Responsible artificial intelligence : a review of current trends

Janan Arslan, Mehdi Ounissi, Gabriel Jiménez, Anuradha Kar, Daniel Racoceanu

RESPONSIBLE ARTIFICIAL INTELLIGENCE: A REVIEW OF CURRENT TRENDS

Arslan, J., Ounissi, M., Jimenez, G., Kar, A., Racoceanu, D.
Paris Brain Institute | NSERM U 1127 | Sorbonne Université, Paris, France

INTRODUCTION

- With the advent and rise of artificial intelligence (**AI**), its social, legal, and ethical implications in society have come into question.
- Chiefly amongst these concerns are the safety of AI, the treatment and handling of big data to train AI, particularly when dealing with personal data and privacy, and the ethics of AI, with concern over AI-related bias and inequity.
- High-risk and high-stake areas, such as security, finance, healthcare, medicine, and the criminal justice system, are of particular concern, as all these areas impact on human lives, interests, and rights [1].
- Responsible AI (**RAI**) has emerged as a new discipline of the AI domain to ensure AI are developed to produce equitable outcomes.
- The objective of RAI is to ensure that AI applications are developed with good intentions, that any applications are safely integrated, pose no bias (e.g., the use of AI predictions in insurance which may render some individuals ineligible) and are designed to serve humanity [1].
- RAI is still a hot topic in the social sciences. However, RAI concepts have not been fully embedded in the development of AI-based products. Additionally, terminologies surrounding RAI are often confusing, thus delaying its complete integration in real-life applications.

OBJECTIVE: To understand the current trends in opinions surrounding AI, propose RAI definitions with the hope they may become universal, and briefly summarize some attempts made in quantifying RAI.

WHAT SHOULD RAI ENCOMPASS?

- While there is no consensus regarding definitions surrounding RAI, **Figure 4** offers an alternative perspective to what RAI should include:
  - *Accountability*: Determining whether the responsibility lies with the AI, AI developer, or data providers in the event the AI does not behave in an expected way.
  - *Authority*: Which outcome should have more authority – the one determined by a human or the one determined by the AI?
  - *Fairness & equity:* Ensuring the AI functions in an equitable manner and does not bias any group of individuals.
  - *Data Privacy & Governance:* Prevention of adversarial attacks on AI-based systems by using mathematical systems, such as differential privacy.
  - *Social & Environmental Impact:* Ensuring the AI is designed with good intentions and improves our everyday lives for the better.
  - *Mechanism & Translations:* This includes explainable AI (**xAI**), intelligible AI (**intellAI**), interpretable AI (**interAI**), and transparent AI (**tAI**) and directly deals with understanding, translating, and improving AI systems uisng qualitative and quantitative measures.
  - *Safety & Security:* Ensure the AI poses no real danger or threat.
  - *Ethics:* Ensuring AI developed are in line with not only legal processes but human values.

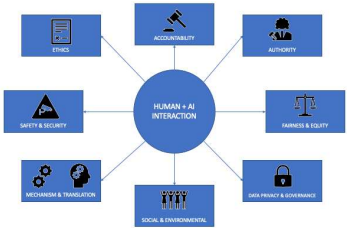

Figure 4. For a harmonious human and AI interaction, the incorporation of RAI is essential. RAI should consist of multiple components, including: accountability, fairness, safety, security, transparency, privacy and ethics.

SOME CONCERNS AROUND AI

- One of the primary concerns is AI bias (**Figure 1**)
- Prominent case of AI bias - *Amazon's Hiring Algorithm*: A computer program designed to assist with hiring future Amazon staff revealed to have a bias towards female candidates in 2015.
- Another prominent case of AI bias: the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). A tool designed to predict future crimes in courtrooms in the United States, in 2016 the COMPAS was revealed to be biased towards African American defendants.
- Search terms 'AI bias' started gaining interest in Google searches around 2016-2017 (**Figure 1**).
- In parallel, search terms "responsible AI" and "explainable AI" started gaining interest in Google searches around 2017, with interest continually increasing up until 2021 (**Figure 2**).
- Fast and Horvitz's (2017) study assessed public response to AI trends over the course of 30 years [3]:
  - Initially, AI was celebrated by the general public
  - Over time, this initial optimism morphed into outright concern as examples of AI bias rose, and the public started wondering "*When would AI takeover?*"
- The world has begun a countdown of sorts to AI takeover, as is illustrated by the book "2062: The World that AI Made" by Prof. Toby Walsh (**Figure 3**). The 'deadline' year of "2062" is an average estimate, calculated based on the interviews conducted by Prof. Toby Walsh, with interviewees including prominent members of the AI community.

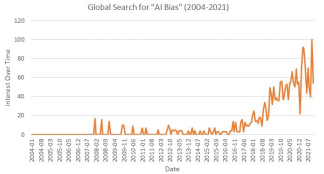

Figure 1. Google Trends Data for search terms "AI bias" from 2004-2021

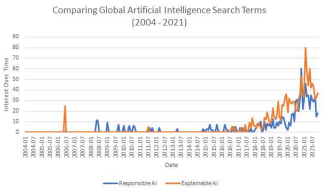

Figure 2. Google Trends Data for search terms "Responsible AI" and "Explainable AI" from 2004-2021

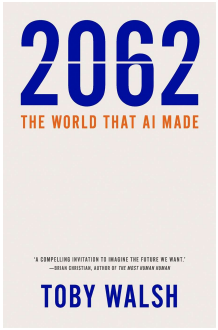

Figure 3. Book by Prof. Toby Walsh, in which the title reflects an estimate to when AI will completely encompass our everyday lives.

CAN WE QUANTIFY RAI?

- At the moment, RAI and subdisciplines are predominantly assessed from a social sciences perspective. However, for RAI to truly be effective, quantification of RAI is necessary. Such quantification can be integrated into a logical AI framework with ease and will ensure that future AI machines and software developed will adhere to expected standards.
- The only part of RAI currently most assessed in terms of quantification is *fairness*, and this is most likely attributable to the awareness built around AI bias in recent years. Some examples of fairness quantification include:
  - Individual and group fairness – statistical parity: For any two individuals x, y that are at distance $d(x,y) \in [0,1]$ map to distributions $M(x)$ and $M(y)$, respectively, such that the statistical distance between $M(x)$ and $M(y)$ is at most $d(x,y)$ [4]. In other words, the smaller the distance, the more likely fairness is achieved.
  - Predictive fairness and parity: This is to ensure same level of fairness is demonstrated, irrespective of group membership, i.e.. $P(Y = 1 \mid S = s, R = b) = P(Y = 1 \mid S = s, R = w)$ [5].
- These methods are all statistical based. Future works in this space can include quantification of other aspects of RAI, including accountability, ethics, and authority. These concepts are discussed extensively, but no consideration has been given to their quantification.

SOME FINAL TAKE AWAY MESSAGES

- Ultimately, RAI development should include engineers, data scientists, regulators, legal experts, as well as sociologists.
- Burden should not be placed on the shoulders of data scientists and engineers alone. RAI requires a multi-disciplinary collaboration to ensure AI-based systems developed within an institution are not only responsible but are aligned with what the organization has defined for itself (i.e., organizational values and mission statements).
- In lieu of any quantifiable metrics specifically designed for use in RAI (i.e., detection and correction of any bias), some cautionary measures can be taken in the interim. For example, frontline developers can check available data for any inadvertent discrimination *before* delving into AI development.

REFERENCES

[1] Islam, S. R., Eberle, W., Ghafoor, S. K. (2019). Towards quantification of explainability in explainable artificial intelligence methods. arXiv: https://arxiv.org/abs/1911.10104
[2] Shermer, M. (2017). Why artificial intelligence is not an existential threat. *Skeptic (Altadena, CA)*, 22(2):29-36.
[3] Fast, E., Horvitz, E. Long-term trends in the public perception of artificial intelligence. arXiv: https://arxiv.org/abs/1609.04904
[4] Dwork, C., et al. (2011). Fairness through awareness. arXiv: https://arxiv.org/abs/1104.3913
[5] Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv: https://arxiv.org/abs/1610.07524

CONTACT AUTHOR