EDITOR: Usama Fayyad, u.fayyad@northeastern.edu

COLUMN: AI INSIGHTS

Responsible AI: An Urgent Mandate

Ricardo Baeza-Yates and Usama M. Fayyad, *The Institute for Experiential AI at Northeastern University, Boston, MA, 02115, USA*

AI is rapidly becoming essential in various industries, raising societal expectations. AI's societal consequences include impacts on mental health; misinformation; workforce displacement; and economic, regulatory, and law enforcement challenges. Indeed, the regulation of AI usage is on the horizon, with the European Union and China already taking big steps, while the United States drafted its first AI-related bill of rights last year. Professional associations and other nonprofits are also contributing to AI ethics and regulations, increasing the urgency and criticality of this area. In this new context, public services and regulated institutions must ensure responsible AI to avoid biased or inaccurate decision-making. Similarly, companies using AI responsibly can stand out, increase efficiency, and avoid future legal problems. This article highlights the issues and problems that result in many organizations not knowing how to do responsible AI in practice, as they need to identify potential problems, set up safeguards, and conduct ethical impact assessments, among other actions. We present the issues to consider toward a comprehensive approach to responsible AI that should include defining a responsible AI strategy road map; assessing models, processes, and products; and training individuals at different levels. By covering the pressing issues related to the urgent need for adopting responsible AI, we hope to highlight the importance for corporations to seriously consider responsible AI as they rush to adopt this technology for competitive advantage.

AI is rapidly becoming an essential technology for companies in every industry. With the increased attention that generative AI has been receiving, the interest and expectations of large swaths of society and economy have been raised. As all of these trends accelerate, the companies that can effectively and responsibly adopt AI will have a competitive edge against the companies that can't. Responsible AI (RAI) is not only important because it is the right thing to do but also because governments are already recommending instrumental principles (the United States)[1] as well as proposing (China)[2] or starting to pass regulations on the use of AI [European Union (EU)].[3] Part of the issue is that there is no "standard definition" of RAI. For illustrative purposes, Technopedia.com defines it as follows: "RAI is the development and use of AI in a way that is ethically and socially trustworthy." The issue, of course, is that this pushes the definition on understanding what "ethically" and "socially trustworthy" mean—no easy feat. So, in this article, we stay at this level of reference, noting that one of the challenges is a better definition and understanding of what is meant beyond social and academic "convention."

We believe that the adoption of RAI will resemble what happened in the past with organic food or fair pricing in the food industries. However, for all its potential, there are still huge limitations in AI technologies today, and the massive misconceptions and hype about generative AI only make matters worse. In fact, the hype around artificial general intelligence exterminating humankind in the near future is just a smokescreen that distracts us from the real problems that we have today—as we expose later—that are much more important and urgent than science fiction dystopias.

1541-1672 © 2024 IEEE
Digital Object Identifier 10.1109/MIS.2023.3343488
Date of current version 7 February 2024.

All of this underscores the need for institutions to have an ethically rooted AI strategy that is Earth-centric and assumes much input and guidance/intervention from humans—what we refer to as "experiential AI" at the Institute for Experiential AI (EAI) at Northeastern University. The companies that are really using and producing AI tools and algorithms, the "AI haves" (e.g., Amazon, Apple, Google, Meta, Microsoft, OpenAI, etc.) know that human intervention is the only way to keep algorithms on track with today's AI. They also realize that capturing detailed data of the interventions—actions, outcomes, and context—is the primary fuel to power the working AI technology of today. Most organizations do not realize this and often let much of this intervention data escape in the proverbial "data exhaust."

This article aims to emphasize the importance of RAI and why organizations should pay more systematic attention to consider this aspect to avoid future problems and ensure a right culture of thinking about consequences as they employ the AI technology for competitive advantage. We cover the new aspects introduced by generative AI solutions as well as factors they need to have in their consideration set within their AI strategy. Finally, we highlight that new regulation is coming, which will require issues of compliance and at least an ability to demonstrate that companies have been thoughtful about the risks and dangers of the technology.

GENERATIVE AI ADDS NEW CHALLENGES

The generative AI currently dominating headlines is powered by a large language model (LLM) called GPT4 (initially GPT3 and GPT3.5) on which ChatGPT is built as a chatbot interface. Such models suffer from problems such as biases and other toxicities in their output, expose private data, easily commit copyright infringements during their training, do (unintended) implicit censorship, and are unable to explain where they get information (e.g., Stahl and Eke[4]). ChatGPT specifically has been shown to present wrong information as fact and give unpredictable responses when facing new situations. While such errors may seem minor problems in a school homework or draft advertising copy, when it comes to health, law, financial services, and other regulated arenas, they may severely affect the reputation, public trust, and revenues of a company.

While we are all fascinated with the apparent eloquence and "fluency" of chatbots like ChatGPT, it is important that we stay lucid and not confuse these with "intelligence" or "cognition"—we are far from systems that have a semantic understanding of what they are saying.[5] We are also far from systems that have reasoning capabilities—including common-sense reasoning—which remain elusive for machines and strictly in the domain of humans so far. In such an environment, it is particularly important that we create guardrails for how to use the technology while avoiding serious legal and ethical issues. One good practice is to use human relevance feedback to train models with more curated and well-labeled data designed to provide "adversarial" feedback as the LLM is built and used. This is how working LLMs, including ChatGPT, are pretrained and, in fact, maintained as they operate. However, this only partially solves the problem, as it does not scale to the size of the space of potential issues. It also requires too much time, computational energy, and capital.

In addition to all of these issues, there are other societal consequences that cannot yet be foreseen in their entirety:

- *Impact on mental health*: Humans have had the cognitive bias of humanizing objects from the beginning of time. Chatbots now provide the illusion of consciousness and sentience—e.g., people talking to fine-tuned models of dead persons,[6,7] an engineer that worshipped one chatbot,[8] an attempt to murder Queen Elizabeth,[9] and a suicide triggered by the content of the conversation.[10] Five days before the suicide event, Jaron Lanier clairvoyantly stated, "The danger isn't that AI destroys us. It's that it drives us insane."[11]
- *Misinformation*: The capacity of generative AI to create false images and videos with perfect voices has the potential to completely undermine the current trust in digital media. If we add to this the ease of creating such content, an avalanche of fake news may imminently severely disrupt democracy.[12]
- *Workforce impact*: Generative AI can clearly improve productivity in some professions. This acceleration of knowledge worker tasks will also naturally reduce the number of such positions and has already impacted Hollywood.
- *Law enforcement*: The breadth of issues is potentially large. A recent report by Europol states that it has already seen cases of fraud, impersonation, social engineering, and cybercrime. It provides a set of recommendations that include awareness, training, and coordination with other stakeholders.[13]

WHY ORGANIZATIONS NEED AN RAI STRATEGY

The stakes are especially high for public services, health care, or education as well as companies working on fintech, insurance, banking and financial services, telecom, or other services. Such tightly regulated institutions make decisions that can have major consequences for individuals, families, and society. Inaccurate or biased algorithms can be problematic in social media feeds, but things get much more serious when they are supporting or deciding whether to give out a loan, grant a scholarship, or approve a public subsidy. Institutions deploying models with unintended biases or without proper safeguards are, thus, exposing themselves to many potential problems, including socially harmful actions.

AI is being used to supercharge operations for companies across many industries. Institutions are using it to create faster, more relevant customer experiences, enhance market forecasting and identify macro trends, and enable a host of new operational efficiencies. Although many cases are true digital transformations, in most cases, the incentive is reducing costs, increasing—at the same time—the digital divide, since, for many services, you need to have Internet access, and it is almost impossible to speak to a live person.

However, as everyone rushes to implement AI, the companies that deploy it responsibly can stand out from competitors while avoiding legal problems and other mistakes that can erode customer trust and damage reputations. Thus, such companies can benefit from the competitive edge offered by the speed-up of work while steering clear of problematic issues stemming from the dangers and limitations of AI.

A set of considerations that organizations must account for are highlighted in the next section on potential regulation. However, we believe that if organizations can show that they are paying attention to issues of bias and inclusivity, understanding and measuring potential harms, and considering ethical aspects and dangers, then these are examples of aspects that can be addressed. At a minimum, organizations should map what the risks are in their use of AI technology, and this will give some ability to mitigate for or compensate against such risks. Unfortunately, often companies proceed while blind to these risks, resulting in unintended harm or legal exposure. We have been working on a framework for risk assessment that could extend all the way to algorithmic audits when needed.[14]

REGULATION IN THE USE OF AI IS COMING

If AI ethics is not sufficient reason to deploy AI responsibly, in the near future, it will become a matter of legal compliance. On 9 December 2023, the European Union finally reached a consensus on a revised version of the EU AI Act,[3] which forbids social scoring, emotion recognition in the workplace and educational institutions, the cognitive manipulation of people, and real-time facial recognition with just three law enforcement exceptions. This included a new article on generative AI. Next is approval by the Council of the EU and the European Commission, which will lead to formal legislation. Nevertheless, these will surely affect multinational companies like the EU data protection regulation has already done. China also published for comments a regulation for generative AI in April 2023, which will probably be the first legislation in actual operation.[2]

Meanwhile, last October, the White House's Office of Science and Technology published the "Blueprint for an AI Bill of Rights,"[1] which "applies to automated systems that have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services." The fact that these recommendations apply to all automated systems is a subtle but important detail, as it does not leave a loophole for non-AI systems. The five instrumental principles recommended for RAI are

- Safe and effective systems.
- Algorithmic discrimination protections.
- Data privacy.
- Notice and explanation.
- Human alternatives, consideration, and fallback.

In addition, the Federal Trade Commission (FTC) has also published recommendations to avoid false claims for AI products[15] as well as their concerns with "AI harms such as inaccuracy, bias, discrimination, and commercial surveillance creep." On the other hand, last March, the U.S. Copyright Office published its guidance for registering works containing material generated by AI,[16] but it is not clear yet whether there will be copyright enforcement when training AI models.

Finally, professional associations and nonprofits are also contributing. In October 2022, the Association for Computing Machinery published its new principles for responsible algorithmic systems.[17] Again, they apply to all systems (whether they use AI or not). They include a new principle on legitimacy and competence that requires an ethical impact assessment showing

that the system's benefits clearly justify the risks and that all of the needed competencies are in place: covering the spectrum from administrative to technical expertise. These principles have been complemented by a technical brief on safer algorithmic systems and an extension of the principles to generative AI. In the nonprofit category, the AI Now Institute proposed five considerations to regulate the general use of AI,[18] while the Center for AI and Digital Policy filed a formal complaint to the FTC against Open AI for violating Section 5 of the FTC Act (unfair methods of competition), the FTC guidance for AI products, and the rules for governance of AI.[19]

A DOUBLE-EDGED SWORD

The transformative nature of AI algorithms will likely lead to an algorithmic arms race in many industries. However, firms thoughtlessly rushing to deploy AI face new risks, leading them into new types of minefields: unintended biases, creating new risks, and (unintentionally) breaking regulatory issues. We already have more than 2000 examples (https://incidentdatabase.ai) of mostly unintended harms. These are just the tip of the iceberg.

Unfortunately, even as many companies agree that deploying AI responsibly is an imperative, tightening budgets are causing mass layoffs[20] of in-house ethics teams of the few companies that have such teams. Such developments are not only disheartening; they serve as motivation for us to double down on our own RAI practice at the Institute for EAI and to delve deeper into what it takes to improve trust in AI.

Even with the myriad challenges associated with AI, companies will have no choice but to adopt it to survive. While some argue that this may slow the rate of innovation, we actually do not believe that RAI is a reason to slow down but, rather, that we should move faster within guardrails that minimize risk and exposure. The only way to do that responsibly is for organizations to establish a comprehensive strategy that allows them to reap the benefits of this world-changing technology while minimizing or containing its risks. Hence, we need to take an Earth-centric and sustainable strategy, which is not only necessary for RAI but also crucial for effective AI. That implies that (ethical) humans are in charge (not just in the loop), with machines serving to help them safely achieve tasks faster, more accurately, and more efficiently.

A RAI STRATEGY THAT WORKS

A comprehensive approach to RAI involves identifying areas where problems could arise before deploying AI models. This should not slow a company's adoption of AI. Companies should also set up guardrails and conduct periodic assessments of their data, algorithms, and the operations they influence. The goal for companies is to minimize risks while showing regulators and the public that they are being thoughtful about AI's possible consequences and are, hence, worthy custodians of the trust we place in them.

At the Institute for EAI, we have been learning much about this area by partnering with different companies to ethically implement AI through our RAI services.[14] One of the reasons we do this is to understand what a teaching curriculum should cover but also what the unsolved research problems are. For example, when is an ethical impact assessment needed? How do you measure, assess, and mitigate risk? When and where are algorithmic audits needed?

Our process starts by designing an AI ethics road map. The road map includes three different areas:

1) Define an AI ethics strategy that specifies the AI governance process and the people involved in it.
2) Analyze which products/projects need to be assessed to fully understand their benefits as well as their risks.
3) Define who needs to be trained to execute successfully the preceding two areas.

The analysis in point 2 may trigger the need to implement an AI system registry that will store all data and models as well as log all steps of the process. This registry is crucial if a technical audit is needed to assess model correctness, accuracy, bias, privacy, security, efficiency, and more. To help organizations with such challenges, we have assembled an AI Ethics Advisory Board, which is made up of world-class experts in AI ethics that act as an independent advisory board, providing top private advice without internal conflicts of interest. One of the areas of concern is addressing the talent gap in this area through student co-ops, expert consulting, and training to help partners upskill their employees. Together, these services help our partners better understand the risks and capitalize on the opportunities of AI.

We believe that taking into consideration the aspects highlighted before, at a minimum, provides a checklist toward a healthy approach for adopting AI technology more safely and responsibly. Delving in without consideration of these issues will typically result in major setbacks and potentially debilitating reputational harm. Such unintended consequences, while not possible to eliminate, can at least be minimized or mitigated once identified. This puts organizations in a

better position to defend themselves against mistakes and mishaps.

For example, being aware of biases allows companies to mitigate against them (perhaps manually). Being aware of algorithmic issues allows guardrails for use in certain cases to be placed. Generally, identification of risks allows for a more informed risk appetite and a better risk–benefit tradeoff decision. Most importantly, it can minimize unpleasant and very dangerous surprises resulting from the deployed solutions as well as help companies create a more balanced and inclusive approach to development team selection and the selection of data, which are so critical to making the AI algorithms work correctly. A final benefit is an ability to measure performance and track potential risks so that interventions can happen early before matters spiral out of control.

ACKNOWLEDGMENTS

We thank many teams and colleagues at Northeastern University and the Institute for Experiential AI.

REFERENCES

1. "Blueprint for an AI bill of rights," The White House, Washington, DC, USA, 2022. [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf
2. "Measures for the management of generative artificial intelligence services," Cyberspace Administration of China, Beijing, China, 2023. [Online]. Available: https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/
3. European Parliament, "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI," Dec. 9, 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
4. B. C. Stahl and D. Eke, "The ethics of ChatGPT – Exploring the ethical issues of an emerging technology," *Int. J. Inf. Manage.*, vol. 74, Feb. 2024, Art. no. 102700.
5. R. Baeza-Yates. "Language models fail to say what they mean or mean what they say." *Venture Beat*, Mar. 29, 2022. Accessed: Dec. 21, 2023. [Online]. Available: https://venturebeat.com/ai/language-models-fail-to-say-what-they-mean-or-mean-what-they-say/
6. M. Growcott. "Man uses Midjourney and ChatGPT to resurrect his dead grandmother." PetaPixel, Apr. 17, 2023. Accessed: Dec. 21, 2023. [Online]. Available: https://petapixel.com/2023/04/17/man-uses-midjourney-and-chatgpt-to-resurrect-his-dead-grandmother/
7. M. MacCall. "A man used AI to bring back his deceased fiancée." *Business Insider*, Jul. 21, 2023. Accessed: Dec. 21, 2023. [Online]. Available: https://www.businessinsider.com/man-used-ai-to-talk-to-late-fiance-experts-warn-tech-could-be-misused-2021-7
8. L. de Cosmo. "Google engineer claims AI chatbot is sentient: Why that matters." *Scientific American*, Jul. 12, 2023. Accessed: Dec. 21, 2023. [Online]. Available: https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/
9. P. A. Media, R. Hall, and N. Badshah, "Man who broke into Windsor Castle with a crossbow to kill Queen jailed for nine years," *The Guardian*, Oct. 5, 2023. [Online]. Available: https://amp.theguardian.com/uk-news/2023/oct/05/man-who-broke-into-windsor-castle-with-crossbow-to-kill-queen-jailed-for-nine-years
10. L. Walker, "Belgian man dies by suicide following exchanges with chatbot," *The Brussels Times*, Mar. 28, 2023. [Online]. Available: https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt
11. J. Lanier, "'The danger isn't that AI destroys us. It's that it drives us insane'," *The Guardian*, Mar. 2023. [Online]. Available: https://www.theguardian.com/technology/2023/mar/23/tech-guru-jaron-lanier-the-danger-isnt-that-ai-destroys-us-its-that-it-drives-us-insane
12. Y. N. Harari, "AI has hacked the operating system of human civilisation," *The Economist*, Apr. 28, 2023. [Online]. Available: https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation
13. "The impact of large language models on law enforcement," Europol, The Hague, The Netherlands, Apr. 2023. [Online]. Available: https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement
14. R. Baeza-Yates and C. Canca, "EAI's responsible AI practice," Institute for Experiential AI – Northeastern University, Boston, MA, USA, 2022. [Online]. Available: https://ai.northeastern.edu/responsible-ai-services/
15. "Keep your AI claims in check," Federal Trade Commission, Washington, DC, USA, 2023. [Online]. Available: https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check

16. "Copyright registration guidance: Works containing material generated by artificial intelligence," US Copyright Office, Washington, DC, USA, 2023. [Online]. Available: https://www.copyright.gov/ai/ai_policy_guidance.pdf
17. R. Baeza-Yates et al., "ACM principles for responsible algorithmic systems," ACM, New York, NY, USA, 2022. [Online]. Available: https://www.acm.org/articles/bulletins/2022/november/tpc-statement-responsible-algorithmic-systems
18. "Five considerations to guide the regulation of 'General Purpose AI' in the EU's AI act," AI Now Institute, New York, NY, USA, 2023. [Online]. Available: https://ainowinstitute.org/wp-content/uploads/2023/04/GPAI-Policy-Brief.pdf
19. "Formal Complaint to FTC," Center for AI and Digital Policy, Washington, DC, USA, 2023. [Online]. Available: https://www.caidp.org/cases/openai/
20. C. Criddle and M. Murgia, "Big tech companies cut AI ethics staff, raising safety concerns," *Financial Times*, London, U.K., Mar. 29, 2023. [Online]. Available: https://www.ft.com/content/26372287-6fb3-457b-9e9c-f722027f36b3

RICARDO BAEZA-YATES is the director of research at the Institute for Experiential AI of Northeastern University, Boston, MA, 02115, USA, where he is also a professor of the practice in the Khoury College for Computer Sciences. Contact him at rbaeza@ieee.org.

USAMA M. FAYYAD is the inaugural executive director of the Institute for Experiential AI at Northeastern University, Boston, MA, 02115, USA, where he is also a professor of the practice in the Khoury College for Computer Sciences. He is also the chair of Open Insights, a company he founded in 2008. Contact him at fayyad@acm.org.