Introduction to Responsible AI

Ricardo Baeza-Yates
rbaeza@acm.org
Institute for Experiential AI, Northeastern University
Silicon Valley, CA, USA

ABSTRACT

In the first part of this tutorial we define responsible AI and we discuss the problems embedded in terms like ethical or trustworthy AI. In the second part, to set the stage, we cover irresponsible AI: discrimination (e.g., the impact of human biases); pseudo-science (e.g., biometric based behavioral predictions); human limitations (e.g., human incompetence, cognitive biases); technical limitations (data as a proxy of reality, wrong evaluation); social impact (e.g., unfair digital markets or copyright, mental health and disinformation issues created by large language models); environmental impact (e.g., indiscriminate use of computing resources). These examples do have a personal bias but set the context for the third part where we cover the current challenges: ethical principles, governance and regulation. We finish by discussing our responsible AI initiatives, many recommendations, and some philosophical issues.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Social and professional topics → Computing / technology policy**.

KEYWORDS

Responsible AI, AI Ethics, discrimination, bias.

ACM Reference Format:
Ricardo Baeza-Yates. 2024. Introduction to Responsible AI. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3616855.3636455

1 INTRODUCTION

Responsible AI (RAI) is becoming more and more important due to the increased social impact of the unethical and incompetent usage of AI. From applications that exhibit race or gender bias to self-driving car accidents, going through fake news and mental health issues due to generative AI [10].

Responsible AI covers all the process and governance for designing and developing an application, from the idea to the deployment, including maintenance, algorithmic audits, and accountability. The main instrumental principles and tools are described in Figure 4. On the other hand, we do not use ethical or trustworthy AI to not humanize technology. Also, we know that AI does not work all the time, so we shouldn't put the burden on the user.

During 2023 we saw the beginning of regulating the use of AI. First, USA's presidential executive order on the safe, secure, and trustworthy development and use of AI followed by the recent consensus for the final version of the European Union AI Act that was proposed in 2021 (but still the final text was not known the first days of 2024). In the realm of generative AI, China also proposed a regulation the same year, which is quite complete.

One of the main challenges of RAI is its multidisciplinary nature. The technical core is computer science but also includes philosophers expert on AI ethics, social scientists, interaction designers, technology policy experts and lawyers, among others. The stakeholders are not only designers and developers, but also owners of the technology, government regulators, and (impacted) users.

2 DETAILED CONTENT

The content of the tutorial has the following four parts:

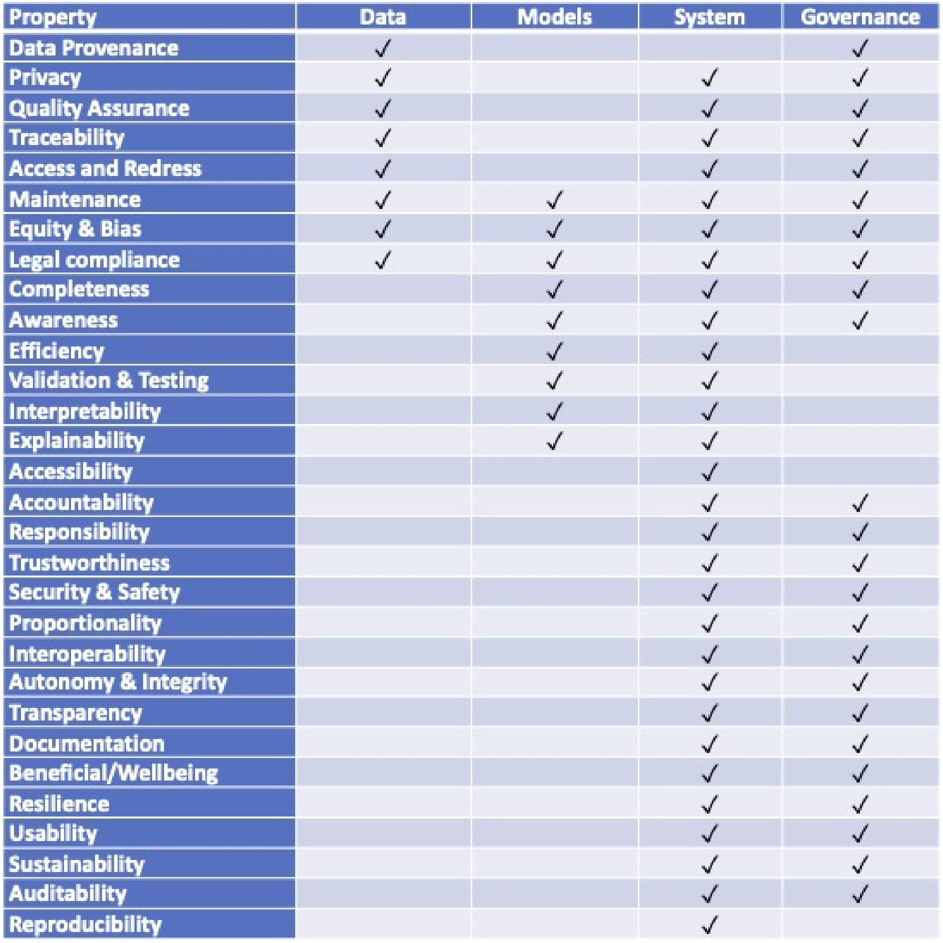
(1) Introduction
- What is responsible AI and why is it important? [10]
- The issues with ethical or trustworthy AI.
- Human intelligence vs. AI [4, 39].

(2) Irresponsible AI
- Discrimination: concepts [21], sources of bias [5], bias amplification [26], noise [24].
- Pseudo-science: physiognomy [1], predictive optimization [49].
- Human limitations: cognitive biases, technical incompetence, lack of ethics.
- Technical limitations: data issues, evaluation issues [7], errors [6].
- Social impact: unfair digital markets [9], political instability [47], disinformation [28], mental health issues [48], evaluation toxicity [44], impersonation [19], copyrights [46], etc.
- Environmental impact: indiscriminate use of resources [11].

(3) Responsible AI
- Ethical values [13] and their usage [12].
- Software properties: to which part of the AI system they apply and to which stakeholder they matter. (see Figures 1 and 2). A possible clustering of these properties based on this analysis (see Figure 3).
- Instrumental principles (OECD [33], UNESCO, ACM [8]). ACM's extended principles for generative AI [22].
- Legitimacy and competence [8]. Benefits and risks impact assessments.
- RAI governance (see Figure 4) and risk management [32].
- Algorithmic audits [17].

WSDM '24, March 4–8, 2024, Merida, Mexico.

ACM ISBN 979-8-4007-0371-3/24/03
https://doi.org/10.1145/3616855.3636455



Property	Data	Models	System	Governance
Data Provenance	✓			✓
Privacy	✓		✓	✓
Quality Assurance	✓		✓	✓
Traceability	✓		✓	✓
Access and Redress	✓		✓	✓
Maintenance	✓	✓	✓	✓
Equity & Bias	✓	✓	✓	✓
Legal compliance	✓	✓	✓	✓
Completeness		✓	✓	✓
Awareness		✓	✓	✓
Efficiency		✓	✓	
Validation & Testing		✓	✓	
Interpretability		✓	✓	
Explainability		✓	✓	
Accessibility			✓	
Accountability			✓	✓
Responsibility			✓	✓
Trustworthiness			✓	✓
Security & Safety			✓	✓
Proportionality			✓	✓
Interoperability			✓	✓
Autonomy & Integrity			✓	✓
Transparency			✓	✓
Documentation			✓	✓
Beneficial/Wellbeing			✓	✓
Resilience			✓	✓
Usability			✓	✓
Sustainability			✓	✓
Auditability			✓	✓
Reproducibility			✓	

Figure 1: Software properties and where they apply.

- Regulations on the use of AI: EU’s AI Act [20], Blueprint for an AI Bill of Rights [51], Biden’s AI executive order [50], and China’s proposal for generative AI [15].
- Interpretability [37] and explainability [3, 23, 31].
- Accountability [42].

(4) Conclusions

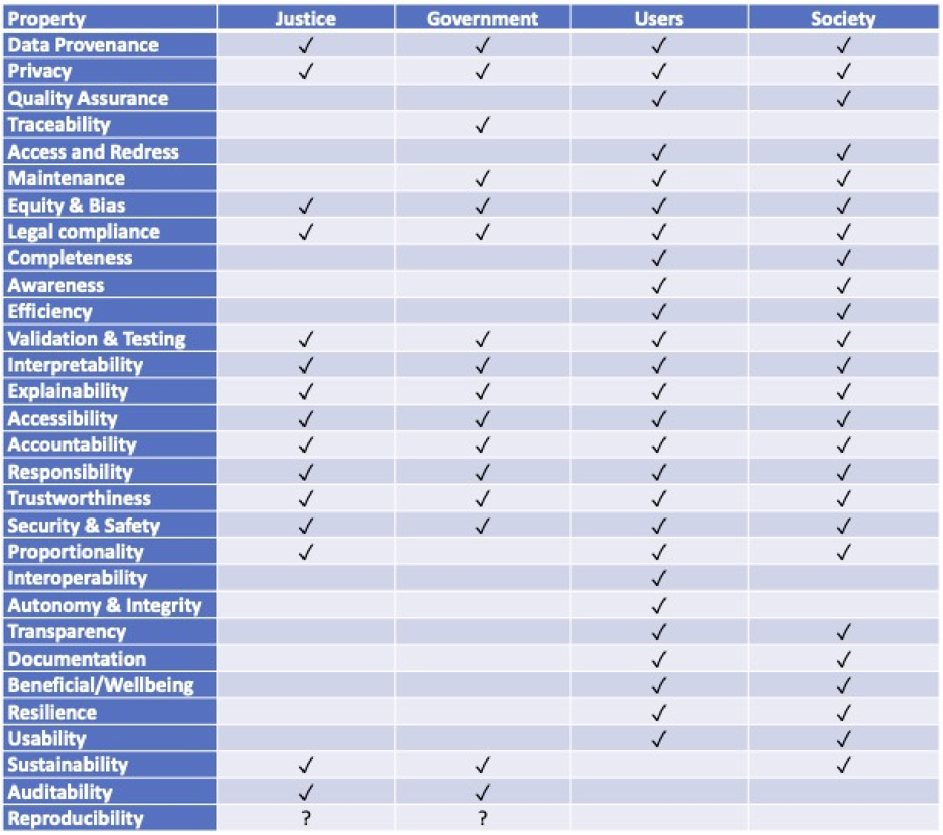
- A holistic view [35, 43].
- Recommendations [2, 27, 40].

In addition to the references above, there are many books that touch upon some of the problems previously outlined [14, 18, 25, 29, 30, 34, 36, 38], and only one of the newest books focuses on this problem [41].

3 SPEAKER BIOGRAPHY

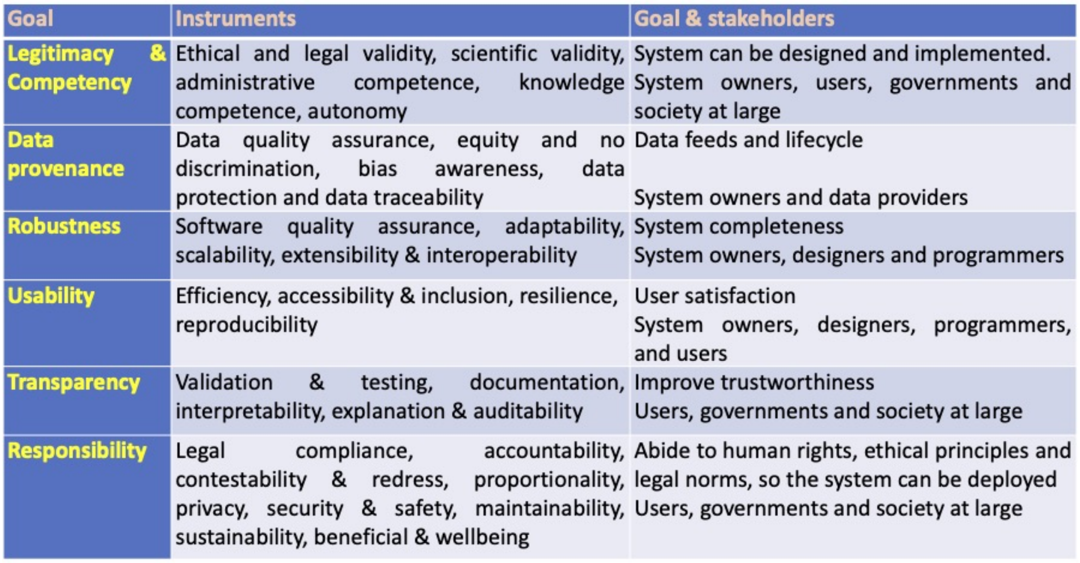
Ricardo Baeza-Yates is Director of Research at the Institute for Experiential AI of Northeastern University. He is also a part-time Professor at Universitat Pompeu Fabra in Barcelona and Universidad de Chile in Santiago. Before he was the CTO of NTENT, a semantic search technology company based in California and prior to these roles, he was VP of Research at Yahoo Labs, based in Barcelona, Spain, and later in Sunnyvale, California, from 2006 to 2016. He is co-author of the best-seller Modern Information Retrieval textbook published by Addison-Wesley in 1999 and 2011 (2nd ed), which won the ASIST 2012 Book of the Year award. From 2002 to 2004 he was elected to the Board of Governors of the IEEE Computer Society and between 2012 and 2016 was elected to the ACM Council. Since 2010 he has been a founding member of the Chilean Academy of Engineering. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow, among other awards and distinctions. He obtained a Ph.D. in CS from the University of Waterloo, Canada, in 1989, and his areas of expertise are web search and data mining, information retrieval, bias and ethics on AI, data science and algorithms in general.

Regarding responsible AI, he is actively involved as expert in many initiatives, committees or advisory boards all around the world: Global Partnership on AI, ACM’s US Technology Policy Committee, IEEE’s AI Committee and IADB’s fAIr LAC Initiative (Latin America and the Caribbean). He is also a co-founder of OptIA in Chile, a NGO devoted to algorithmic transparency and inclusion and a member of the editorial committee of the new Springer’s AI and Ethics journal, where he co-authored an article highlighting the importance of research freedom on AI ethics [16].



Property	Justice	Government	Users	Society
Data Provenance	✓	✓	✓	✓
Privacy	✓	✓	✓	✓
Quality Assurance			✓	✓
Traceability		✓		
Access and Redress			✓	✓
Maintenance		✓	✓	✓
Equity & Bias	✓	✓	✓	✓
Legal compliance	✓	✓	✓	✓
Completeness			✓	✓
Awareness			✓	✓
Efficiency			✓	✓
Validation & Testing	✓	✓	✓	✓
Interpretability	✓	✓	✓	✓
Explainability	✓	✓	✓	✓
Accessibility	✓	✓	✓	✓
Accountability	✓	✓	✓	✓
Responsibility	✓	✓	✓	✓
Trustworthiness	✓	✓	✓	✓
Security & Safety	✓	✓	✓	✓
Proportionality	✓		✓	✓
Interoperability			✓	
Autonomy & Integrity			✓	
Transparency			✓	✓
Documentation			✓	✓
Beneficial/Wellbeing			✓	✓
Resilience			✓	✓
Usability			✓	✓
Sustainability	✓	✓		✓
Auditability	✓	✓		
Reproducibility	?	?		

Figure 2: Software properties and to which stakeholders they matter.



Goal	Instruments	Goal & stakeholders
Legitimacy & Competency	Ethical and legal validity, scientific validity, administrative competence, knowledge competence, autonomy	System can be designed and implemented. System owners, users, governments and society at large
Data provenance	Data quality assurance, equity and no discrimination, bias awareness, data protection and data traceability	Data feeds and lifecycle System owners and data providers
Robustness	Software quality assurance, adaptability, scalability, extensibility & interoperability	System completeness System owners, designers and programmers
Usability	Efficiency, accessibility & inclusion, resilience, reproducibility	User satisfaction System owners, designers, programmers, and users
Transparency	Validation & testing, documentation, interpretability, explanation & auditability	Improve trustworthiness Users, governments and society at large
Responsibility	Legal compliance, accountability, contestability & redress, proportionality, privacy, security & safety, maintainability, sustainability, beneficial & wellbeing	Abide to human rights, ethical principles and legal norms, so the system can be deployed Users, governments and society at large

Figure 3: A possible clustering of properties based on the previous tables.

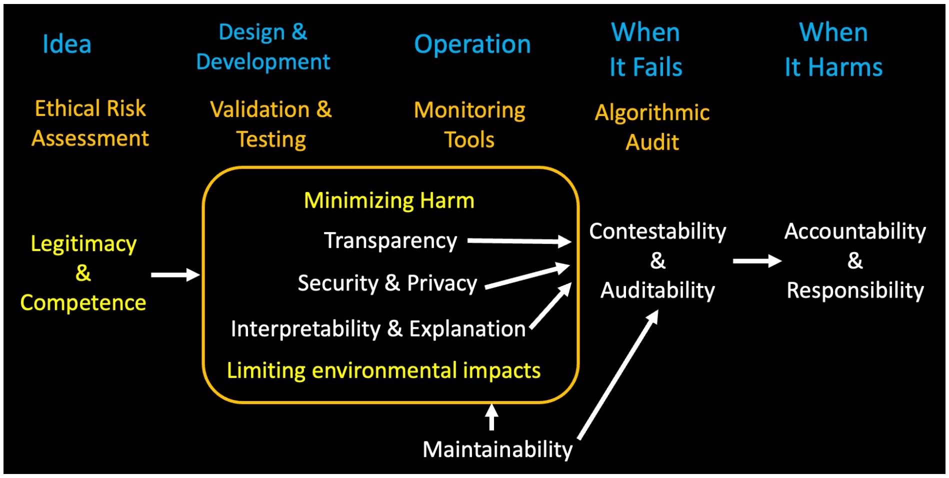


Figure 4: AI governance timeline based on newest ACM principles [8] from [10].

REFERENCES

[1] Agüera y Arcas, B., Mitchell, M. and Todorov, A. Physiognomy's New Clothes, Medium, 2017.
[2] AI Now Institute. Five considerations to guide the regulation of "General Purpose AI", 2023.
[3] Babic, B., Gerke, S., Evgeniou, T., and Cohen, I.G. Beware explanations from AI in health care, Science 373, 2021.
[4] Baeza-Yates, R. and Villoslada, P. Human vs. Artificial Intelligence. IEEE 4th Int. Conf. on Cognitive Machine Intelligence, 2022.
[5] Baeza-Yates, R. Bias on the Web, Communications of ACM, 2018.
[6] Baeza-Yates, R. Language models fail to say what they mean or mean what they say. Venture Beat, 2022.
[7] Baeza Yates, R., and Estévez Almenzar, M. The relevance of non-human errors in machine learning. Proceedings of the Workshop on AI Evaluation Beyond Metrics (EBeM 2022). Vienna, Austria. CEUR-WS, 2022.
[8] Baeza-Yates, R., Matthews, J., et al. Principles for Responsible Algorithmic Systems, ACM US TPC, 2022.
[9] Baeza-Yates, R., and Delnevo, G. Exploration Trade-offs in Web Recommender Systems. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 1-9), 2022.
[10] Baeza-Yates, R. An Introduction to Responsible AI. European Review, 31(4), 406-421, 2023.
[11] Bender, E.M., Gebru, T., McMillan-Major, A., and Mitchell, M. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ACM Conference on Fairness, Accountability, and Transparency, March 2021.
[12] Canca, C. Operationalizing AI ethics principles. Communications of ACM , 2020.
[13] Coeckelbergh, M. AI ethics. MIT Press, 2020.
[14] Crawford, K. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press, 2021.
[15] Cyberspace Administration of China. Measures for the Management of Generative Artificial Intelligence Services (Draft for Comments), 2023.
[16] Ebell, C., Baeza-Yates, R., Benjamins, R., Cai, H., Coeckelbergh, M., Duarte, T., Hickok, M., Jacquet, A., Kim, A., Krijger, J. and MacIntyre, J. Towards intellectual freedom in an AI Ethics Global Community. AI and Ethics, 1, pp.131-138, 2021.
[17] Eticas Consulting. Guide to Algorithmic Auditing, 2021.
[18] Eubanks, V. Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press, 2018.
[19] Europol. The impact of Large Language Models on Law Enforcement, 2023.
[20] European Union. The AI Act, 2021 (revised version in 2023).
[21] Friedler, S., Scheidegger, C., Venkatasubramanian, S.. The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. Communications of the ACM 64(4), 2021.
[22] Jain, R., Matthews, J., Saucedo, A., et al. Principles for the Development, Deployment, and Use of Generative AI Technologies. ACM US TPC, 2023.
[23] Jones, H. Geoff Hinton Dismissed the Need for Explainable AI: 8 Experts Explain Why He's Wrong, Forbes, 2018.
[24] Kahneman, D., Sibony, O., Sunstein, C. Noise: A Flaw in Human Judgment, Little, Brown Spark, 2022.
[25] Kearns, M., Roth, A. The ethical algorithm: The science of socially aware algorithm design. Oxford University Press, 2019.
[26] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S. Human Decisions and Machine Predictions, NBER 23180, 2017.
[27] Knott, A., Pedreschi, D., Chatila, R., Chakraborti, T., Leavy, S., Baeza-Yates, R., Eyers, D., Trotman, A., Teal, P.D., Biecek, P., Russell, S., Bengio, Y. Generative AI models should include detection mechanisms as a condition for public release. Ethics and Information Technology 25, 55, 2023.
[28] Kreps, S. and Kriner, D. How generative AI impacts democratic engagement, Brookings Institute, 2023.
[29] Marcus, G., and Davis, E. Rebooting AI: Building artificial intelligence we can trust. Vintage, 2019.
[30] Mitchell, M. AI: A guide for thinking humans. Penguin UK, 2019.
[31] NIST. Four Principles of Explainable AI, 2020.
[32] NIST. AI Risk Management Framework, 2023.
[33] OECD. AI Principles Overview, 2019.
[34] O'Neil, C. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, 2017.
[35] Pedreschi, D., Pappalardo, L., Baeza-Yates, R., Barabasi, A. L., Dignum, F., Dignum, V., ... & Vespignani, A. Social AI and the Challenges of the Human-AI Ecosystem. arXiv preprint arXiv:2306.13723, 2023.
[36] Reich, R., Sahami, M. and Weinstein, J. SYSTEM ERROR: Where Big Tech Went Wrong and How We Can Reboot by New York: Harper Collins, 2021.
[37] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intell. (1), 2019.
[38] Russell, S. Human compatible: Artificial intelligence and the problem of control. Penguin, 2019.
[39] Searle, J. Minds, Brains, and Programs, Behavioral and Brain Sciences, 1980.
[40] Shneiderman, B. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Transactions on Interactive Intelligent Systems 10(4), 2020.
[41] Shneiderman, B. Human-centered AI. Oxford University Press, 2022.
[42] Smiley, L. Aftermath of a Self-Driving Tragedy, Wired, 2022.
[43] Tahaei, M., Constantinides, M., Quercia, D., Kennedy, S., Muller, M., Stumpf, S., Liao, Q.V., Baeza-Yates, R., ... & Olteanu, A. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-4), 2023.
[44] Time. OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, 2023.
[45] UNESCO. Recommendations on the Ethics of AI, 2021.
[46] US Copyright Office. Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 2023.
[47] van Bekkum, M., Borgesius, F.Z.. Digital welfare fraud detection and the Dutch SyRI judgment. European Journal of Social Security 23(4), 2021.
[48] Walker, L. Belgian man dies by suicide following exchanges with chatbot, The Brussels Times, 2023.
[49] Wang, A., Kapoor, S., Barocas, S., and Narayanan, A. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. ACM FAccT, 2022.
[50] White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023.
[51] Whitehouse's OSTP. Blueprint for an AI Bill of Rights, The White House, USA, 2022.