DEFINING RESPONSIBLE AI

Jun 24, 2024

Roberta Calegari (Bologna University), Virginia Dignum (Umeå University)

What is Responsible AI?

Currently, for most practical uses, Artificial Intelligence (AI) is first and foremost a technology that can automatise tasks and decision making processes. However, considering its societal impact and need for human contribution, AI is much more than a technique but can best be understood as a socio-technical ecosystem, recognising the interaction between people and technology, and how complex infrastructures affect and are affected by society and by human behaviour. As such, AI involves the structures of power, participation, and access to technology that determine who can influence which decisions or actions are automated, which data, knowledge, and resources are used for learning, and how interactions between decision-makers and those impacted are defined and maintained.

The main focus of Responsible AI is ensuring that AI systems are developed, deployed, and used in a manner that is ethically sound, respects human rights, and considers societal implications. This encompasses not just ethical and legal considerations, but also the socio technical aspects that ensure that accountability for the development and use of the AI system is guaranteed. Responsible AI practices often involve processes and guidelines that organisations follow during the design, development, and deployment stages of AI systems. This could include impact assessments, reviews, and monitoring of AI systems in real-world applications.

Trustworthy AI emphasises the reliability, safety, and robustness of AI systems, as well as their ethical implications. The goal is to ensure that users and stakeholders can have confidence in AI systems' decisions and behaviours. This might involve ensuring an AI system functions correctly under various conditions, is robust against adversarial attacks, and can explain its decisions in understandable terms. Trustworthiness often requires technical solutions, such as robustness testing, adversarial training, and explainability methods, in addition to governance and ethical guidelines.

Generally, Responsible AI practices encompass Trustworthy AI requirements. A responsible, ethical, and trustworthy approach to AI will ensure transparency about how adaptation is done, responsibility for the level of automation on which the system is able to reason, and accountability for the results and the principles that guide its interactions with others, most importantly with people. In addition, and above all, a responsible approach to AI makes clear that AI systems are artefacts manufactured by people for some purpose, and that those which make these have the power to decide on the use of AI.

In this sense, AI ethics is not, as some may claim, a way to assign responsibility to machines for their actions and decisions, thereby absolving people and organizations of their own responsibility. On the contrary, ethical AI imposes greater responsibility and accountability on the

people and organizations involved: for the decisions and actions of the AI applications, and for their own decision to use AI in a given context.

Guidelines, principles and strategies to ensure trust and responsibility in AI refer to the socio-technical ecosystem in which AI is developed and used. It is not the AI artefact or application that needs to be ethical, trustworthy, or responsible. Rather, it is the people, organisations and institutions involved that can and should take responsibility and act in consideration of an ethical framework such that the overall system can be trusted by users and society.

In a nutshell, we can recap the main definitions as follows:

"Responsible AI" refers to the concept of developing and deploying AI systems in a way that aligns with ethical principles, societal values, and legal requirements. Overall, responsible AI seeks to foster the development and adoption of AI technologies in a way that promotes ethical values, respects human rights, and contributes to the well-being of individuals and communities.

"Trustworthy AI," on the other hand, refers to the concept of developing and deploying AI systems that are reliable, ethical, lawful, and transparent, thereby earning the trust of users, stakeholders, and society at large. By embodying these principles and characteristics, trustworthy AI inspires confidence and trust among users, stakeholders, and society, facilitating the responsible adoption and utilization of AI technologies for the benefit of society.

So in a way, trustworthy AI is the enabler for responsible AI. While the former is more focused on the technical aspects to build systems reliable, transparent, accountable, and ethical, thereby earning the trust of users, stakeholders, and society, the latter emphasizes the ethical and moral dimensions of AI development and deployment, aiming to promote ethical behavior, respect for human rights, and the well-being of individuals and communities.