Ethics for AI in Business

Saïd Business School, University of Oxford:
Felipe Thomaz (felipe.thomaz@sbs.ox.ac.uk), Natalia Efremova, Francesca Mazzi,
Greg Clark, Ewan MacDonald, Rhonda Hadi, Jason Bell and Andrew Stephen

Report commissioned by the ICC Research Foundation.

Ethics in AI for Business

Introduction

At this point in time, it would be a severe understatement to simply point out that the recent rapid development and adoption of artificial intelligence (AI) processes has been transformative in all facets of business operations. For a variety of reasons, from the digital transformation of businesses and the accompanying greater data availability to the acceleration in computing power; the variety of methodologies and approaches that are broadly classified as AI have drastically shifted the way in which firms ingest information, process and analyse data, and augment or automate tasks and decisions. In fact, more broadly, the advent of AI is likely to have a substantial impact on the nature of consumer behaviour, firm competition, and future successful business models (Thomaz et al 2020).

While customers and firms have benefited from these technologies, the overall value of AI has been significantly larger. Beyond purely economic, cost, and efficiency advantages, the application of these technologies has brought vast social benefits, including advances in new product development that take into account improved accessibility for individuals with disabilities, and reduced waste. AI has been applied in healthcare for both the development of new treatments and the improvement of diagnostics, and amongst others in the development of a vaccine to contrast the COVID-19 pandemic. In natural sciences, it has been used to analyse the pattern of natural fires or improving irrigation in farming. It has even been used in conservation for the identification of individual animals and their habitats.

However, we also have seen a number of unintended consequences and negative outcomes, with numerous reports of innovative attempts by companies and governments falling short in spectacular and public fashion. Among these are chatbots who learned hate speech online, human resources processes that unwittingly enshrined sexual discrimination in code, or school testing scores being assigned on basis of socioeconomic condition rather than student achievement. Invariably, this has led to new policy debates on how to curb negative outcomes and the concept of ethical use of AI.

Given the tremendous incentives for businesses to develop and implement AI systems, the private sector is a critical player in these policy debates. And consequentially, any delay or inability for the private sector in adopting ethical use of AI further motivates stronger regulation of the technology and its uses, which in turn would have an expected dampening effect on further innovation and advancement in AI. Unsurprisingly, many companies are tackling this issue head on; especially those whose business models are more strongly intertwined with the development and use of AI processes. These firms have taken the lead in ethical AI thought and leadership, setting up governance structures, statements of purpose, and guidance materials for their own organizations and peers. Similarly, IGOs and governments have added their own voices and guidance on the ethical application of AI as it pertains to the citizens or a combined vision of what an Ethical AI society should aspire to be.

However, while firms, governments, intergovernmental and non-governmental organizations have developed and published materials on ethical AI principles, a series of significant issues have become apparent. In a review of 84 existing ethical AI guidelines, Jobin and colleagues (2019) identify that there remains a significant divergence in understanding and interpretation of the principles being discussed. Simply, there is no common language and meaning across many of these documents, no agreement on the issues, domains, or actors to be considered. Furthermore, this analysis highlights a sense of confusion due to unintegrated sectorial guidelines and strong geographic bias, with the bulk of guidelines representing North American and European perspectives, rather than a global framework. And lastly, but perhaps most importantly, none of the existing work presents a viable candidate application strategy on how to achieve and manage AI ethics.

Here we aim to resolve these issues. This effort motivated the partnership between the International Chamber of Commerce (ICC) and the University of Oxford's Saïd Business School to collaborate and provide the voice of business and actively promote the ethical use of AI. We start from the foundational belief that ethical design of AI is intrinsically good for business, and that its application is ultimately good for broader society. Businesses of all sizes, sectors and locations must be empowered to seize the many new opportunities that AI technologies offer. This report provides the necessary guidance to help avoid the pitfalls associated with these technologies and ensure the effective implementation of ethical AI for the benefit of the business, its stakeholders and the wider community. We especially recommend this report, to the attention of small and medium enterprises (SMEs) thinking of applying AI to their business or as a thought starter for companies of all sizes just setting out on their AI journey.

In order to summarize and organize the existing guidance on Ethical AI in a practical framework we first collected any items that were published in academic journals on the topic, but also considered business statements (e.g. Microsoft Responsible AI Principles, Google AI Principles and Responsibilities), governmental guidance documents (e.g. G20 AI Principles), as well as those prepared by non-governmental and intra-governmental agencies (e.g. OECD AI Principles and AI Policy Observatory, A Framework for Ethical AI at the United Nations), and even articles found in the popular press. For those interested in specifics, the literature in AI ethics in computer science from 1954 to 2021 is described and reviewed in the Annex 1 of this document. This review identified eight key areas of concern when discussing AI ethics, from Privacy to Algorithmic fairness and beyond. These are highlighted and described more completely in Annex 2 of this document. Annex 3 summarizes the key insights from this review.

Lastly, in order to fully take advantage of the collective knowledge and the wealth of information contained in all of these documents, we turned to AI ourselves and used Natural Language Processing techniques to process, organize, and extract relational information from the documents available (the corpus). Details on text normalization, frequency analysis, vector embeddings, named entity recognition and knowledge graph generation, as well as and some results, are given in Annex 4 of this document.

But importantly, by scouring the existing knowledge base on Ethical AI and combining the analysis of this corpus with operating ethical frameworks, we were able to organize the critical components into a defensible and implementable workflow. We discuss this next.

Our AI Implementation Roadmap

1. Set of principles to be implemented throughout the lifecycle of AI
2. Definition of the principles and their hierarchy
3. Assessing ethical risks (general)
4. Practical Implementation
   - Mission Statement
   - Ethics design plan
     A. Phase and application specific ethical impact assessment
     B. Mitigation strategies
     C. Monitoring system

1. Set of principles to be implemented throughout the lifecycle of AI

There are a number of issues that become apparent when we consider the field of ethical AI and its application to business, especially when we must translate discussions into actual tangible steps and processes in our operations. First and foremost is the nature of ethical principles themselves. These are invariably intangible, grand ideas with rather fuzzy boundaries, which makes them perhaps easy to grasp in general, but difficult to pin down with any sort of precision. For example, consider the idea of "justice." We might have an intuitive sense of what being "just" might be, or we might know it when we see it, but the specifics of what makes something "just" becomes more difficult.

Relatedly, there is ample confusion among similar terms and concepts. For example, do you need to provide "justice" or "fairness"? Are they the same thing? What about Ethical AI, versus Responsible AI, or Trustworthy AI? These terms are often used interchangeably, depending on the speaker, the document, or domain. Sometimes, the same word might be used for an ethical principle and the reverse meaning in computer science applications (e.g. "transparency"). We might then be particularly worried that different individuals or organisations might read the same statement; but leave with entirely different interpretations of its meaning.

Additional confusion about enabling, creating, and maintaining ethical AI applications could be attributed to the significant, and perhaps overwhelming, number of ethical principles that are used to explain ethical considerations and design. In the aforementioned overview of ethical guidelines for AI (Jobin, Ienca and Vayena 2019), the authors' analysis of 84 guidance documents identified 1180 distinct ethical concepts, belonging to 13 ethical themes and categories. As such, the ethical designer must contend with a host of difficulties to define concepts.

Lastly, and further complicating the issue, these concepts are always presented in an unintegrated, non-relational form. Meaning that while we might have guidance with lists of concepts to consider, we do ultimately lack any frameworks with which to guide the design of ethical AI business applications and governance systems. By analogy, a firm's directory contains all of the people that make up the organization, but hardly captures the tasks, teams and structure that allows the business to operate successfully. And similarly, a directory of ethical principles can be useful, but perhaps insufficient for adaptation and inclusion into specific and diverse business contexts.

2. Definition of the principles and their hierarchy

In order to develop our framework and organization of ethical principles, we conducted a review of ten descriptive and normative ethical theories (Utilitarianism, Deontology, Virtue Ethics, Ethics of Care, Egoism, Divine Command, Natural Law, Social Contract, Theory of Justice, and Moral Relativism) in order to understand their arguments and structures. Additionally, we reviewed and took inspiration from the Belmont report, a guiding document on Protection of Human Subjects in Biomedical and Behavioural Research. This report represents a longstanding gold standard in the academic community for the design of processes where individuals will have less power, or agency in their environments. Lastly, we derived insights from our own Natural Language Processing analysis of AI ethics documents, policy discussions, guidelines, and academic publications as described earlier. Importantly, this last analysis allows us to better select the most relevant ethical principles for our context, and to begin organizing them in a coherent structure.

Figure 1 below presents the resulting framework, and more precise definitions of each component follows.

Importantly, we suggest that these ethical principles are hierarchical; meaning, that many are components and subcomponents of other larger and more complex principles. We begin at the highest level of the hierarchy with "Ethical." The second layer of the framework, then provides the means by which one would achieve this highest-order ethical behavior. Specifically, it suggests that ethics requires one to be both "Responsible" and "Accountable." More precise definitions follow below, but for now, responsibility relates to the execution of one's duties and responsibilities faithfully. Accountability; however, reflects one's ability and willingness to explain and justify their actions and decisions. Therefore, while responsibility speaks to the means by which ends are achieved, accountability is concerned with the ends themselves (intended, perceived, and actual). Ethical AI applications in business then require both a responsible (how) and accountable (what outcomes) approach.

Why might each of these components be insufficient on their own, when in fact, a number of organizations globally have adopted a "responsible AI" approach and language? Our approach looks at two domains: being accountable with respect to end results, and being responsible, with respect to the execution of your duties and overall behaviour. Conversely, simply executing duties and responsibilities faithfully, is necessary, but insufficient. Attention must be paid to the outcomes of those actions. Our framework suggests that Ethical AI in business should satisfy both conditions: responsible means, and accountable outcomes. Taken together, the two concepts aim at safeguarding not only that organisations ensure the proper functioning of the AI systems that they design, develop, operate or deploy throughout their entire lifecycle, in accordance with their roles and applicable regulatory frameworks, but also that they demonstrate this commitment through their actions, behaviour and decision-making process more broadly.

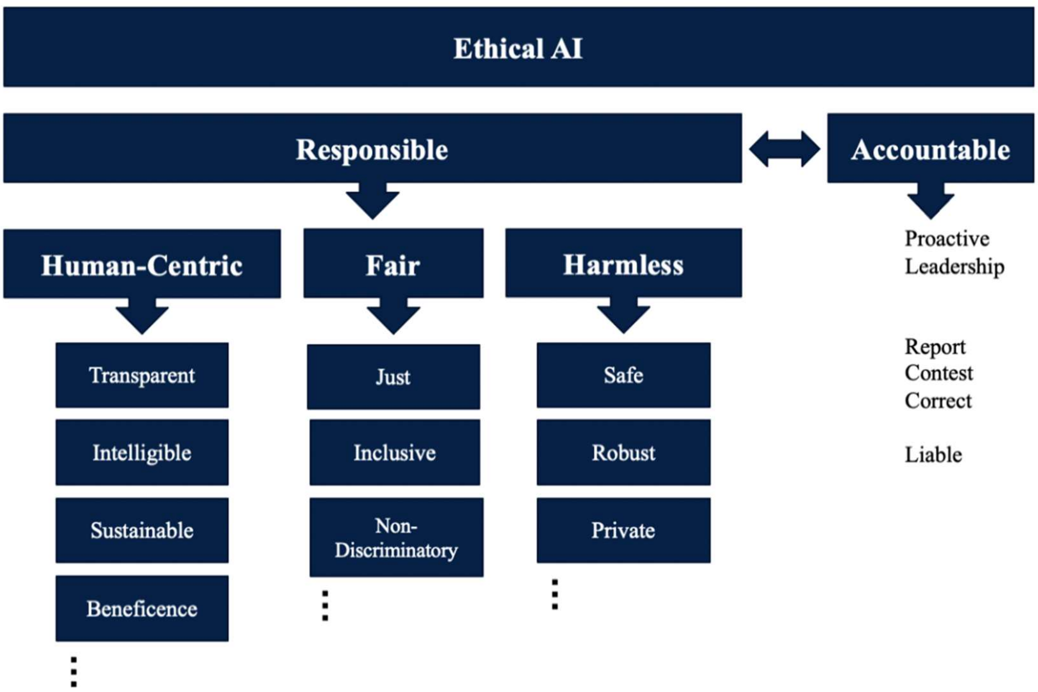
The next layer in our hierarchy of principles proves to be the most important, and one in which we'll focus most of our attention. Which principles are required in order to have a "Responsible" application? These are: "Human-Centric," "Fair" and "Harmless." You'll note that these principles are also quite vague, which suggests that each will in turn still be made up of several smaller but increasingly more tangible ideas. And precisely, the next set of principles in our hierarchy provide the mean by which one achieves these goals. For example, "Human Centric" is achieved via transparent, intelligible, and sustainable systems, but also includes to concept of beneficence. "Fair" processes are those that can be classified as just, inclusive, and non-discriminatory. In turn, "Harmless" systems are safe, robust, and private. Importantly, these lower-level characteristics are not only more easily understood and defined, but also more measurable, meaning that at this level we might be able to start introducing controls and assurances of performance within acceptable bounds. For example, we might easily be able to quantify the safety of a system based on its failure rate, as well as its robustness based on the scope of consequences given a failure. Those values, together with an assessment of the systems privacy risk, aggregate upwards toward a defensible classification of a "Harmless" AI process. If we can equally support the idea that the process is "Fair" and "Human-Centric", then we can aggregate from there and summarize the process as "Responsible." And so on until an overarching stance of it being an ethical AI application.

"Accountable" deserves special attention as it featured heavily in our NLP analysis of the body of work discussing AI ethics in business. While the concept appears prominently in legal and regulatory texts, it is featured a lot less in in business-based publications. That should not be altogether surprising, as regulators will invariably focus on the need for new regulation. That is their function after all. Our focus was to unpack the reasons for the term's relative unpopularity in business literature. It is our belief that this disconnect comes from the close association or perhaps even confusion of this concept with that strictly of liability and culpability. In that sense, to be held accountable would be a discussion on being legally liable for the consequence of AI processes and certain outcomes. And while that is certainly one component of accountability, which is made more complicated by regional legal differences and buyer/supplier of AI solution relationships; accountability is more complex and more important than this suggests.

Our framework places a number of principles and activities as components of accountability: proactive leadership, reporting, contesting, correcting, and liability. Focusing on this first component, proactive leadership, we recognize that accountability is a leadership function and not strictly reactive where the concern lies in addressing culpability when something goes wrong. Leadership, and specifically with the business and new AI development space, is tied to a forward-looking view, scenario mapping, as well as an awareness of consequences such that the firm is accountable to its various stakeholders, internal and external. In that sense, accountability is not strictly a legal concern, but rather an obligation to one's customers, employees, communities, as well as the public at large.

Lastly, we note that the framework itself and the list of items aggregating up to the main and critical concepts of "human centric," "fair," and "harmless" is open-ended. As noted earlier, with over one thousand candidate ideas, a comprehensive and exhaustive framework would be both unwieldy and perhaps unwanted. While additional ideas and ethical concepts exist, they will invariably fit as components of one of these three main general principles, or perhaps as a component of accountability itself. This organization then presents a reduction and a simplified view of the field of AI ethics for business applications, allowing us to build practices and governance structures to ensure ethical processes as well as outcomes.

Figure 1: Hierarchy of Principles



<u>Principles – definitions</u>

In this section we provide brief definitions and explanations of the principles presented in our framework:

Ethical - moral principles that govern a person's behaviour or the conducting of an activity.

Responsible - having an obligation to do something, or having control over or care for someone, as part of one's job or role.

Accountable - required or expected to justify one's intentions, actions or decisions.

Human-Centric - respects individuals' value and agency, including any necessary. information for self-determination, and care for the environment and conditions where individuals exist.

Transparent - Open to necessary scrutiny.

Intelligible - Able to be understood; comprehensible.

Sustainable - Able to operate in ecological balance, avoids depleting resources.

Beneficence – Accomplishing a good outcome, increasing the welfare of the user.

Fair - an ability to judge without reference to one's feelings or interests, treating people equally without favouritism or discrimination.

Just - deserved or appropriate treatment in the circumstances, well founded opinions and appraisals; justifiable.

Inclusive - not excluding any groups of people.

Non-Discriminatory - no prejudicial distinctions between different categories of people.

Harmless – Not able, or likely to cause harm.

Safe - operation/use does not cause harm, protected from risk.

Robust - The process or application is said to be robust if it fails "gracefully" rather than "spectacularly." Usually it means the system is not sensitive to extreme examples, outliers, and adversarial attacks, combined with fail-safe measures that allow it to fail (if it does) without causing harm.

Private - provides the ability to control access to information about the self

3. Assessing Ethical Risks

Three risk "buckets": Data, Algorithm, Business Use

(1) Data - The term "ethical AI" implies a commitment to verify and ensure that the entire workflow of AI respects ethical principles. Indeed, such term shall be seen as a label, a guarantee of adequacy that the business is willing to achieve and maintain, and that consequently contributes to build trust between users and firms that utilises AI. The required commitment shall cover all the relevant phases and all the aspects surrounding the use of AI, through a so-called "ethical impact assessment".

For the purpose of designing a meaningful impact assessment, a first macro-categorization of areas that need to be considered when assessing ethical risks includes data, algorithm and business use.

The first category, data, represents an essential component of AI that provides substance to the use of AI in terms of content. From the training of the AI to the actual deployment of it, data plays the role of nourishing the AI with the relevant information that will initially provide the AI with actual knowledge and reference, and then contribute in a determinant way to the final outcomes of its deployment.

To provide an example, the AI can learn how to differentiate apples from oranges based on provided data, which can include pictures of oranges and apples and related labels to identify it. Moreover, when the AI is used to practically differentiate apples from oranges, it will be able to do so in relation to images or pieces of information of oranges and apples that it is provided with.

As a consequence, the selection of data that the machine is provided with (so-called data set) represents a fundamental aspect of the AI functioning, but also a potential source of unethical outcomes. For example, if the AI is trained only on a certain type of apples, it will not be able to recognise other types of apples, which will result in a discrimination. If the AI is used only in relation to these two types of fruit, it should not be deployed in environments or for purposes that might include other kind of food.

Hence, the first identified macro-area of ethical risk, data, requires a substantial analysis of the selection of information provided to the AI. Such analysis shall be performed in light of the general purpose of the use of AI within the business, the environment of the actual deployment, and the ethical principles identified, and it shall concern all data provided to the AI from the training phase to the actual use.

Moreover, ethical concerns that specifically arise in relation to the category of data relate to data protection law and the right to privacy. With the proliferation of digitalised information, the legislation regarding data protection flourished in many countries in the past years. Laws and regulations on data protection shall be carefully assessed and considered not only in relation to the immediately applicable legislation, for example the legislation of the country where the firm is legally based, but also in relation to the geographical scope of operation of the business. Indeed, for example, legislation such as the General Data Protection Regulation is also applicable to businesses based outside the EU that process personal data of EU citizens.

Consequently, an adequate ethical assessment of AI requires an evaluation of the type of data that is processed by AI in order to evaluate whether it involves information that concerns individuals, whether the individuals are identifiable through the processed information, whether the information concerns sensitive aspects of individuals such as health or religion. Moreover, such assessment will need to consider the requirements of the applicable regulations, the potential risks for individuals, if any, from a privacy perspective, and the required measures to prevent risks and ensure an effective exercise of data subjects rights.

(2) Algorithms - The second category, the algorithm, represents the basic component of AI. The algorithm is, indeed, a set of instructions, and the AI is composed of algorithms that can be changed and adapted based on the intelligent learning of the AI. Hence, algorithms represent the bone structure of the AI. As such, they do represent a potential source of unethical output as well. The entirety of the lifecycle of algorithms should be monitored to verify the respect of ethical principles.

Indeed, for example, biases in the AI can result from the phase of development of the algorithm. Biases can arise due to unethical coding influenced by pre-existent biases, coming from social institutions, cultural practices, and personal attitudes, perpetuated by the programmer. Additionally, during the development phase, technical biases can be incorporated in the design of the algorithm. Moreover, during its entire lifecycle, the AI can be biased due to emergent biases, since as it was mentioned earlier the algorithms that compose the AI can be changed and adapted through the learning of AI. Hence, if the AI learns an unethical pattern, which can be caused by the real-world use or by processing biased data, it might adapt its algorithms based on such unethical information, therefore incorporating, and perpetuating such biases. Although certain AI systems, such as neural networks, contain a high number of layers and algorithms cannot be monitored closely during their deployment due to the so-called black-box effect, in such cases appropriate measures should be in place to constantly monitor the ethical deployment of such models.

Consequently, an adequate ethical impact assessment requires a close monitoring of the different phases of the algorithms, from their development to their deployment, to the extent that biases can be detected and corrected.

(3) Business Use - The third category, business use, comprises the purpose and the use of the AI that its owner intends to pursue, and the actual deployment in the real world.

The ethical principles set out in relation to AI should be in principle applicable to the business as well. In order to use the "ethical AI" label as described above, adjectives such as "non-harmful" and "fair" should not only be referable to the AI within a business, but also to the business model employing the AI *per se*. For example, an AI could be well-trained to avoid gender biases, but it will not be ethical if used to monitor employees in working conditions that violate human rights.

In this sense, the business use of AI can be a source of unethical behaviours for two main reasons. The first, immediate reason concerns the business goal, as described in the example. If the purpose of the development of AI is the creation of a lethal weapon, the outcome of the AI is foreseeably unethical.

The second reason relates to the actual deployment of the AI in the real world. The AI can be built to achieve ethical goals, but it can prove to produce unethical outputs when opened to the public if risks are not adequately prevented. As an example, discriminations or harms can derive from the interaction of users with the AI or between themselves through the AI. A mismatch of values and intentions between the function that the AI is intended to perform and its actual deployment in the real world can result in unethical output and consequently lead to liability issues.

Hence, in order to prevent liability issues and to properly address potential risks deriving from the business use of the AI, a careful consideration of the societal consequences of the use of AI is fundamental. An adequate ethical impact assessment requires an evaluation of the objectives and goals to be achieved with AI, and of the risks associated with the real-world use of the AI to achieve such a purpose.

4. Practical Implementation

While potentially interesting as a thought exercise alone, we propose that the framework for ethical principles presented in this work is much more impactful in practice if used for the design of an ethical plan for the firm's AI ambitions and for each of its specific applications. Specifically, this approach would require two steps:

First, the creation of a general, external and internal facing statement or certification of the firm's stance as an adherent of AI ethics and guaranteeing the integrity of its practices. In terms of scope and tone, this document might be seen an analogous to a firm's purpose or mission/vision statement. In fact, ethical AI policies should be a logical extension of, and entirely consistent with, corporate purpose. This initial statement should serve as blanket declaration that the firm's AI technology is consistent as such, and ethically applied. For the purpose of enhancing transparency, the statement should include the hierarchy of values of the firm, that are used to address trade-offs.

As a second step, we create a set of supporting documents that provide validity to the general statement. Each is an internal application-specific ethical design plan, where the designer carefully audits the AI process, raising questions and flagging risks as well as concerns around data, algorithm, and business use. Once specific ethical concerns and risks are identified, specific controls as well as management and mitigation strategies for each risk can be introduced. As part of the supporting documents, the firm should keep record of decisions concerning trade-offs, that should reflect the values indicated in the mission statement, to facilitate external auditing and enhance transparency. The exercise provides the opportunity to carefully consider the sources of ethical tension, depending on the application and context, and to create a manageable control system that increases the likelihood of human centric, fair, and harmless outcomes; therefore, responsible and more easily accountable processes.

For the firm, the first document is a “flight plan” by analogy. A guiding clearance that all is in order in terms of trajectory and the underlying engineering. Keeping to the same analogy, the second set of documents provides us our “flight checklist”, or a more precise listing of important safety checks and considerations pre-flight along with processes for in-flight monitoring. While the first document, "the flight plan", incorporates the mission statement as a stable long-term trajectory, the second set of documents is dynamic in nature as it evolves with the use of AI. By monitoring the use of AI, new tensions or risks can be identified, and the ethical "checklist" can be updated accordingly with related safeguards and preventive measures. In this way, the second set of documents works in loop cycles that ensures constant monitoring and updating of the ethical plan.

As a governance process, this approach provides a number of benefits. Broad adoption would legitimize the practice and severely minimize the threat of regulatory force, who would otherwise need to step in to consolidate practices and protect end users. A well-studied, thought out, and implemented ethical design should minimize the risk exposure of the firm by identifying, minimizing, and mitigating issues before they have an opportunity to become unpleasant and large issues. And lastly, as AI becomes as increasingly strong competitive battleground for businesses, the evidence of good ethical practices should serve as a competitive advantage over firms that choose to forego such measures.

As such, the firm would have the following document workflow:

A) A general AI Ethics Statement for the whole Organization (external facing)

B) An application specific control document for each use-case (internal), identifying

   a. Ethical risk identification and impact assessment

   b. Mitigation strategies for identified concerns

   c. Monitoring system for the application that account for identified risks.

Any changes to operations would then trigger a quick check on whether new risks are introduced, or potential impact is altered, against the noted components of the framework given in Figure 1. Anything that threatens the execution of Ethical AI then requires some work on mitigation strategies and monitoring updates in order to continue operating in a way that benefits the firm and its innovation, but also keeping with its responsibility and accountability to stakeholders. And altogether, this should move AI applications towards a much more familiar management practice around precise controls and margins of acceptable operation, rather than vague aspirational requirements.

References

1. Thomaz, F., Salge, C., Karahanna, E. et al. Learning from the Dark Web: leveraging conversational agents in the era of hyper-privacy to enhance marketing. J. of the Acad. Mark. Sci. 48, 43–63 (2020). https://doi.org/10.1007/s11747-019-00704
2. Jobin, Anna, Ienca, Marcello, and Vayena, Effy (2019), 'The global landscape of AI ethics guidelines', Nature Machine Intelligence, 1 (9), 389-99.

Annex 1 – Literature Review

This section describes literature on ethics, fairness and transparency in AI and related fields, such as machine learning (ML) and data science (DS). In order to provide a full description of literature on this topic(s), we include the following items: (1) an overview of the volumes of the publications over the period of time, covered with electronic publishing; (2) a comparison of the literature from computer science versus other fields; and (3) a classification and a brief description of the main topics in computer science literature.

In order to scope a body of literature on all three topics of interest, we conducted a scoping review of the existing corpus of documents on AI ethics. This included a search for literature containing principles and guidelines for ethical AI from academic sources. We analyse non-academic sources of literature with knowledge graphs in the second part of this research (see NLP model for knowledge representation with graphs).

A scoping review is a method aimed at synthesizing and mapping the existing literature (Arksey, H. and O'Malley 2005) that is considered particularly suitable for complex or heterogeneous areas of research (Pham et al. 2014).

First, we performed a keyword-based search on Scopus using the following keywords: 'AI ethics', 'AI fairness' and 'AI transparency'. First search returned 858 results; second search returned 156 and the third search returned 316 results (Table A1.1). The overlapping results constituted only 7 studies, which indicates that these keywords are used in different scientific domains. The results below confirm this assumption.

Additionally, we analysed unique source titles (journal or conference titles). In the first group, we have found 478 unique source titles, in the second group we have found 112 sources and in the third one 249.

Table A1.1 Number of papers per keyword group: AI ethics, AI fairness and AI transparency. We analyse the total number of publications and additionally we distinguish between computer science literature versus other types of publications.

Keywords	Results	Source titles	Years	CS papers	non-CS papers
AI ethics	858	478	1954 - 2021	280	578
AI fairness	156	112	1999 - 2020	117	39
AI transparency	316	246	1983 - 2020	176	140

For every category, we have checked whether the source was from computer science literature or not: we performed binary labelling on the retrieved data. We marked as technical literature only the sources, where the scope of the publisher was explicitly technical, and marked as non-technical literature everything else[1]. The results showed significant growth of publications for all fields over time. To demonstrate the growth in the number of publications per year, we have displayed the following data below:

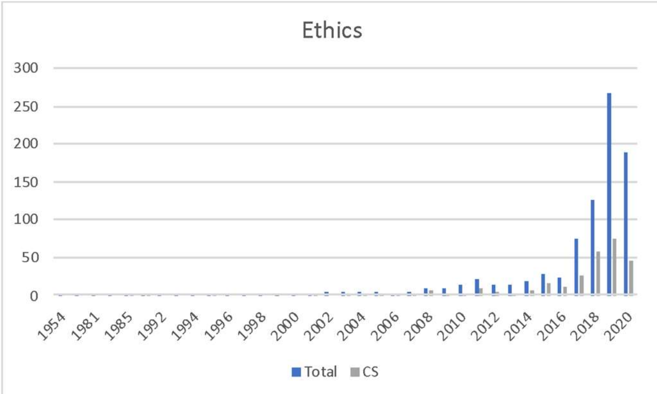
- number of sources per year (growth) per keyword pair;
- number of technical sources (computer science, CS) per year (growth) per keyword pair.

To provide a clearer picture of the growth of the volumes of publications, we have combined them into four groups: (1) everything that was published before 2000; (2) 2000-2009; (3) 2010- 2018 and (4) 2019-2020.

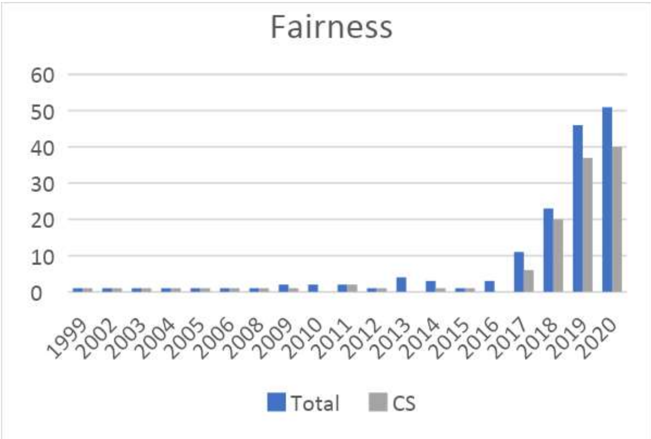
1 For example, Ethics and Information Technology journal was marked as non-technical (https://www.springer.com/journal/10676).

Table A1.2. Publications per year group. Total: total number of publications per year; CS - publications in computer science per year.

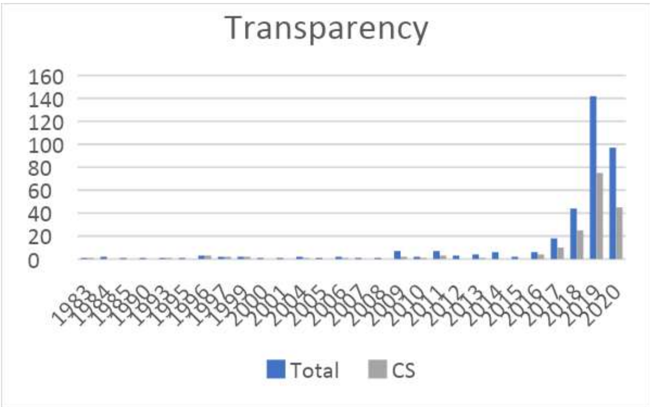
a. AI ethics



b. AI fairness



c. AI transparency



From tables A1.1 and A1.2, we can see that AI ethics discussion shifted significantly from computer science to adjacent fields. Before 1999, the related literature in computer science provides over 99% of the whole corpus, whereas over the last 2 years 40% of published studies moved to other domains, including medicine, sociology, etc.. A similar picture can be observed for non-computer science literature on the "AI Transparency" category, showing a shift from 0.02% to 38% of studies. However, for AI Fairness, the percentage of non-computer science papers is still quite low, representing only 13% of publications.

In order to fully describe the literature on AI ethics in computer science literature, we reviewed multiple sources, including workshops on the high-impact computer science conferences (NeurIPS, ICML, ICLR, CVPR, IEEE, AAAI and IJCAI), popular online courses (on AI and Data Science Ethics) and all available journal titles. Then, after consultation with researches in ethics and computer science, we constructed the following list of the main research topics, related to AI ethics:

1. Data Ownership
2. Privacy
3. Anonymity
4. Data Validity
5. Algorithmic Fairness
6. Model Transparency
7. Platform Responsibility
8. Data Licensing

Annex 2 – Key areas of AI Ethics

1. Data Ownership. The first topic that we can distinguish in the field of AI ethics, fairness and transparency is data ownership.

The terms "ownership" and "property rights" in the context of data science ethics are defined as following: "property rights (as I will deal with them here, are the rights of ownership" (Becker 1980). Property refers to the thing(s) to which property rights apply (Hummel et al 2020). In broader context, it covers questions on whom the data belongs to (data ownership), how to store it and how to use it correctly (aggregation, ownership management and privacy protection). Literature on data ownership covers such questions as direct data ownership; data aggregation operations in data analytics, machine learning and artificial intelligence (Rosa et al 2020); dynamic ownership management and privacy protection (Janeček 2018), (Yuan et al. 2018).

2. Privacy. This topic overlaps with the previous one (data ownership). However this category is more related to privacy-preserving algorithms and secure processing of data (Ryffel et al, 2018, Adams. 2014, Boyd 2011, Davidson 2015, Walter 2002). This field covers GDPR, privacy policies and regulations in different fields in different countries, therefore it has a very large volume of documents.

3. Anonymity. This field covers such topics and de- and re-identification and related issues (Witten, 2011). De-identification is the removal of identifiable information from data. Re-identification is the process by which anonymized personal data is matched with its true owner. In order to protect the privacy interests of consumers, personal identifiers, such as name and social security number, are often removed from databases containing sensitive information. This field also includes studies on face recognition. In this context, re-identification problem describes matching faces across multiple sources for the purpose of sequential authentication over space and time (Knyaz 2019, Wechsler and Li 2014).

4. Data Validity. Data validity covers appropriate use of data science methods, specifically from the field of statistics in real-world application, for example for medical records and adjacent fields. Additional sub-items within Validity can also include data accuracy, relevance and timeliness to action/need, as well as its completeness.

5. Algorithmic Fairness. This field covers identifying, measuring and improving algorithmic fairness when using AI algorithms and heavily concentrates on the algorithmic bias (including bias in images, texts, spatial data etc.) that can lead to multiple problems, such as unfairness in decision making and racial or class discrimination (Pessach and Shmueli 2020).

6. Model transparency. This field covers issues of machine learning models, related to both model transparency and data transparency (Ananny and Crawford 2018). The data transparency problem is related to the fact that end users of a model are often unaware of what data was used to train the model. Since the majority of ML models require vast amounts of data, this encourages situations where data is collected and combined from across various and unrelated sources. This might lead to class imbalance, which affects the resulting performance of the model. And while regulation and policies on data gathering might be seen to solve this problem in the long term; there is currently no way with which we can ensure that the model producer uses the reliable data for training ML models. Another issue is the transparency of the algorithm itself. A model, especially a deep learning model, is very difficult to decode once it's been trained. In many cases, there is a possibility to look inside the model and visualise what is happening inside deep learning model, giving rise to a whole separate field of computer vision who works solely on this problem (Olah 2015, Erhan et al. 2008, Simonyan et al. 2013, Zeiler and Fergus 2014).

7. Platform responsibility. This field regards platforms as moral agents, covering translating of the basic moral principles; analysis of the literature from different fields, such as philosophy and sociology in application to artificial agents, including robots, ML models, AI models and machines in general. A separate subfield deals with the representation of social media platform as moral agents and discusses the responsibilities that platforms should carry versus acting as mediators in human communication (Awad 2018, Floridi and Sanders 2004, Moor 2006).

8. Data licensing. This field covers issues of licensing of data in the fields of artificial intelligence and machine learning. There is no common framework for data licensing akin to the licensing of open source software. Two main issues can be distinguished: lack of data transparency and conceptual ambiguities in existing licensing language. Solving these issues may help foster fairer and more efficient markets for data through bringing about clearer tools and concepts that better define how data can be used in the fields of AI and ML (Benjamin 2019).

Annex 3 – Key insights from AI ethics documents

One immediate and clear issue that we can see is the fairly obvious division in understanding, and ultimately in priorities, when it comes to AI, the underlying technology, its applications, and its consequences. This finding confirms the conclusions of the narrower review done by Jobin and colleagues (2019), and as such, it is not entirely surprising. However, in this analysis we are better able to identify the nature and themes of this divergence. First, there is the issue of time. Business discussion appears to de-emphasize the impact of real time applications, focusing instead on decision support. The distinction is important, as it goes to the heart of the discussion around automation, and specifically the proportion of system suggestions that are immediately accepted by managers (thus shrinking the difference between decision support and automation). Real-time applications encompass customer-facing applications where data, systems, and processes are encoded and enshrined inside of AI umbrella, and leaves little room for course correction or adjustment by humans. Therefore, it should be clear why regulators might be more interested in these situations, and more likely to take regulatory actions against such activity.

The second crucial distinction lies in an overall underrepresentation of "accountability" as a topic in the business discourse around AI, and its narrow use in the legal sense of the word, or as "liability" equivalent. This is understandable from a number of perspectives, including the outdated legal schemes to address liability that prove to be inefficient when applied to black box AI. However, as we discuss in the main document, accountability is also associated with a leadership component that can serve as a base for proactive engagement in ethical AI during a transition phase.

By closing these gaps in understanding and refocusing the discussion, business will be better able to assume the leadership not only in the development of AI, but also in discussions on its governance.

Annex 4 – Analysis of Ethical AI Documentation

NLP Analysis of Ethics in AI documentation

Since the literature on AI ethics is significant, we used a machine learning approach to retrieve relevant information from the documents and to understands similarities and differences between those texts.

We processed 2 text corpora: business and regulatory literature corpus (divided by groups of stakeholders) and news corpus (divided by industry sectors).

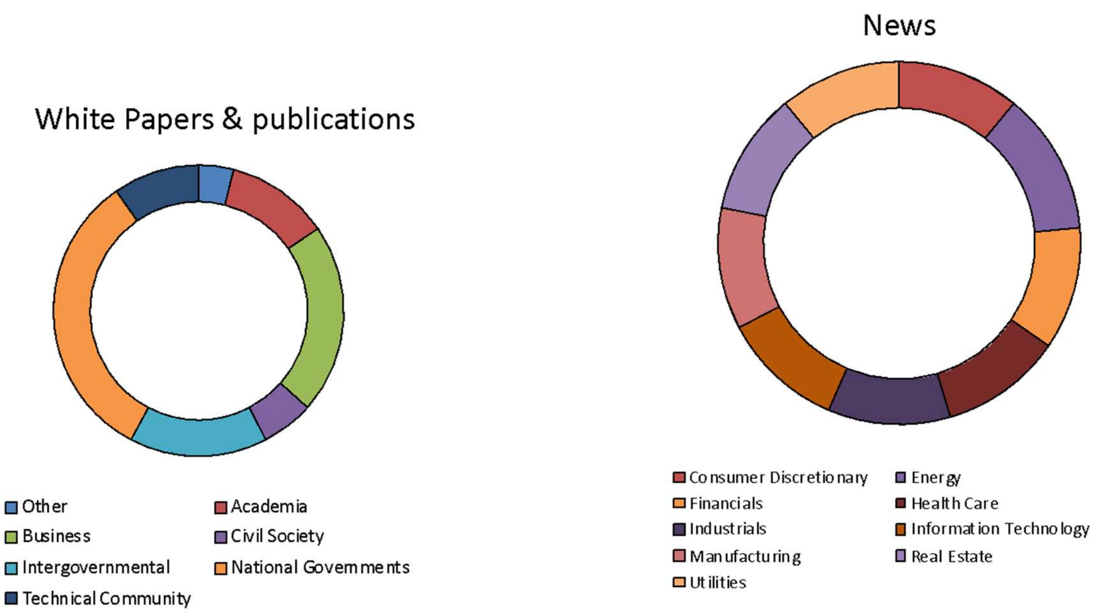


Figure 1 Text corpora used in NLP analysis consists of "Business" publications and "News" corpus.

In order to extract meaningful data from the large text corpus, we perform two types of processing: information retrieval (IR) and knowledge graph (KG) creation. Text processing was divided into 3 main parts: (1) pre-processing of the text to make it readable for ML algorithms, (2) IR and (3) KG creation. These techniques were performed on each text corpus and the results were compared between the industry sectors in case of news corpus and groups of stakeholders in case of business and regulatory literature corpus.

First, we normalized texts and created a dictionary of words from all texts to perform frequency analysis. It allowed us to understand what words (unigrams[2]) and pairs (bigrams) or triplets (trigrams) of words are used the most, thus uncovering the most important topics from the text (word-level text processing). Next, we analysed unigrams and bigrams in the text, extract text frequencies and created word clouds. In the third part, we extracted information from text to create knowledge graphs. In this task, we work on the sentence level to extract main relationships from the text. These relationships were then further used to build knowledge graphs (Jurafsky and Martin 2008).

For information retrieval, we first performed a series of operations, collectively called text normalisation. We begin this process with cleaning all texts with regular expressions: removing special symbols, hashtags, emoticons, etc., since they are not important for this research. We then separate our texts into words (or tokenize words from running text, the task of tokenization). English words are often separated from each other by whitespace, but whitespace is not sufficient in this case: we also separate punctuation (pad punctuation symbols with whitespaces etc.), expand clitic contractions that are marked by apostrophes (using Penn Treebank tokenization standard). Another part of text normalization is lemmatization, the task of determining that two words have the same root, despite their surface differences. For example, the words sang, sung, and sings are forms of the verb sing. The word sing is the common lemma of these words, and a lemmatizer maps from all of these to sing. Stemming refers to a simpler version of lemmatization in which we mainly remove suffixes from the end

[2] An n-gram is a sequence of N n-gram words: a unigram is a separate word ('intelligence"); a 2-gram (or bigram) is a two-word sequence of words like "artificial intelligence", "generalised", and a 3-gram (or trigram) is a three-word sequence of words like "generalised artificial intelligence". We use bigrams and unigrams for extracting frequencies from text.

of the word. In this case we use lemmatization. Next, we performed information retrieval[3] to extract frequent words and phrases from texts. Finally, we used sentence segmentation to break up our texts into individual sentences, using cues: question marks, periods and exclamation points.

In the description below, we will use language models (models that assign probabilities to sequences of words, LMs). We first removed the stop words from our list of tokens, created in the previous step. Stop words are very frequent words like "the" and "a". We use frequency–inverse document frequency (TFIDF) vectorizer for frequency analysis of bigrams and unigrams. TFIDF, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TFIDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. We use this method to count all the occurrences of each word in the text corpus (for bigram analysis, we do the same for bigrams) and analyse the results. Results are displayed in fig. 2 through 5 below.

---

[3] Information retrieval (IR) is the task of finding the document d from the D documents in some collection that best matches a query q. For IR we'll therefore also represent a query by a vector, also of length |V|, and we'll need a way to compare two vectors to find how similar they are.

This process of information extraction (IE), turns the unstructured information embedded in texts into structured data, for example for populating a relational database to enable further processing.

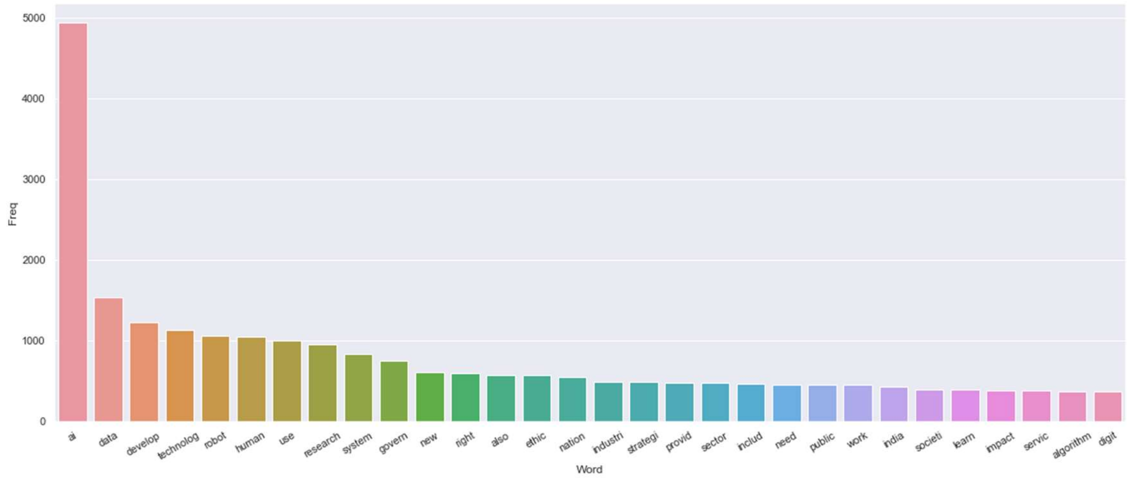


Figure 2 Unigram analysis on the text corpus (business literature)

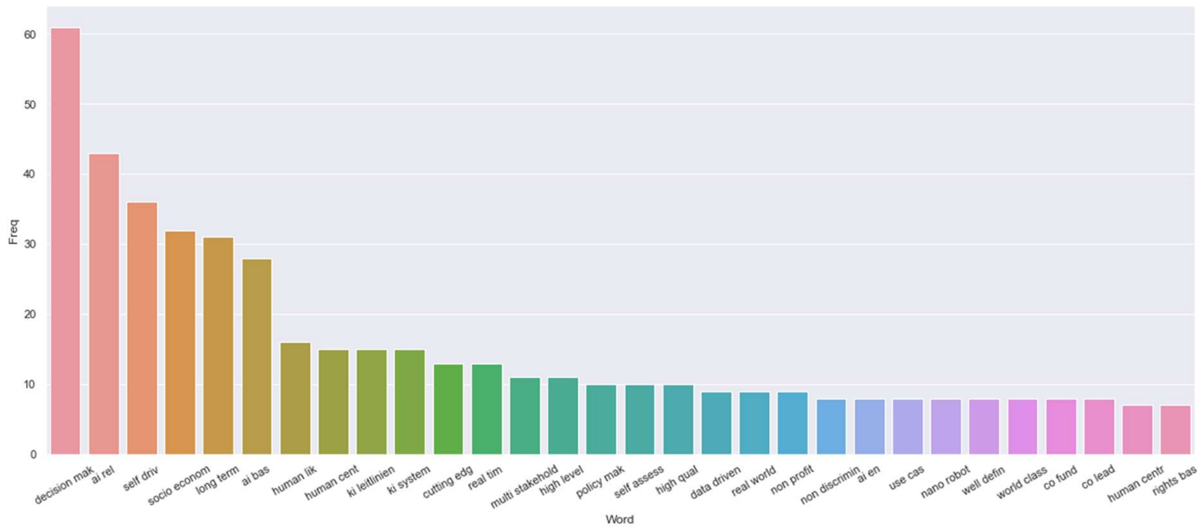


Figure 3 Bigram analysis on the text corpus (business literature)

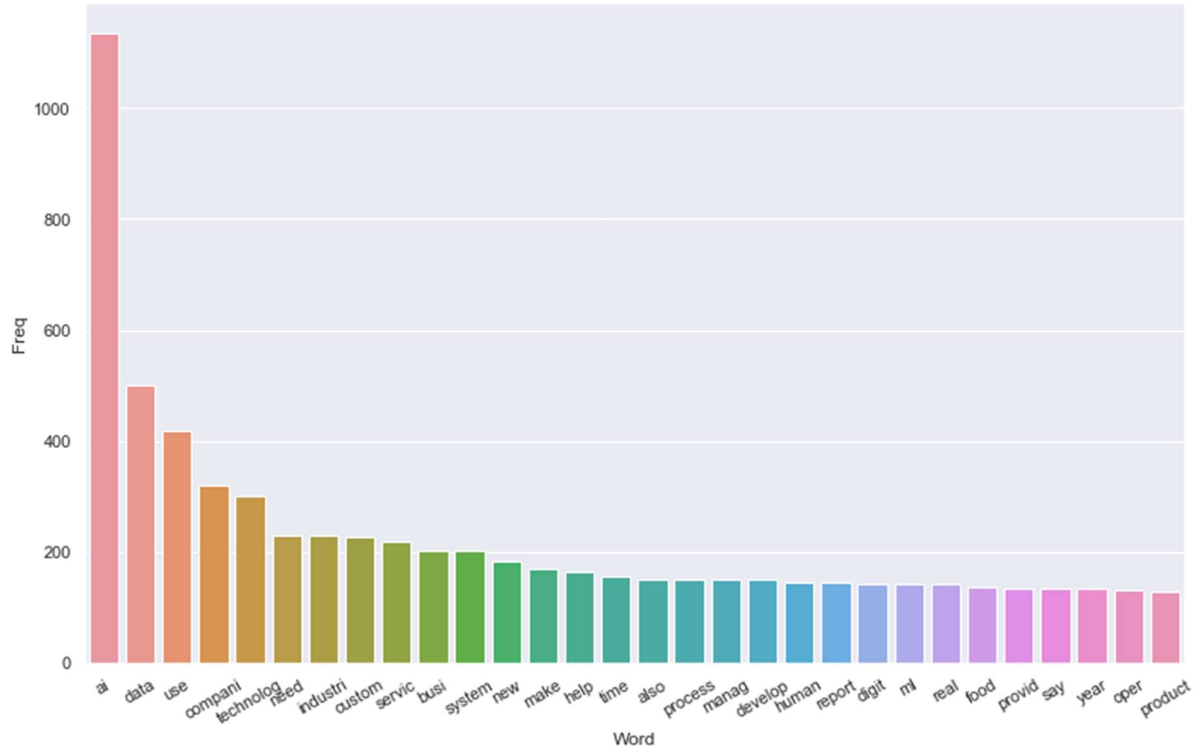


Figure 4 Unigram analysis on the text corpus (news)

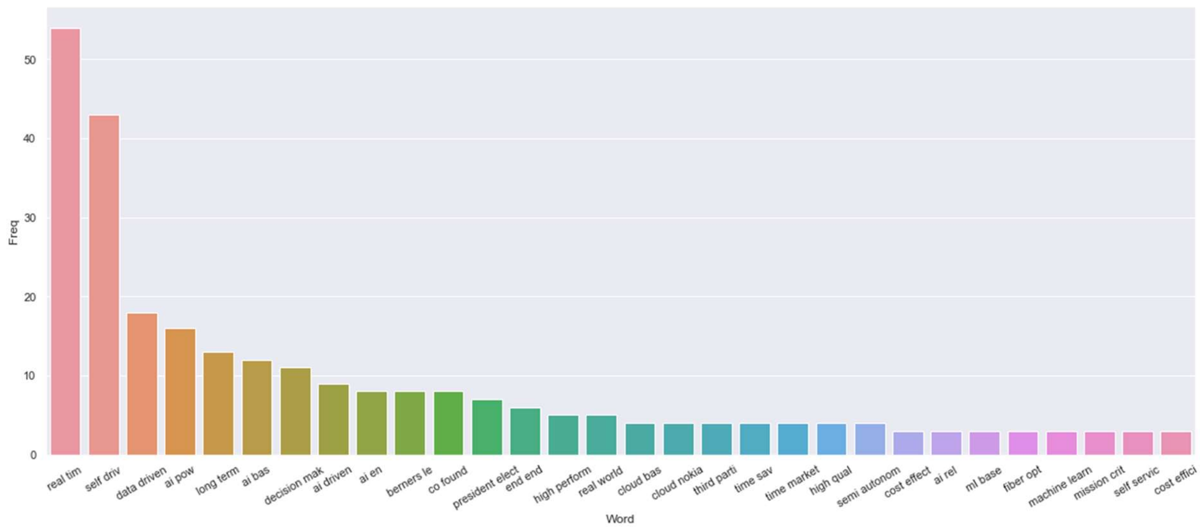


Figure 5 Bigram analysis on the text corpus (news)

Knowledge graphs are used to visually represent a text or a collection of texts, based on main topics that are discussed in the text. To build a knowledge graph, we first find all names or named entities in a text. The task of named entity recognition (NER) is to find each mention of a named entity in the text and label its type. We use long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) models to extract following entities: people, places, and organizations. Next, we extract relations: find and classify semantic relations among the text entities. These are often binary relations like “child-of”, “employment”, part-whole, and geospatial relations. For simplicity, we work on a sentence level and extract only the basic relations from the text (source, target and relation triplets).

We construct a data frame, that contains extracted entities and build knowledge graphs based on the extracted entities. Sample graph is depicted in Fig. 6. Nodes represent named entities and connections represent relations between the nodes. For our goal, relations are unimportant, since we want to understand what is discussed, not to extract the sentiment.

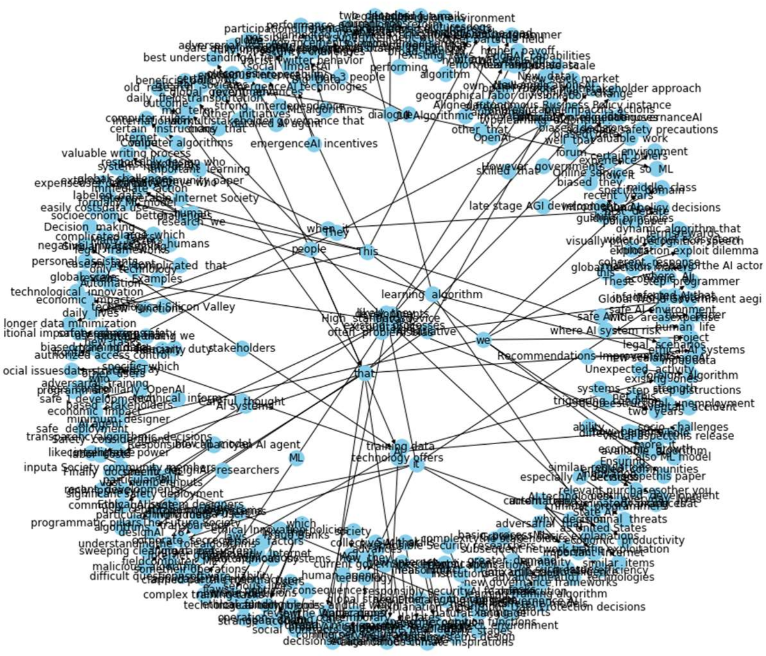


Figure 6 Knowledge graph, extracted from a subset of texts.

Since the whole graph is difficult to read, we extract only the relations, where the source is ‘AI’ (Fig.7). We do the same for each group of texts in each text corpus. Resulting graphs allow us to quickly visualise connections between the most important concepts in the text.

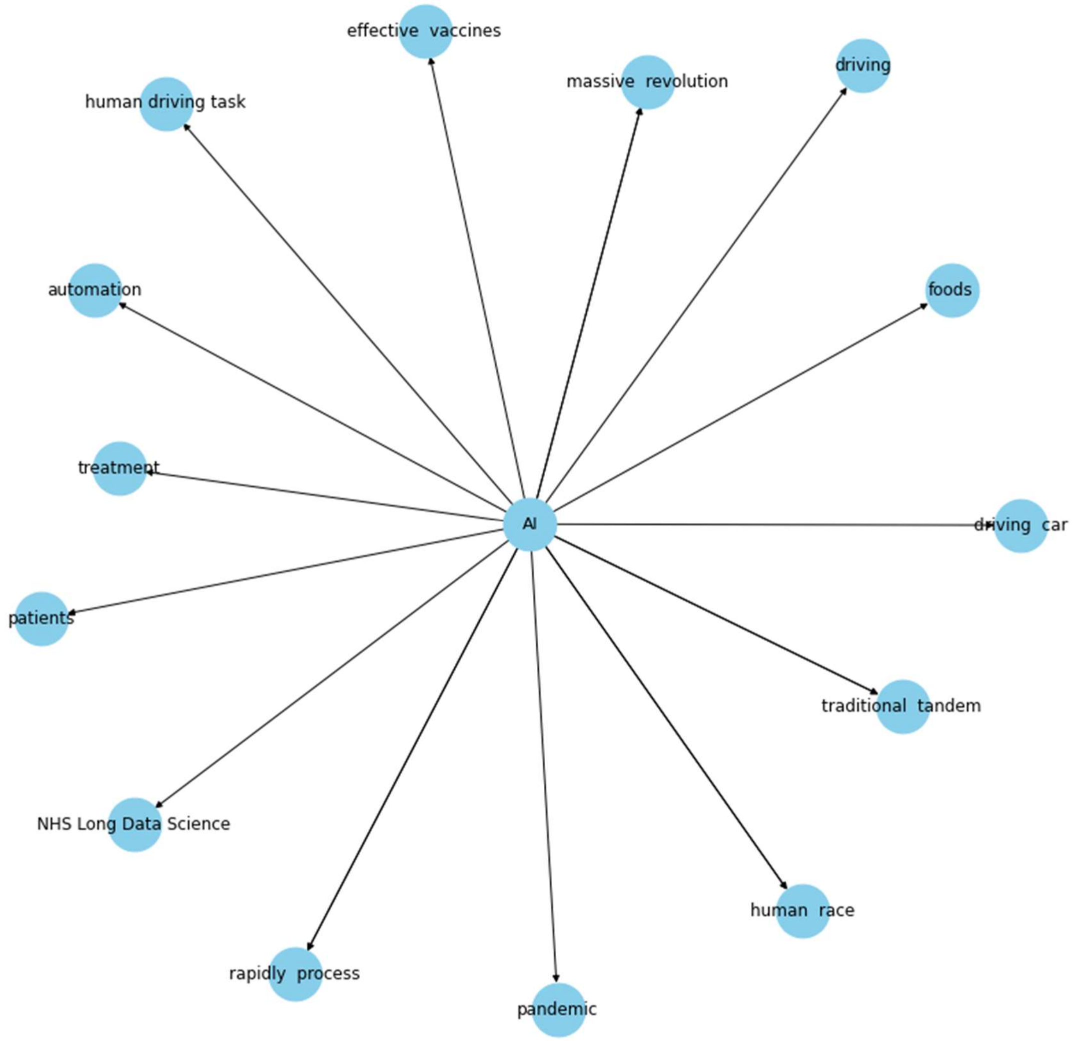


Figure 7 Knowledge graph depicting the relationships of the term 'AI' in the "News – Healthcare" corpus.

The proposed method allowed us to process the large corpus of texts and define differences and similarities between different groups of texts in an automatic manner. This approach allows not only to deal with large text corpus, but seamlessly add new data and repeat the analysis as frequently as required.

References

1. Arksey, Hilary and O'Malley, Lisa (2005), 'Scoping studies: towards a methodological framework', International Journal of Social Research Methodology, 8 (1), 19-32.

2. Pham, Mai T., et al. (2014), 'A scoping review of scoping reviews: advancing the approach and enhancing the consistency', Research Synthesis Methods, 5 (4), 371-85.

3. List of conferences: https://neurips.cc; https://icml.cc; https://iclr.cc; http://cvpr2021.thecvf.com; https://www.ieee.org/conferences/index.html; https://ijcai-21.org ; https://aaai.org/Conferences/AAAI-21/

4. Becker, Lawrence C (1980), 'The moral basis of property rights', NOMOS: Am. Soc'y Pol. Legal Phil., 22, 187.

5. Becker, Lawrence C (1980), 'The moral basis of property rights', NOMOS: Am. Soc'y Pol. Legal Phil., 22, 187.

6. Hummel, Patrik, Braun, Matthias, and Dabrock, Peter (2020), 'Own Data? Ethical Reflections on Data Ownership', Philosophy & Technology.

7. Rosa, Marco, Cerbo, Francesco Di, and Lozoya, Rocio Cabrera (2020), 'Declarative Access Control for Aggregations of Multiple Ownership Data', Proceedings of the 25th ACM Symposium on Access Control Models and Technologies (Barcelona, Spain: Association for Computing Machinery), 59–70.

8. Janeček, Václav (2018), 'Ownership of personal data in the Internet of Things', Computer Law & Security Review, 34 (5), 1039-52.

9. Yuan, Haoran, et al. (2018), 'DedupDUM: Secure and scalable data deduplication with dynamic user management', Information Sciences, 456, 159-73.

10. Ryffel, Theo, et al. (2018), 'A generic framework for privacy preserving deep learning', arXiv preprint arXiv:1811.04017.

11. Adams, A. 2014. Report of a debate on Snowden's actions by ACM members. SIGCAS Computers & Society 44, 3: 5-7.

12. Boyd D., Crawford K. 2011. Six provocations for big data. Retrieved from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1926431

13. Davidson, Roei and Poor, Nathaniel. 2015. The barriers facing artists use of crowdfunding platforms: Personality, emotional labor, and going to the well one too many times. New Media & Society 17, 2: 289-307. https://journals.sagepub.com/doi/abs/10.1177/1461444814558916

14. Joseph Walther. 2002. Research ethics in internet-enabled research: Human subjects issues and methodological myopia. Ethics and Information Technology 4, 3: 205-216. https://link.springer.com/article/10.1023/A:1021368426115

15. Ian H. Witten, Mark A. Hall. What's It All About?, in Data Mining (Third Edition), 2011

16. Knyaz, VA, Maksimov, AA, and Novikov, MM (2019), "Vision Based Automated Anthropological Measurements and Analysis", International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.

17. Wechsler, Robert T, et al. (2014), 'Conversion to lacosamide monotherapy in the treatment of focal epilepsy: results from a historical-controlled, multicenter, double-blind study', Epilepsia, 55 (7), 1088-98.

18. Pessach, Dana and Shmueli, Erez (2020), "Algorithmic fairness", arXiv preprint arXiv:2001.09784.

19. Ananny, Mike and Crawford, Kate (2018), "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability", new media & society, 20 (3), 973-89.

20. Olah, Christopher (2015), "Understanding LSTM networks". https://colah.github.io/posts/2015-08-Understanding-LSTMs/

21. Erhan, D., Bengio, Y., Courville, A. and Vincent, P., 2009. Visualizing higher-layer features of a deep network University of Montreal, Vol 1341, pp. 3.

22. Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew (2013), 'Deep inside convolutional networks: Visualising image classification models and saliency maps', arXiv preprint arXiv:1312.6034.

23. Zeiler, Matthew D and Fergus, Rob (2014), 'Visualizing and understanding convolutional networks', European conference on computer vision (Springer), 818-33.

24. Awad, Edmond, et al. (2018), 'The moral machine experiment', Nature, 563 (7729), 59-64.

25. Floridi, Luciano and Sanders, Jeff W (2004), 'On the morality of artificial agents', Minds and machines, 14 (3), 349-79.

26. Moor, James (2006), 'The Nature, Importance, and Difficulty of Machine Ethics', IEEE Intelligent Systems, 21, 18-21.

27. Benjamin, Ruha (2019), 'Race after technology: Abolitionist tools for the new jim code', Social Forces.

28. Jobin, Anna, Ienca, Marcello, and Vayena, Effy (2019), 'The global landscape of AI ethics guidelines', Nature Machine Intelligence, 1 (9), 389-99.

29. Jurafsky, Daniel and Martin, James H (2008), 'Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing', Upper Saddle River, NJ: Prentice Hall.

30. Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735