PROCEEDINGS OF SCIENCE

Trustworthy AI. The AI4EU approach

Ulises Cortés[*][†]
Universitat Politècnica de Catalunya-Barcelona Supercomputing Center
E-mail: ia@cs.upc.edu

Atia Cortés
Barcelona Supercomputing Center
E-mail: atia.cortes@bsc.es

Cristian Barrué
Universitat Politècnica de Catalunya
E-mail: cbarrue@cs.upc.edu

The Ethical, Legal, Socio-Economic and Cultural aspects of Artificial Intelligence have been a burning issue over the last five years. In the EU, especially this has been brought into focus by the High-Level Experts Group on Artificial Intelligence (AI HLEG) which produced a document with guidelines for this purpose. This paper presents an interpretation of the application of the EU guidelines to promote a Trustworthy AI for Europe.

Artificial Intelligence for Science, Industry and Society, AISIS2019
October 21-25, 2019
Universidad Nacional Autónoma de México, Mexico City, México

[*] Speaker.

[†] Prof. Cortés and Dr. Atia Cortés are partially supported by the project A European AI On Demand Platform and Ecosystem (AI4EU) H2020-ICT-26 #825619. The views expressed in this paper are not necessarily those of the consortium AI4EU. Prof. Cortés is a member of Sistema Nacional de Investigadores (SNI-III), (CONACyT, Mexico).

1. Introduction

Artificial Intelligence (AI) has recently raised to the point where it has a direct impact on the daily life of billions of individuals. AI technology is moving incredibly fast, and the marketing around it does not help in finding reasonable ways to try to understand the phenomena behind. AI is used almost in any applications sectors: transportation, health, justice, security, warfare, influence, insurance, finance, recruitment, management, personal service, assistance, *etc*.

Many of those applications are *critical*, like those in Health, Transport, Human-Robot Interaction so there is concern about AI applications that are maybe threatening for human rights, well-being, fairness or the Ethics behind the AI systems. About some of those concerns (see [7],[8] [9],[28]), we have to include technical challenges for reliability, safety, robustness. Also, we have to aware that Data-driven Machine Learning is not contextual, and therefore it lacks semantics.

We explore some examples of past unethical research in general, see §2.1. Also, we address some of the well-known examples of the misuse of AI-based technologies, see §2.1 that are in the core of the actual rise of awareness about the lack of regulation and control in the use of AI. What is clear is that the multiple ways to implement it are a real challenge for regulators that have no means to predict its Legal, Social, Economical, Cultural and Ethical (ELSEC) impacts of AI-based technologies.

As a consequence of this growing concern the AI community, first, and then the Society is articulating a response to raise greater awareness of the need to produce reliable AI, fair in its application, and that respects the laws and customs of the Society that uses it. All over the world governments have been generating reports and official documents surveying the state of the art of AI applications and suggesting possible directions, see §3. We have dedicated a section of this article, see §2.2 to discuss some of the existing initiatives, without wanting to be exhaustive. We commit an entire section to the European initiative, see §3, that is the main reason for this paper.

This article ends by discussing our vision of the need to have a reliable AI that allows a harmonious development. And this kind of trustworthy AI does not impede technological innovation, provided that these novel technologies and their integration are not harmful to the Society, in general, nor can they be used to discriminate to any of its citizens. This is one of the main objectives of the AI4EU[1] EU-funded project, see §1.1.

1.1 The AI4EU platform

Europe has an excellent tradition in AI research, and many of the most used methods and tools originated in European universities and research institutes. Current AI assets, however, often lack features that are crucial to the future need of having a human-centred AI, namely in terms of safety, usability, and respect of the ethical values that are at the centre of European culture [31]. AI4EU has identified five interconnected priority areas that are pivotal for the achievement of human-centred AI and where fundamental technological gaps exist:

- Explainable AI: An AI system should allow humans to understand the reasons behind its recommendations or decisions. It should be possible to know the data, rationale and arguments that lead to a result, to question them and to correct them.

[1] https://www.ai4eu.eu/

- Verifiable AI: It should be possible to guarantee fundamental properties (e.g. safety, privacy and security) of an AI system both before deployment and at run time.

- Collaborative AI: An AI system should be capable of operating in collaboration with humans, share knowledge with them, and take decisions collaboratively with them. This will lead to AI systems that are safer, more usable, can learn new knowledge, and can adapt to their users.

- Integrative AI: To achieve the above features, AI systems will need to integrate different AI methods and tools into new hybrid techniques, for example, combining data-driven and knowledge-based methods or symbolic and sub-symbolic techniques.

- Physical AI: An AI system should be able to interact with the physical world of humans. To do so, it must go beyond the simplifying assumptions often made in AI, and deal with issues like uncertainty, limited data availability, and limited computational resources.

The overall goal of the AI4EU project is to build a comprehensive European AI-on-demand Platform that provides innovators in all areas of society with access to expertise, knowledge, algorithms and tools for developing, deploying and funding AI-based innovations.

AI4EU will create a competing innovation ecosystem that adheres to European standards for ethical and Socio-economic impact [15]. In AI4EU societal concerns over the use and misuse of AI will be addressed with the organisation of an Ethical, Legal, Socio-Economical and Gender-Aware observatory to provide the AI community as well as European and National authorities with detailed, accurate and up to date information regarding the consequences of use and misuse of AI. Lessons from the AI4EU pilot applications, the research challenges and contributions from the AI4EU community will be used to drive the creation of a Strategic Research Innovation Agenda for AI as a major document to shape the AI European Strategy in the next 20 years.

2. Unethical research

Unethical research is not a new phenomenon, in literature, it is possible to find plenty of well-documented cases, in particular in medicine. Scientific research on human beings has been reported since the 18^{th} century when prisoners at Newgate were pardoned if they agreed to undergo variola vaccination (1721), and Edward Jenner began a series of cowpox vaccinations in children (1776) [27]. After notorious research abuses in the United States, exemplified by the Tuskegee syphilis experiments (1932 to 1972) [19] and the Willowbrook hepatitis study (1956 to 1972) [20], the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research was established in 1974, not far away in time [27].

Due to those experiences at the end of last century, Ethics become an essential dimension of human research, and it is considered both as discipline and practice. Still, even if the research has been developed under the strictest ethical parameters, the application of these results is far from the researcher's control and supervision. This effect gets amplified when some scientific results are integrated with others, and the result had not been considered initially.

2.1 Unethical uses of Artificial Intelligence

The explosion of AI technologies with connected benefits and risks are at the centre of a global discussion. The primary concern is not emergent consciousness in IA-Based systems but simply their ability to make *high-quality* and ethical decisions.

The generalisation of the use of IA-based technologies in almost every aspect of life produces real examples of biased or unethical uses of these. Some of those as the Cambridge Analytica [32], COMPAS [11], Uber's fatal crash and YouTube's recommendations of divisive and misleading content [13], has solicited a profound reflection on AI's impact among experts but also in the public sphere. We will overview some of the most notorious.

- **Algorithmic Bias**. The rapid growth of algorithm-driven services has led to growing concerns among civil society, legislators, industry bodies, and academics about potential unintended and undesirable *biases* within intelligent systems that are largely incomprehensible *black boxes* for users [3], [11]. The IEEE Standards Project on Algorithmic Bias Considerations is designed to provide individuals or organisations creating algorithms with methods to provide clearly articulated accountability and clarity around how algorithms are targeting, assessing, and influencing the users and stakeholders affected by the algorithm. [21].

- **Cambridge Analytica**. The Facebook-Cambridge Analytica affair was a major political scandal that revealed how the company Cambridge Analytica had inadvertently harvested data from millions of Facebook users to be used for political advertising without their consent. The company targeted to intentionally sway voters towards a specific candidate using AI-personalised political marketing using the personal data form their profiles. The data was detailed enough for Cambridge Analytica to create geolocalised psychographic profiles of the subjects of the data. For a given political campaign, the data was detailed enough to create a profile which suggested what kind of advertisement would be most effective to persuade a particular person in a particular location for some political event [32]. With the revelation that Facebook handed over personally identifiable information of more than 87M users to Cambridge Analytica, it is now imperative that comprehensive privacy policy laws be developed and enforced [17] [25]. After situations like this, citizens ought to have the *right to explanation* about the decisions algorithms make.

- **Disinformation and Fake News**. Disinformation is verifiably false or misleading information created, presented and disseminated for economic gain or to intentionally deceive the public[2]. It may have far-reaching consequences, cause public harm, be a threat to democratic political and policy-making processes. It may even put the protection of EU citizens' health, security and their environment at risk [22]. Alcott and Gentzkow define fake news as intentionally, and verifiably wrong or false news produced to earn money and/or promote ideologies. Their definition explicitly excludes *slanted* news, conspiracy theories, rumours and *false statements by politicians*. They argue that there is a market for verifiably false news because (1) it is cheaper to produce false than accurate news, (2) it is costly for consumers to distinguish between accurate and fake news, and (3) consumers may enjoy reading fake

[2] https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation

news because it confirms their beliefs [1].

This is the case of the YouTube algorithm [5], that has reached a controversial status for promoting unethical recommendations through its auto-played video-clip convey belt. YouTube has 1.5 billion users, and what they watch is shaped by YouTube's recommendation algorithm. There have been several scandals related to the contents provided by this algorithm, including the promotion of conspiracy theories [14] or disturbing and violent contents for children in their protected app YouTube Kids [30][13]. Unsupervised AI-based recommendation algorithms with the only goal of maximising the time a users spends in the platform can lead to the promotion of untruthful or dangerous contents.

2.2 Responses to unethical research

Social pressure and awareness is making clear the importance of Ethics for the conduct of research and for the application of its results, it should come as no surprise that many different professional associations (e.g. [29]), government agencies, and universities have adopted specific codes, rules, and policies relating to research ethics and its application.

As a reaction to these cases of unethical use of AI, there has been a backlash from a part of our society. People is everyday more aware of the possible (mis)uses of their personal and other types of sensible data. Moreover, we can see the potential reach of personalising content thanks to second uses of personal data, data aggregation or data re-identification. All of this is generating a lack of trust in AI systems, which could cause a big impact for organisations in economical terms. To achieve trust, AI system designers or practitioners need to create accurate, reliable systems with informative, user-friendly interfaces, while the operators must take the time for adequate training to understand system operation and limits of performance [25].

In recent years there has been a boom in publications of ethical guides, lists of recommendations or codes of good practice (a complete review on 84 guidelines was published in 2019 [18]). There are several governmental initiatives, such as the reports from China [4], the White House[25], the House of Lords in the UK [16], or the AI Governance and Ethics initiatives of Singapore [23] which has been awarded by the World Summit of Information Society in the field of Ethical Dimensions. Some of the largest IT companies such as Google, IBM or Microsoft have published their own list of principles of ethical AI or have set a list of requirements for a specific field (*e.g.* Microsoft's Facial Recognition principles[3]).

Along with the European initiative (see Section §3.1), other supranational groups have create their own guidelines, such is the case of the OECD [24], the IEEE [29] or the G20 [12]. The publication of the General Regulation on Data Privacy (GDPR [10]) supposed a big effort to any private and public institution gathering or using personal data to enforce the new law. Articles such as the right to be informed (Art. 13 & 14), the right to be forgotten (Art. 17) or the restriction on automated decisions and profiling (Art. 22) represent a first step in prioritising citizens' rights to privacy and transparency over performance or utility of the system. However, the GDPR lacks on actionable mechanisms to ensure such rights.

The apparition of all these guidelines, norms and regulations is a direct consequence of the way our societies emphasise penalising unethical behaviour through law-enforcement and other punitive

[3] https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work/

actions, rather than rewarding ethical behaviour or education. Still, these initiatives are voluntary commitments and maintain a focus on self-regulatory solutions, which might no be enough for our Society.

3. Ethical, Legal, Socio-Economic and Cultural aspects of AI

One of the most important issues raised by AI is its impact on jobs and the economy [25]. We have to observe that the responsibility to build trust in AI relies in different stakeholders [6],[26] and not only on the AI system. These stakeholders include different professional profiles, such as IT/AI experts, lawyers and regulators, experts in other domains using AI systems, human resources, management or executive boards and the end-users (*i.e.* the Society). We need to promote values for a responsible AI that *(i)* puts upfront the respect for fundamental rights over performance or economical benefit; *(ii)* that provides means to empower people interacting with AI by educating users and up-skill or re-skill them [4]; *(iii)* that are bind to a regulation without killing innovation. The main question remains, how can AI systems be designed to align with ethical, legal, and societal principles?

3.1 Trustworthy AI Guidelines

In 2018, the European Commission created the High Level Expert Group on AI (AI HLEG from now on), composed of 52 representatives of academia, civil society and industry. The aim of this group is to shape the European Strategy on AI by proposing guidelines and actionable questions related to ethical, legal and societal aspects of AI. The aim of the guidelines is to establish the main principles and requirements that any AI system should fulfil in order to build trust to the society. The EU approach of Trustworthy AI is based on three main concepts, which should be met throughout the system's entire life cycle:

1. it should be lawful, complying with all applicable laws and regulations;
2. it should be ethical, ensuring adherence to ethical principles and values; and
3. it should be robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.

Each component in itself is necessary but not sufficient for the achievement of Trustworthy AI. According to the AI HLEG, all three components work in harmony and overlap in their operation. If, in practice, tensions arise between these components, society should endeavour to align them [15]. The Trustworthy AI guidelines focus on the ethical and robust concepts with a human-centric approach that aims to raise different questions from the early stage of design of an AI system to the final phase of usage. The AI HLEG has identified 4 principles and 7 requirements that organisations should follow in order to build trustworthy AI systems. The document is complemented with a set of 131 questions with objective to help organisations to operationalize the requirements.

[4] See the EU strategy with the Digital Skills and Jobs Coalition https://ec.europa.eu/digital-single-market/en/digital-skills-jobs-coalition

The ethical principles are derived from the EU Charter of Fundamental Rights and the EU Treaty [31], and have been selected by their direct relation with the impact that an AI system might have over the society, or a part of it. These principles are:

- *Respect for Human Autonomy:* AI systems must respect the freedom and autonomy of individuals.
- *Prevent of Harm:* AI systems should not cause any harm that may affect human beings.
- *Fairness:* includes among others to avoid individual or group bias that can lead to discrimination and stigmatisation, ensure an equal distribution of costs and benefits, provide means of redress.
- *Explicability:* improve transparency and communication of AI systems to provide means of verification and explanation of decisions.

In order to implement these ethical principles, the HLEG-AI defines seven concrete requirements that should be covered to build Trustworthy AI systems: *(i)* Human Agency and Oversight; *(ii)* Technical Robustness and Safety; *(iii)* Privacy and Data Governance; *(iv)* Transparency; *(v)* Diversity, Non-discrimination and Fairness; *(vi)* Societal and Environmental Well-being; and *(vii)* Accountability. These requirements are applicable and should be reviewed during the life-cycle of the system (*i.e.* development, deployment and use) and can involve different stakeholders at each phase. Table §1 provides a description of the seven requirements of Trustworthy AI along with their relation to the before-mentioned ethical principles.

Table 1: Requirements of Trustworthy AI

Requirement	**Description of Topics**	**Relation to Ethical Principles**
Human agency and oversight	AI systems should respect 1) human *fundamental rights* or undertake impact assessments and 2) *human agency* by supporting their autonomy, instead of prioritising deceiving functionalities of AI systems. *Human oversight* is required in decision making to ensure that there are no ethical, legal or social conflicts. AI practitioners should develop governance mechanisms that allow human intervention at any moment of the AI system life cycle; the level of safety and control is directly related to the potential risk the AI system could cause of humans.	Respect for human autonomy

Technical Robustness and Safety	AI systems should be designed to behave as intended, minimising unintentional or unexpected harm and preventing unacceptable harm. AI systems must ensure *resilience to attack and security* by being protected against software or hardware attacks; have safeguards such as *fallback plan and general safety*, assessing the level of risk that could suppose to living beings or the environment; define *accuracy* measures that can help mitigating or correcting possible risks coming from inaccurate predictions, especially when this has a direct impact on human lives,; provide *reliability and reproducibility* techniques to prove the consistency of the AI systems and the outcomes generated.	Prevention of harm
Privacy and Data Governance	AI systems must prioritise the respect for *privacy and data protection* by enforcing existing regulations such as GDPR, using a privacy by design approach. This can be enhanced with data governance, which includes 1) guaranteeing *quality and integrity of data*, especially the data used for training (whether it is in-house or external), looking for any possible misleading or malicious data; and 2) offering protocols that define who and under which circumstances has the right to *access to data*.	Prevention of harm
Transparency	In order to build trust, an AI system should be transparent during the whole process, from data gathering and labelling to the algorithms used and the decisions made by the system. A documentation of the *traceability* process can facilitate a future auditability. AI practitioners should invest time in exploring *explainability* methodologies (both technical process and human decisions, adapted to different levels of expertise), even though this affects the performance of the system. Finally, and based on the right to be informed (GDPR Art. 13 & 14 [10]), AI systems should be identified as so when interacting with humans. Only by improving the *communication* channels, informing about capabilities and limitations of the system, there will be full transparency.	Explicability

Diversity, Non-discrimination and Fairness	One of the main consequences of the lack of trust in AI systems is the replication of socially constructed or historic biases. The *avoidance of unfair bias* is crucial to ensure that automated decisions will not be made on basis of any discrimination or prejudice. AI systems should be user-centric, guaranteeing *accessibility and universal design* for all collectives and physical and/or cognitive diversities and include *stakeholder participation* through it life cycle to obtain feedback about its design and usage.	Fairness
Societal and Environmental Well-being	AI systems should be built under a broader vision of their impact and be aligned with the Sustainable Development Goals. To ensure a *sustainable and environmentally friendly AI*, it is necessary that AI practitioners raise awareness on the resource usage and energy consumption during training phase. In addition, the *social impact* that an AI system can cause needs to be monitored and alert when they produce a negative impact (e.g. professional deskill or replacement) in order to protect humans physical and mental well-being. Bsed on recent cases of fake news and political influence, *society and democracy* should also be protected from AI systems that could lead to any unwanted behaviour from a societal perspective	Fairness, Prevention of harm
Accountability	Responsibility and accountability mechanisms are needed to ensure that the rest of requirements are being applied: 1) *auditability* processes (both internal, external or even independent) to assess the algorithm, data and design and availability of evaluation reports; 2) *minimisation and reporting of negative impact* by identifying, evaluating and documenting any possible risk; 3) Documenting and evaluating *trade-offs* between requirements with a clear mechanism and focus on ethical principles and fundamental rights; and 4) *redress* mechanisms for any foreseen situation with unwanted impacts, especially for vulnerable groups.	Fairness

Each requirement has different scopes and impacts on different stakeholders, depending on their implication in a given development phase. Even if most of the definitions in Table §1 describe *how* should an AI system behave, the responsibility should rely on all the stakeholders at different levels, where each role will have a major influence for a given requirement. Those related to traditional technological aspects such as technical robustness and safety, traceability, bias or accountability are already covered by other documents, going from internal best practices or codes of conducts provided by the organisations, to international standards and regulations. This is especially the case of the Privacy and Data Governance requirement, but also for transparency, which is strongly related to the GDPR [10]. Other requirements suppose new research challenges, such as transparency and explainable AI (also known as XAI [2]), which is considered the core value to move towards a Trustworthy AI. AI developers will need to change their programming paradigm, providing techniques of interpretability to any Machine Learning, and especially Deep Learning models.

Europe will need experts with a cross knowledge between AI, laws and fundamental rights able to assess organisations or research institutions in developing methodologies to ensure compliance with Trustworthy AI.

4. Conclusions

As seen along the paper it has never been more imperative to have an open discussion about the proliferation of AI-based technology in our lives and *how* it will affect our human rights, our privacy rights and our security. Therefore, we would like to discuss some broad suggestions for the future directions of AI research aimed at counteracting the problems presented in this paper.

There are several reasons why, we believe, it is essential to adhere to ethical norms in research and in producing Trustworthy AI-based applications. While AI technologies can provide great benefit for European Society, misuse can pose grave risks. To protect European Society from abuse of AI, we need to steer AI technology development and exploitation with clear guidelines that respect European ELSEC values.

Even if in the following enumeration we refer to researchers, the same principles apply to citizens, governments and industry when they are using scientific results for their application.

- First, norms promote the fundamental aims of scientific research, such as knowledge, truth, and avoidance of error. The social benefit of scientific results should also be included.
- A second reason, in our opinion, is that scientific research often involves a great deal of cooperation and coordination among many different people coming from various disciplines, cultural origins and institutions. Soon, we may need to consider the participation of autonomous intelligent systems. Common ethical standards promote the values that are essential to collaborative work, such as trust, accountability, mutual respect, and fairness.
- A third reason is that many of the ethical norms for scientific research help to ensure that researchers can be held accountable to the public. In the case of the research towards a Trustworthy AI, as discussed in §3.1, accountability and explainability are a crucial element.

Considering the examples discussed in §2.1, if ethical criticisms to those were to occur only in the literature, now we see an explosion of those, they would reach only a small fraction of all scientific readers (arguably the wrong ones). They could reinforce the idea that ethical issues are not intrinsic to research planning.

Even at this stage of AI adoption, public and industrial corporations need to take ELSEC and responsible approaches when creating AI systems because the industry and public bodies are already starting to see a backlash against AI implementations that play loose with ethical concerns. It is clear that creating an ethical culture among citizens, thus, requires not only thinking about ethics as a belief problem, but also as a design problem [6].

References

[1] Hunt Allcott and Matthew Gentzko. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2018.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward Responsible AI. *Information Fusion*, 58:82 – 115, 2020.

[3] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[4] China. Beijing AI Principles. https://www.baai.ac.cn/blog/beijing-ai-principles, 2019.

[5] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, New York, NY, USA, 2016.

[6] Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise A. Dennis, Gonzalo Génova, Galit Haim, Malte S. Kließ, Maite López-Sánchez, Roberto Micalizio, Juan Pavón, Marija Slavkovik, Matthijs Smakman, Marlies van Steenbergen, Stefano Tedeschi 0001, Leon van der Torre, Serena Villata, and Tristan de Wildt. Ethics by design: Necessity or curse? In Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi, editors, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 60–66. ACM, 2018.

[7] Tony Doyle. Weapons of math destruction: How big data increases inequality and threatens democracy. *Inf. Soc*, 33(5):301–302, 2017.

[8] G.L Drescher. *Good and real.* MIT Press, 2006.

[9] Virginia Eubanks. *Automating Inequality*. St. Martin's Press, New York, 2018.

[10] European Commission. *Regulation (EU) 2016/679: General Data Protection Regulation (GDPR)*. 2016.

[11] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

[12] G20. Ministerial Statement on Trade and Digital Economy. https://www.mofa.go.jp/files/000486596.pdf, 2020.

[13] Paul Lewis.The Guardian. Fiction is outperforming reality': how youtube's algorithm distorts truth. https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth, 2018.

[14] Sam Levin.The Guardian. Las vegas survivors furious as youtube promotes clips calling shooting a hoax. https://www.theguardian.com/us-news/2017/oct/04/las-vegas-shooting-youtube-hoax-conspiracy-theories,
2017.

[15] High-Level Expert Group. *Ethic Guidelines for Trustworthy AI.* European Union, Brussels, 2019.

[16] Artificial Intelligence Committee House of Lords. AI in the UK. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf, 2017.

[17] J. Isaak and M. J. Hanna. User data privacy: Facebook, cambridge analytica, and privacy protection. *Computer*, 51(8):56–59, 2018.

[18] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nat Macth Intell*, 1:389–399, 06 2019.

[19] J.H. Jones. *Bad Blood: The Tuskegee Syphilis Experiment*. The Free Press, New Year, 1993.

[20] J. Katz, A. M. Capron, and E. S. Glass. *Experimentation with human beings: The authority of the investigator, subject, professions, and state in the human experimentation process.* Russell Sage Foundation., 1972.

[21] Ansgar Koene. Algorithmic bias: Addressing growing concerns [leading edge]. *IEEE Technology and Society Magazine*, 36(2):31–32, 2017.

[22] B. Martens, L. Aguiar, E. Gomez-Herrera, and F. Mueller-Langer. *The digital transformation of news media and the rise of disinformation and fake news. An economic perspective.* Digital Economy Working Paper 2018-02. JRC, 2018.

[23] Monetary Authority of Singapore. Principles to promote Fairness, Ethics, Accountability and Transparency in the use of AI and data analytics in Singapore's financial sector. https://www.mas.gov.sg/publications, 2018.

[24] OECD. Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449, 2019.

[25] United States. Executive Office of the President and John Podesta. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.

[26] Albert Rizzo. *Ethically Aligned Design, Version 2*. IEEE, December 2017.

[27] Robert Sade. Publication of unethical research studies: The importance of informed consent. *Ethics in Cardiothoracic Surgery*, 75(2):325–328, 2003.

[28] Luc Steels. What needs to be done to ensure the ethical use of AI? In *Artificial Intelligence Research and Development - Current Challenges, New Trends and Applications, CCIA 2018, 21st Int. Conf. of the Catalan Association for Artificial Intelligence.*, pages 10–13, 2018.

[29] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. IEEE, First edition, 2019.

[30] Sapna Maheshwari.The New York Times. On youtube kids, startling videos slip past filters. https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html, 2017.

[31] European Union. Charter of Fundamental Rights of the European Union. https://www.refworld.org/docid/3ae6b3b70.html, 2012.

[32] Wikipedia. Facebook–Cambridge Analytica data scandal. https://en.wikipedia.org/wiki/Facebook–Cambridge_Analytica_data_scandal, 2019.