, discrimination, and stigmatization. Popular notions of fairness include demographic parity (also called statistical parity, e.g., women and men have the same chance to get a loan), equalized odds (women and men who all meet certain other requirements have the same chance to get a loan), or the well-calibrated ness (among those who got a loan, women and men are equally represented as in any random sample).

Sources of bias are likely to be present in the data utilized for training predictive models in financial and healthcare services. Therefore, it is crucial to identify these sources of bias and implement effective measures to mitigate their impact. In this regard, we produce a guide of best practices aimed at minimizing undesired bias and ensuring the reliability and validity of the models. To frame this in context, we set out two domains concerning the population as a whole that can be highly useful for illustrative purposes.

Firstly, AI is becoming an essential tool for financial services such as fraud detection, risk prevention, credit scoring, loan approval, or insurance underwriting. Moreover, given the nature of data in the banks, AI has a significant role in processing data to predict the future of the economy and banking industry.

Secondly, regarding healthcare, AI systems are used for population and individual segmentation, personalized screening, diagnosis, massive data treatment, and personalized interventions. Technology derived from wearable devices can be applied for disease management and monitoring. In addition, artificial intelligence has the potential to revolutionize biomedical research and drug development, including immunological therapies for rare diseases and less frequent types of cancer. Furthermore, AI is effectively applied in clinical management such as prediction of demand and intelligent use of healthcare resources, optimization of operating rooms, or intelligent scheduling. The adoption of AI in the healthcare field presents a series of HCAI particularities. On the one hand, it demands an important level of data privacy and algorithmic robustness that exceeds those in other domains. On the other hand, it entails the need to establish clear accountability while not discouraging medical professionals from utilizing these tools [19].

5.1 Illustrative case in financial sector

Thinking from the big picture of ethical AI in financial services, a model for automating credit decisions, the results of which affect human lives and are publicly visible, should be free of unwanted bias and meet requirements for model transparency.

In some legislations, credit customers even have the explicit right to request an explanation of the reasons behind credit decisions pertaining to themselves, whether the actual decision was positive or negative [21].

A typical example would be if we suppose that the training data for a credit pricing model show that men have higher salaries on average than women, which is actually a societal fact. Any bank should be aware that this gender bias can arise in models even though gender itself is not an explanatory variable in the model. It may be that a higher loan rejection rate for women would be statistically justifiable from the training data (and might comply with an equalized odd definition), but a bank should reject that model for ethical (or reputational) reasons. Another known example of bias has been racial discrimination in mortgage approvals [22]. Minority applicants were found to have a significantly lower chance of receiving algorithmic approval to receive a mortgage from race-blind government automated underwriting systems compared to Caucasians.

5.2 Illustrative case in healthcare sector

In the case of unwanted bias and fairness, while AI is improving diagnosis, treatments, and lowering the costs to discover and develop drugs, it has also introduced biases detrimental to demographic minorities in automated decision-making. These biases are partly due to the disproportionate overrepresentation of Caucasian and higher income patients in electronic health records datasets [22]. If training data consists predominantly of medical records from white males, an AI clinical decision support system may perform poorly or be less accurate when making diagnoses or treatment recommendations for women or people from racial minorities. This is because the AI model has not been trained on enough heterogeneous data to account for the diverse ways diseases occur in different demographic groups. Such bias can have profound consequences on patient outcomes and exacerbate existing healthcare disparities.

In view of the risk of undesired bias in AI-based products for the financial and health sectors, we propose a good practice guide on bias and fairness in AI. These principles are defined in such a way that they can be applied to all business sectors, as a one-size-fits-all guide.

6. Good practice guide for addressing Bias in AI-based solutions

6.1 Ensure that the data sets used to train and evaluate the prediction models are representative and high quality

Ensuring a representative dataset and high-quality data is crucial to avoid social biases in the predictions made by an AI product. For example, if a dataset is composed mainly of one demographic group, the AI product may not accurately capture the needs or behaviors of other groups, leading to biased predictions. Additionally, if data quality is low, the AI product may generate inaccurate predictions, leading to social bias. Therefore, it is essential to take measures to ensure that the data set used to train an AI model is diverse, balanced, and high quality to minimize the risk of social biases in the predictions made by the AI product.

Several analytical procedures can be included to achieve a representative high-quality data set. Such as including samples from all relevant groups and populations, use statistical methods to identify and remove any bias in the dataset before training the AI model or use multiple sources of data to train the AI model to reduce the

likelihood of bias from an only source. Many actionable guidelines for those actively involved in the development, evaluation, and implementation of AI-based prediction models are available (e.g., [23]) and adapted to specific sectors. For illustration, de Hond et al. [24] provide a general guide for healthcare. When it comes to the financial sector, several guidelines can be used depending on the specific case to face. For instance, Zampino et al. [25] offer an example of a guide for creditworthiness.

6.2 Include testing and mitigating bias as a routine in the development and deployment of algorithms

To ensure that an algorithm is unbiased and fair, it is important to thoroughly evaluate its inputs and outputs against known biases and check for fairness in decision-making. This can be achieved by analyzing the data and examining the algorithm's decision-making process to identify any potential biases. Once identified, strategies can be implemented to mitigate these biases. These strategies may include adjusting the algorithm parameters, adding more data, or changing the data collection process.

Open-source tools such as the IBM AI Fairness 360 [26] or the Holistic AI library can be useful in this process. These tools include a comprehensive set of metrics that can be used to assess biases in both datasets and models. Additionally, it provides explanations for these metrics and algorithms that can be used to mitigate bias in datasets and models. Using these tools and techniques, data scientists can help guarantee that their algorithms are unbiased and fair, promoting greater equity and inclusion in the decision-making process.

6.3 Build multidisciplinary and collaborative teams in charge of the AI models

A multidisciplinary team, in which close cooperation of IT staff with experts of relevant sectors such as financial or health occurs, could be one way to adjust the tradeoff between the predictability of the model and explainability and respond to the legal and regulatory requirements for auditability and transparency. There may be a need to build bridges between disciplines that currently work in silos, such as deep learning and symbolic approaches, with the latter involving rules created through human intervention [27].

This approach is formally known in the literature as developing collaborative machine learning. Collaborative machine learning allows the development of machine learning models that involve multiple stakeholders working together to create, test, and deploy the model. The stakeholders can include data scientists, domain experts, and/or end users. A team composed of individuals with diverse backgrounds enables varied perspectives in analyzing data, thereby reducing the likelihood of overlooking biases in datasets and the methods used for developing predictive models. This underscores the relevance of diversity in mitigating potential biases in the whole pipeline of the predictive models. Similarly, if end users are surveyed to provide feedback on the model, they can identify any issues or biases that may not have been apparent to the developers and help them make decisions about the model.

6.4 Monitor and review

Continuously monitor and review the algorithm's performance to ensure it remains unbiased over time and update it as needed. The goal is to identify potential

issues or biases that may arise and correct them before they cause harm or inaccuracies. Continuous testing of AI models is indispensable to identify and correct model drifts. Model drift occurs when the model's performance starts to deteriorate over time because of changes in the data it is processing or other external factors. Capturing and correcting model drifts early allows to maintain the algorithm's accuracy and avoid unintended consequences.

The frequency of review and validation may need to be defined depending on the complexity of the model, the pace of new data generation, and the relevance of the decisions made by such a model. For instance, an algorithm that is used to make high-stakes decisions on people may require more frequent reviews than one used for less critical tasks.

6.5 Ensure AI transparency and governance

Be transparent about the algorithm and its limitations. Explain how it works and provide clear explanations about how decisions are made. Transparency in AI transparency requires the availability of model and system documentation that is understandable and trustworthy. This allows a consumer of the model to determine if it is appropriate for their situation. AI governance allows companies to specify and enforce policies describing how an AI model or service should be constructed and deployed. This can prevent undesirable situations such as a model training with unapproved datasets, models having biases, or models having unexpected performance variations. Several methodologies have been developed to assure accountability and transparency in the development of AI models and systems. Among them, IBM FactSheets [28] and the model card framework [29] are widely known. These methodologies share a common approach of using document templates that mimic 'nutrition labels,' which contain basic information on the purpose of the model, data selection and preparation, algorithm selection and adjustment, and testing for accuracy, bias, or privacy risks. Templates can be customized to suit a diverse range of stakeholders, including risk officers, end users, affected subjects, or bank officers, among others. Additionally, there are instructional materials, guidelines, and case studies available for financial AI products and medical decision systems.

6.6 Ensure traceability of our models

Traceability and quality management are important aspects of business performance in the industry. Requirements for businesses to report in writing operational details and design characteristics of the models used were already in place before the advent of AI. Documentation of the logic behind the algorithm, to the extent feasible, is being used by some regulators to ensure that the outcomes produced by the model are explainable, traceable, and repeatable [27, 28].

Traceability in AI is considered a key requirement for trustworthy AI outputs, related to the need to maintain a complete account of the provenance of data, processes, and artifacts involved in the production of an AI model. A comprehensive approach to traceability would require on one hand a repeatable execution of the computational steps, but also to capture aspects as metadata that may not be explicit or evident in the digital artifacts. To ensure traceability, a documentation mechanism must be incorporated to the best possible standard. A review of existing methods and tools to do this documentation can be further consulted by Mora-Cantallops et al. [30].

6.7 Implement practices of algorithmic auditing by internal or external parts

The area of 'algorithmic auditing' is emerging and becoming an important aspect in the adoption of AI products in companies from all sectors as it institutionalizes accountability and robust due diligence in technology. Companies may incorporate formal ethics reviews and model validation exercises in addition to internal and external algorithmic auditing to ensure that the adoption of AI is transparent and has gone through screening and formal validation processes. The broader outcome of an auditing process is to improve confidence or ensure trust of the underlying system and then to capture that in some certification process. After analyzing the system and implementing mitigation strategies, the auditing process assesses whether the system conforms to regulatory, governance, and ethical standards. Providing assurance needs to be understood through different dimensions, and steps need to be taken so that the algorithm can be shown to be trustworthy [29, 31].

6.8 Introduce mechanisms to ensure that humans verify the final decision of the model

AI applications are designed and used by humans, and humans decide the degree of autonomy assigned to an AI application, whether that be human-controlled, semiautonomous, or fully autonomous. Human overseers are supposed to increase the accuracy and safety of AI systems, uphold human values in automated decision-making, and build trust in the technology. Therefore, delegation of autonomy comes with great responsibility, and organizations must remember that technologies such as AI are not a complete substitute for humans [32].

Appropriate emphasis could be placed on human oversight in decision-making when it comes to higher-value use cases (e.g., lending decisions), which significantly affect the population [27]. The final decision about the model to be used, which one needs to be reviewed, which models should be discontinued, and, importantly, the action to be done with the decision of the algorithm (e.g., approval of a mortgage or medical diagnosis), should always be made by a human being. Therefore, AI models should be used as a decision support tool rather than being left to act on their own. This ensures that the responsibility resides with the

Good Practice Guide for Addressing Bias in AI-based Solutions
Ensure representative and high quality in the datasets used to train and evaluate the prediction models
Include testing and mitigating for bias as a routine in the development and deployment of the algorithms.
Build multidisciplinary and collaborative teams in charge of the AI models
Monitor and review
Ensure AI transparency and governance
Ensure traceability of our models
Implement practices of algorithmic auditing by internal or external parts
Introduce mechanisms to ensure that the final decision of the model is verified by humans.

Table 1.
Good practice guide for addressing Bias in AI-based solutions. Source: Prepared by CTIC.

respective human decision-maker but it is also an important control for drift in self-learning models [21]. An overview of the 8 items presented in this section is shown on **Table 1**.

7. Conclusions

The full impact that AI technology may have in business and society is yet to be determined, but several technical, strategic, and stakeholders cooperation questions need to be addressed. In this sense, hot topics to address in the upcoming years include developing new concepts for testing and validation, defining dataset requirements, and ensuring their quality for AI, as well as embedding ethics guidelines to guarantee trustworthy AI.

Summarizing all the above, to achieve equity and fairness in AI, developers must ensure that their algorithms are trained on unbiased data and that they are transparent and explainable. They must also continuously monitor and audit their AI systems to identify and address any potential sources of bias or discrimination. Additionally, it is important to involve diverse stakeholders in the development and deployment of AI systems to ensure that a variety of perspectives and needs are considered.

This best practices guide is proposed to underscore biases and ensure fairness in AI-based products, but many of its principles extend to and positively impact other dimensions of HCAI such as transparency, accountability, or explainability.

Improving people's understanding of how AI models operate and having a clear HCAI strategy that gauges negative potential biases of AI systems will increase user trust, spread, and usage of AI devices, thus ensuring full acceptance of AI in society and the economy.

Acknowledgements

This research was granted by IBERUS (CER-20211003 - Spanish Ministry of Science and Innovation). IBERUS CERVERA Network of Excellence Project (CER-20211003) is funded by the Ministry of Science and Innovation through the Centre for the Development of Industrial Technology (CDTI), under the General State Budget 2021 and the Plan for Recovery, Transformation, and Resilience (Plan de Recuperación, Transformación y Resiliencia) of the Spanish Government.

Conflict of interest

The authors declare no conflict of interest.

Human Factor on Artificial Intelligence: The Way to Ethical and Responsible Economic Growth
DOI: http://dx.doi.org/10.5772/intechopen.111915

IntechOpen

Author details

Helena García-Mieres, Ignacio Pedrosa* and Jimena Pascual
Health, Active Ageing and Wellbeing Unit and Human Factor Unit of CTIC Technological Centre/W3C LATAM, Technological and Scientific Park of Gijón, Spain

*Address all correspondence to: ignacio.pedrosa@fundacionctic.org

IntechOpen

References

[1] BDVA Task Force. Big Data and AI for the Financial Sector: Challenges and Opportunities [white paper]. 2022. Available from: https://www.bdva.eu/sites/default/files/BDVA%20TF7.SG10%20AI%20and%20Big%20Data%20for%20the%20Financial%20Sector%20-%20Whitepaper%20V3_0.pdf. p. 7

[2] McKinsey Company. The State of AI in 2022—And a Half Decade. 2022. Available from: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#/

[3] Maslej N, Fattorini L, Brynjolfsson E, Etchemendy J, Ligett K, Lyons T, et al. The AI index 2023 annual report. In: Ai Index Steering Committee. Stanford, CA: Institute for Human-Centered Amnesty International, Stanford University; 2023

[4] Maslak OI, Maslak MV, Grishko NY, Hlazunova OO, Pererva PG, Yakovenko YY. Artificial Intelligence as a Key Driver of Business Operations Transformation in the Conditions of the Digital Economy IEEE International Conference on Modern Electrical and Energy Systems (MEES). Vol. 20212021. pp. 1-5. DOI: 10.1109/MEES52427.2021.9598744

[5] Ozmen Garibay O, Winslow B, Andolina S, Antona M, Bodenschatz A, Coursaris C, et al. Six human-centered artificial intelligence grand challenges. International Journal of Human Computer Interaction. 2023;**39**(3):391-437. DOI: 10.1080/10447318.2022.2153320

[6] Galdon-Clavell G, Bertelli V, Yalaz Ozen E, Lorente-Martinez T. How Public Money is Shaping the Future Direction of AI: an Analysis of the EU's Investment in AI Development. Ethicas Report commissioned by the European AI & Society Fund. 2023. Available from: https://europeanaifund.org/newspublications/report-how-public-money-is-shaping-the-future-direction-of-ai-an-analysis-of-the-eus-investment-in-ai-development/

[7] Duffy C. Cable News Network Business. 'It's An Especially Bad Time': Tech Layoffs are Hitting Ethics and Safety Teams. 2023. Available from: https://edition.cnn.com/2023/04/06/tech/tech-layoffs-platform-safety/index.html

[8] Zhu J. AI ethics with Chinese characteristics? Concerns and preferred solutions in Chinese academia. AI & Society. 2022:1-14. DOI: 10.1007/s00146-022-01578-w. PMID 36276898002E

[9] Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission; 2019. Available from: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

[10] European Commission. Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts; 2021. Available from: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[11] UK Department for Science, Innovation and Technology. A Pro-Innovation Approach to AI Regulation [white paper]. 2023. Available from: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

[12] OECD. Designing Active Labour Market Policies for the Recovery. OECD Publishing; 2021. Available from: https://www.oecd.org/coronavirus/policy-responses/designing-active-labour-market-policies-for-the-recovery-79c833cf/

[13] Amadeo R. Amazon Alexa Is a Colossal Failure, on Pace to Lose $10 Billion this Year. Ars Technica; 2022. Available from: https://arstechnica.com/gadgets/2022/11/amazon-alexa-is-a-colossal-failure-on-pace-to-lose-10-billion-this-year/

[14] Burgess M. Alexa, How did Amazon's Voice Assistant Rack up a $10bn Loss? The Guardian. Available from: https://www.theguardian.com/commentisfree/2022/nov/26/alexa-how-did-amazons-voice-assistant-rack-up-a-10bn-loss [Accessed: 26 November 2022]

[15] Swearingen J. Amazon's Alexa Unit Reportedly Suffered $10 Billion in Losses, Highlighting the company's Struggle to Break into Business Software. Business insider; 2022. Available from: https://www.businessinsider.com/amazon-alexa-business-failure-10-bn-losses-2022-11

[16] Liao S, Wilson C, Cheng L, Hu H, Deng H. Measuring the effectiveness of privacy policies for voice assistant applications. In: Annual Computer Security Applications Conference (ACSAC '20). New York: Association for Computing Machinery; 2020. pp. 856-869. DOI: 10.1145/3427228.3427250

[17] Noreen U, Shafique A, Ahmed Z, Ashfaq M. Banking 4.0: Artificial intelligence (AI) in banking industry & consumer's perspective. Sustainability. 2023;**15**(4):3682. DOI: 10.3390/su15043682

[18] Yang R, Wibowo S. User trust in artificial intelligence: A comprehensive conceptual framework. Electron Markets. 2022;**32**(4):2053-2077. DOI: 10.1007/s12525-022-00592-6

[19] World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva: World Health Organization. Licence: CC BY-NC-SA 3.0 IGO; 2021. Available from: https://www.who.int/publications/i/item/9789240029200

[20] NTT. DATA and Oxford Economics. AI Accelerated: The Great Shift to Artificial Intelligence and Automation. 2020. Available from: https://mx.nttdata.com/es/engage/ai-study-ai-accelerated

[21] Fritz-Morgenthal S, Hein B, Papenbrock J. Financial risk management and explainable, trustworthy, responsible AI. Frontiers in Artificial Intelligence. 2022;**5**:779799. DOI: 10.3389/frai.2022.779799, PMID 35295866

[22] Neil B, Aurel H, Daniel R. How Much Does Racial bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions. Washington, D.C.; 2022. Available at SSRN from: https://ssrn.com/abstract=3887663 or DOI: 10.2139/ssrn.3887663

[23] Gudivada V, Apon A, Ding J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. International Journal of Advances in Software Engineering. 2017;**10**(1):1-20

[24] de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: A scoping review. npj Digital Medicine. 2022;**5**(1):2. DOI: 10.1038/s41746-021-00549-7 PMID 35013569

[25] Zampino F, Longo A, Zappatore M. A user-centered approach to create realistic datasets for AI. Case study: Creditworthiness in the banking sector. In: CEUR Workshop Proceedings. 2022. Available from: https://ceur-ws.org/Vol-3340/paper41.pdf

[26] Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, et al. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development. 2019;**63**(4/5):1-5. DOI: 10.1147/JRD.2019.2942287

[27] OECD. Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers. 2021. Available from: https://www.oecd.org/finance/artificialintelligencemachinelearningbigdatainfinance.htm

[28] Richards JT, Piorkowski D, Hind M, Houde S, Mojsilovic A, Varshney KR. A human-centered methodology for creating AI FactSheets. IEEE Data Engineering Bulletin. 2021;**44**(4):47-58

[29] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. pp. 220-229. DOI: 10.1145/3287560.3287596

[30] Mora-Cantallops M, Sánchez-Alonso S, García-Barriocanal E, Sicilia M-A. Traceability for trustworthy AI: A review of models and tools. Big Data and Cognitive Computing. 2021;**5**(2):20. DOI: 10.3390/bdcc5020020

[31] Koshiyama A, Kazim E, Treleaven P, Rai P, Szpruch L, Pavey G, et al. Towards algorithm auditing a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. SSRN Journal. 2021. DOI: 10.2139/ssrn.3778998

[32] Kelley S, Levin SY, Saunders D. A Code of Conduct for the Ethical Use of Artificial Intelligence in Canadian Financial Services. 2018. Available from: https://www.stephaniekelleyresearch.com/_files/ugd/614bca_5112c4738e8642b58d46cb4676c46729.pdf