Unravelling Responsible AI: An Umbrella Review

Gisela Reyes-Cruz[1], Elvira Perez Vallejos[1], Pepita Barnard[1], Eike Schneiders[2], Marisela Tachiquin[1], Dominic Price[1], Damian Eke[1], Liz Dowthwaite[1], Aislinn Gomez Bergin[1], Virginia Portillo[1], Joel Fischer[1]

[1]School of Computer Science, University of Nottingham, UK
[2]School of Electronics and Computer Science, University of Southampton, UK
gisela.reyescruz@nottingham.ac.uk

Abstract

The term 'Responsible AI' (RAI) has become widely adopted in various sectors, including industry, research, and policy, and has also entered general public discourse. There are significant similarities and overlap with terms such as Ethical AI and Responsible AI. As the terminology surrounding AI evolves, it is important to untangle the explicit and implicit meanings of RAI, its relationship with other relevant concepts, and the implications for the AI landscape. This paper examines the ways in which RAI has been defined and described in systematic reviews within academic research between 2013 and early 2024, by conducting an umbrella review. Five main questions are explored in these findings: 1) What is RAI? 2) What are the motivations behind RAI efforts? 3) What is its purpose? 4) What are the terms related to RAI and how they are related? and 5) What are the current challenges and future directions of RAI? This review highlights that despite the potential benefits of AI, there remain risks and concerns surrounding it. This in turn calls for a set of computational and human measures, as well as principles that span from risk mitigation to benefiting humans and addressing social problems. However, RAI is conflated, used interchangeably with, or comprises other terms, and 'responsible' and 'responsibility' are often used with different connotations. There is also an interesting contradiction between the proliferation of RAI frameworks, and the need for more actionable or operationalisable ones. We discuss the implications of these findings and offer recommendations in light of current challenges and future directions to elucidate the meanings and understandings of RAI.

Introduction

There is a pressing urgency to ensure that Artificial Intelligence (AI) is developed, deployed and used responsibly, given the numerous problematic examples documented in recent years, such as biases in criminal prediction algorithms (Angwin et al. 2016), racial disparities in healthcare (Obermeyer et al. 2019), and other instances of AI-related issues.

An analysis of the corpus of ethical AI principles and guidelines conducted in 2019 (Jobin, Ienca, and Vayena 2019) found that responsibility was one of the five converging ethical principles across the world alongside transparency, justice and fairness, non-maleficence and privacy. The analysis further found that although rarely defined, responsibility and accountability mainly referred to acting with integrity and the attribution of legal and moral liability, while fewer sources referred to responsibility as addressing the underlying causes leading to potential harm. Over five years later, responsibility has emerged as a salient concept in technology development and Responsible AI (RAI) has become an established term in industry and research. However, the current landscape is characterised by a multitude of approaches and terminology, often lacking coherence and symbiosis.

The aim of this umbrella review is to provide clarity by identifying commonalities and differences in key aspects of what is being referred to as 'Responsible AI', as found in systematic literature reviews in the past decade. Five main questions are explored in this umbrella review:

1. What is RAI?
2. What are the motivations behind RAI efforts?
3. What is RAI's purpose?
4. What are the terms related to RAI and how?
5. What are the current challenges and future directions of RAI?

Related Work

Here we review some of the origins and connections of RAI emerging from industry, academia, and the public discourse, mainly focusing on ethical AI and Responsible Research and Innovation.

Ethical AI

Contrary to what many believe, ethical AI is not a novel concept but reflects similar debates in other emerging technologies characterised by prominent impacts and uncertainties (Ulnicane et al. 2021). Indeed, ethical concerns around technology can be traced as far back as early 20th-century technologies and science fiction, notably in Isaac Asimov's "Three Laws of Robotics" (Jung 2018). These cultural reflections foreshadowed today's ethical dilemmas: can machines be moral? Who is responsible when they cause harm? How are values embedded into the technologies? Whose values? With the rise of AI, particularly generative AI, these questions have gained an increased sense of urgency.

Rooted in moral theories such as utilitarianism, deontology and virtue ethics, ethical AI encompasses discourses on Trustworthy AI and similar concepts focused on addressing the socio-cultural and ethical concerns, risks and impacts of AI (Stahl and Eke 2024). These philosophical foundations emphasise the prioritisation of values and principles such as fairness, justice, accountability and transparency as well as the maximisation of overall well-being, adherence to moral rules and the cultivation of desired virtues.

Today's landscape of ethical AI is characterised by the growing recognition of its importance by a range of stakeholders including researchers, policy makers and industry players across disciplines and sectors. By and large, these stakeholders focus on key ethical principles such as transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin, Ienca, and Vayena 2019). For a number of years, the narratives surrounding these principles centred on views and insights from the Global North. But in recent times, others have pointed to the underrepresentation of perspectives from the Global South and thus emphasise principles such as solidarity and decoloniality as essential principles or requirements for ethical AI (Wakunuma et al. 2025). However, this focus on principles by ethical AI has been criticised by some. For the critics, principles alone are not enough (Mittelstadt 2019). Ethical AI should not simply focus on alignment with principles and the development of technical safeguards but rather it requires effective regulatory frameworks.

Responsible Research and Innovation

Responsible Research and Innovation (RRI) serves as a foundational concept for understanding the broader landscape of responsible technological development, significantly influencing the discourse around RAI. Originating from policy discussions, notably gaining prominence with the European Commission's Horizon 2020 Program launched in 2014 (EC 2014), RRI emphasises an approach where societal actors work together during the whole research and innovation process to better align both the process and its outcomes with the values, needs, and expectations of society. This alignment encompasses critical considerations such as social desirability, sustainability, and ethical implications, aiming to ensure that innovation contributes positively to society and the environment (Von Schomberg 2013). RRI is characterised by key dimensions, including anticipation, reflexivity, inclusion, and responsiveness (Owen et al. 2013). It calls for considering the potential social, ethical, and environmental impacts of research and innovation from the outset, fostering inclusive dialogue with diverse stakeholders, and being responsive to changing circumstances and societal feedback. This emphasis on inclusion and stakeholder engagement draws parallels with the core principles of participatory research (Cornwall and Jewkes 1995), which advocate for active collaboration with those affected by the research throughout its lifecycle. These RRI dimensions align closely with the motivations behind RAI efforts, which seek to mitigate risks and ensure AI development and deployment benefit individuals, groups, and society while addressing social problems.

The relationship between RRI and RAI can be seen as one of specific application. While RRI provides a comprehensive framework for responsible practices across all fields of research and innovation, RAI focuses these principles specifically on the unique challenges and opportunities presented by AI. Implementing RRI in practice, however, presents significant challenges. Responsibility in research and innovation operates at multiple interconnected levels – micro (individual researchers/innovators), meso (institutional/organisational), and macro (policy/societal) (Stahl et al. 2024). Ensuring coherence and effective action across these levels is complex. Furthermore, translating RRI principles into concrete, actionable steps within dynamic research and development environments, particularly in rapidly evolving fields like AI, remains an ongoing challenge. Responsible Innovation frameworks with focus on the ICT sector (AREA Plus framework) (Jirotka et al. 2017) and high-level tools as the Responsible Innovation Prompts and Practice cards (Portillo et al. 2023) have been developed within the digital economy field to support researchers and innovators bridging the gap between the theory and the practice of responsible innovation. Although this presents opportunities for many, it may also create confusion or difficulties in implementation, a challenge observed across the RAI landscape (Portillo et al. 2022).

Methodology

To answer the five guiding questions previously presented, we conducted an umbrella review, also referred to as a review of reviews (Stefanidi et al. 2023). Umbrella reviews synthesise existing systematic reviews, thereby providing an excellent methodological approach for exploring broad questions such as the ones investigated in this paper. This section will outline our search strategy, inclusion and exclusion criteria, our screening process, as well as our data extraction and analysis processes. Figure 1 depicts an overview of the different phases of our umbrella review.

Search Strategy

Papers were obtained using the ACM Digital Library, IEEE Xplore, Scopus, and Web of Science databases. To limit the search scope, we applied a series of inclusion criteria. Firstly, we only included research papers published from 2013 to May 2024. This starting date was chosen as 2013 marks the year when Responsible Research and Innovation (RRI) frameworks were adopted by the EU (European Commission and Directorate-General for Communication and Directorate-General for Research and Innovation 2013). Secondly, only published and peer-reviewed reviews conducting systematic reviews (including narrative, scoping, rapid, systematic, literature reviews and meta-analyses) were included. Thirdly, the key topic had to be related to "responsible AI" (as long as they mentioned "responsible" or "responsibility"). The search query used for all four databases was the same[1]: ``Responsib* AND (AI OR Artificial

[1]While the search query remained the same, its syntax was adopted for the specific database.

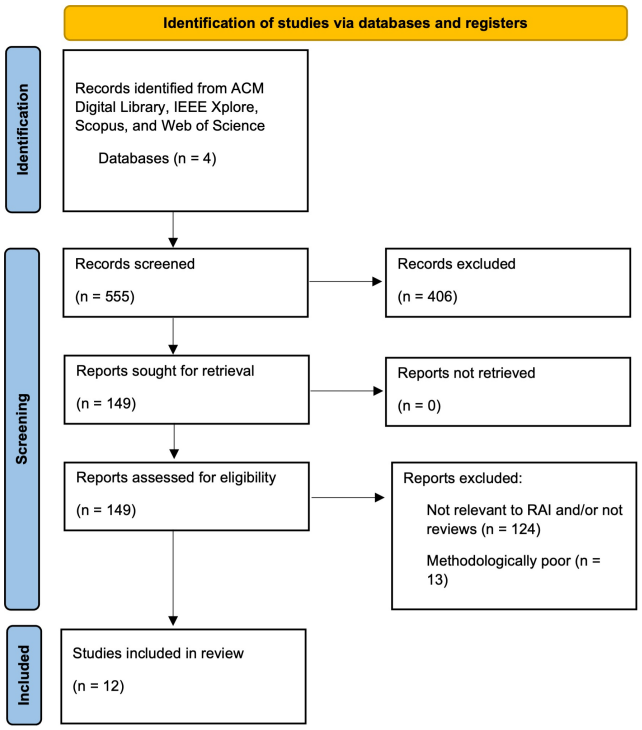

Figure 1: PRISMA flow diagram

`Intelligence) AND (overview OR review OR synthesis OR summary OR analysis OR "meta-review" OR metareview)''`. All articles had to be written in english.

Screening Process

All papers were initially screened for relevance by having two researchers independently read each paper's title and abstract. In cases of disagreement, a third reviewer assessed the title and abstract to determine relevance. Following this initial filtering, full-text screening was conducted. Subsequently, group discussions were held to calibrate individual assessment of papers and solve questions or disagreements. Spreadsheet files were used to collect the details of the selected studies to be reviewed.

- Following the removal of duplicates, the initial search of the four databases resulted in **555 papers**. Their title and abstract were screened for initial inclusion in the review following the criteria described above. This led to the removal of 406 papers.
- **149 papers** were pre-selected after the first pass. Parts of the remaining 149 were read—focusing on the abstract, method, findings, and conclusions—and were subsequently discussed to clarify and calibrate our screening process. For instance, some papers used the word 'responsibility' albeit not in a context related to AI. Papers not relevant to RAI were excluded from the corpus. This led to the removal of 124 papers.
- The remaining **25 papers** were read in full. Key points of interest were extracted and collated using a spreadsheet based approach. Again, each paper was reviewed by two researchers independently. Following the full text reads, an additional 13 papers were removed. Reasons for this included, for instance, the lack of rigorous methodological description allowing the replication of the review.
- This paper presents results based on the remaining **12 papers** (see Table 1). After additional discussions within the team, key findings described in this paper were identified.

Data Extraction

Data extraction was conducted using a spreadsheet with details of interest including definitions and concepts used to describe RAI, domains in which RAI is being applied, people affected, frameworks or guidelines referenced, purposes of RAI, motivation and contribution of the papers, target audience, year of publication, venue, discipline and paper status.

The spreadsheet was piloted by extracting the first three papers; the researchers discussed inconsistencies in the interpretation of the fields. Changes and additions to the extraction details were agreed upon by the team. All researchers extracted data from an equal number of studies to ensure a balanced distribution. Each paper was reviewed by two researchers. Additionally, the team discussed the data extracted by others.

Data Analysis

The data extracted from the selected studies was collated and synthesised in a formal narrative to address the aims and research questions, as described above. Reflexive thematic analysis (Braun and Clarke 2021) was used for the qualitative findings, following a top-down deductive coding approach to answer the five guiding questions presented in the Introduction, which are answered in detail in the results section. Descriptive results are also used to present a brief overview of the data extracted, to complement the qualitative findings.

Descriptive Overview

The domains of the articles reviewed include Education / Higher Education (3), Healthcare (2), AI and Software Development (2), Auditing (1), Management and Economics (1), Policymaking (1), Human Rights (1) and multiple domains (1). The publication years of the reviewed papers include 2021 (2), 2022 (1), 2023 (5), up to May 2024 (4). The types of review included scoping or systematic using the PRISMA flow (3), other systematic process (2), literature or narrative (4), rapid (1), systematic mapping (1), and meta-analysis (1).

Qualitative Results

In this section we describe several subthemes arranged around four main questions that aim to unravel the concept of RAI. Firstly, we attempt to provide key RAI definitions. Then we elaborate on the motivations behind and purposes of RAI, expanding on the relevant terms related to it. Lastly,

Paper title	Authors
Artificial Intelligence and Ethics in Dentistry: A Scoping Review	(Mörch et al. 2021)
Reflections on the human-algorithm complex duality perspectives in the auditing process	(Tiron-Tudor and Deliu 2022)
The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT	(Wach et al. 2023)
Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering	(Lu et al. 2024)
A Rapid Review of Responsible AI frameworks: How to guide the development of ethical AI	(Barletta et al. 2023)
Towards Concrete and Connected AI Risk Assessment (C2AIRA): A Systematic Mapping Study	(Xia et al. 2023)
Harnessing Potential: Meta-Analysis of AI Integration in Higher Education	(Samman 2024)
How to Playfully Teach AI to Young Learners: a Systematic Literature Review	(Gennari et al. 2023)
Investing in AI for social good: an analysis of European national strategies	(Foffano, Scantamburlo, and Cortés 2023)
Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review	(Singhal et al. 2024)
The Risks Associated with AI Chatbots in Teaching Future Engineering Graduates: A Systematic Review	(Moloi et al. 2023)
Fostering Human Rights in Responsible AI: A Systematic Review for Best Practices in Industry	(Baldassarre et al. 2024)

Table 1: Final list of papers included in the review

we describe the current challenges and future directions suggested in this area. Table 2 presents an overview of these findings.

1. What is RAI?

RAI definitions and concepts vary widely and here we provide an overview of how prior literature utilise and explain them, albeit to various explicit and implicit degrees.

RAI is a combination of technical solutions and human efforts. Four articles (cited below) provide a rounded understanding of RAI as a combination of approaches that include technical solutions and human involvement, in the form of human-in-the-loop (HITL) approaches and end-user education (Wach et al. 2023). One of the noted purposes of HITL is to provide an avenue for shared responsibility (which we describe in Theme 4). Moreover, RAI is perceived to entail certain obligations involving direct and indirect users by protecting and informing them to prevent and mitigate negative impacts, as well as maximising long-term benefits (Barletta et al. 2023). Educating people, where possible from young, to use and interact with AI responsibly, entails teaching them how the systems work and enabling them to participate in their AI-driven future (Gennari et al. 2023). The technical solutions developed with an RAI framing, should also follow certain principles or quality metrics that operationalise RAI such as human-centered values, fairness, privacy protection and security, reliability and safety, transparency and explainability, contestability, and accountability (Xia et al. 2023). We return to these principles and discuss them in more detail in Theme 4.

Only some specific definitions provided. It is important to note that five out of the twelve articles reviewed did not provide a specific definition, concept or explanation of what they meant by RAI or responsibility in the context of AI, although all the articles assume an understanding of what it entails. This may largely be attributed to the diverse terminology employed by many of the papers in their discussions or references to RAI. We expand on this in the following sections.

2. What are the Motivations Behind RAI Efforts?

The rapid advancement and inevitability of AI is disrupting domains. Ten out of the twelve articles reviewed noted the increasing integration of AI in their domains, such as healthcare, education, research, industry, public administration and daily life. Some articles mention the disruptions AI is causing to existing professions, for instance completely transforming the auditing practice (Tiron-Tudor and Deliu 2022), and the "*dramatical shift of the educational system's trajectory*" (Moloi et al. 2023). Digital technologies have paved the way for the inevitable integration of AI in various fields; for example, social media platforms that have served to disseminate health care information have resulted in current and future healthcare technologies with underlying AI capabilities (Singhal et al. 2024). Two articles specifically discuss Generative AI (GAI) tools such as ChatGPT, and remark its 'dark sides' (Wach et al. 2023) and market growth (Baldassarre et al. 2024). Lastly, one of the articles focuses on how governments are embracing AI to create societal and economic value, leading to the development of national AI strategies outlining not only policies but investment efforts

Themes	Subthemes	Details
Understandings	A combination of technical solutions and human efforts	Technical solutions, Human-in-the-loop approaches, obligations to protect, inform, and educate direct and indirect users, principles or quality metrics that operationalise RAI.
	Definitions	Only some specific definitions provided. RAI is generally assumed to be understood.
Motivations	Rapid advancement of, and disruption caused by AI	Transforming practices, dramatical shift of domains, current and future technologies with AI capabilities, national AI strategies embracing AI to create societal and economic value.
	Despite AI potential, challenges and risks remain	Poor data transparency, disinterest in RAI, weak quality control, job losses, AI stress, surveillance, overreliance, erosion of critical skills, misinformation, government spending lacking societal proof.
Purposes	Mitigation of AI risks	To prevent loss of public trust, data and privacy breaches, social manipulation, erosion of ethics, violations of human rights.
	Benefitting humans and addressing social problems	Streamlining administration, reducing costs and errors, human protection strategies (e.g., privacy, bias reduction), prioritising the needs of marginalised stakeholders.
Terminology	RAI conflated with or including other terms	As a variation, synonym or component of ethical and trustworthy AI; used alongside or as an umbrella term for other principles.
	Different meanings of 'responsible' and 'responsibility'	Accountability, organisations' commitment to social responsibility, practices and principles, responsible use of and education about AI, personal, professional or external motivations and incentives.
Challenges and directions	Actionable tools, guidelines, practices and processes	Proliferation of frameworks and tools, some with operationalisation shortcomings.
	More empirical evidence of AI effects in society	Impact of AI tools over time, empirical understanding of the relationship and material practices between humans and AI
	Formal training on AI skills inc. ethical and social impacts	Innovative tools for teaching AI, focusing on academic training, critical thinking, problem-solving, ethics and humanities.
	Domain-specific standards and professional codes	Governance structures collaboratively developed by stakeholders across developers, communities, institutions, industries, etc.

Table 2: Overview of umbrella review findings

(Foffano, Scantamburlo, and Cortés 2023).

The radical changes introduced by AI across domains is undeniable, but the framing of such AI proliferation can have different connotations. Five out of the twelve articles provide a positive or neutral framing of the impacts of AI. For instance, a neutral example remarks that as audit firms deal with AI implementation, the interaction between humans and algorithms gets configured and reconfigured in various ways within the auditing practice (Tiron-Tudor and Deliu 2022), with no positive or negative implication explicitly mentioned. Moreover, other articles have found positive impacts of AI, such as substantiated learning experiences, admin streamlining and data-driven decision-making (Samman 2024), significant cost savings and minimal possible human errors (Moloi et al. 2023) in Higher Education. Ultimately, some consider that the monetary investment surrounding AI could have a positive effect in the sphere of AI for social good (Foffano, Scantamburlo, and Cortés 2023). In a similar vein, some consider the value of companies that are putting their efforts into the Research & Development of technologies and methodologies for a range of issues (e.g. privacy preservation and bias reduction, user-friendly tools for managing personal data, etc) (Baldassarre et al. 2024).

Although AI could improve practices across domains, challenges and risks remain. Nine articles noted that despite the current and potential future benefits of AI, complex challenges and risks remain. These include ethical concerns in the health care practice (e.g. lack of transparency in data) (Mörch et al. 2021; Singhal et al. 2024), concerns about the human-AI relations, evolution and implications for the auditing field (Tiron-Tudor and Deliu 2022), risks of using GAI in businesses (e.g. lack of quality control, disinformation, job losses, AI techno stress, surveillance, privacy violation) (Wach et al. 2023), and the pedagogical, ethical and sociocultural challenges and effects in educational contexts, especially in the long term (e.g. overreliance, decreased critical thinking and problem-solving skills, misin-

formation) (Samman 2024; Moloi et al. 2023). Likewise, it is noted that companies require more guidance to address the emerging challenges for social responsibility in the AI context (Baldassarre et al. 2024), and that there is little clarity as to what extent governmental investments in AI contribute to the good of people and society as a whole (Foffano, Scantamburlo, and Cortés 2023).

Although there is a significant global concern about the risks of developing and using AI, there is also an increasing recognition that approaches to promote RAI are needed in order to address or mitigate them (Xia et al. 2023). Nonetheless, two articles stress the lack of information and interest in addressing ethical and Fairness, Accountability, Transparency, Ethics (FATE) challenges in some healthcare areas, despite the myriad of existing and emerging approaches and metrics created for those purposes (Mörch et al. 2021; Singhal et al. 2024).

3. What is the Purpose of RAI?

From risk mitigation to benefitting humans and addressing social problems. RAI has become an increasingly urgent topic of interest in academia, industry and civil society (Xia et al. 2023), with recent research and media reporting on concerns related to AI (e.g. in decision-making) (Barletta et al. 2023).

Four articles establish that mitigating AI risks and avoiding their negative consequences is a crucial means to achieve RAI. For instance, to avoid loss of public trust and unwanted results in healthcare practice (Mörch et al. 2021), mitigating data and privacy risks, as well as avoiding social manipulation and the weakening of ethics and goodwill in society (Wach et al. 2023), and to mitigate risks related to violations of human rights and dignity (Xia et al. 2023).

Five papers note that RAI mainly refers to the practice of development, deployment, and use of AI systems that benefit or have a positive impact on individuals, groups, society and the environment (Lu et al. 2024; Xia et al. 2023), whilst also offering the corresponding guidance for those purposes (Foffano, Scantamburlo, and Cortés 2023). In this view, RAI practices not only involve risk mitigation but human protection strategies or actions (Baldassarre et al. 2024), prioritising all stakeholder needs, especially from those minoritised or disadvantaged (Barletta et al. 2023; Baldassarre et al. 2024). RAI can also be understood as the specific set of practices enacted by different technology companies (Baldassarre et al. 2024).

4. What are the Terms Related to RAI and How?

RAI is often conflated, used interchangeably with, or comprising other terms. The majority of articles reviewed fall in this category. Four of these conflate RAI with other terms, two of them mention responsibility closely linked or alongside other terms, and two do not explicitly mention 'RAI' but other concepts or principles that are related to the responsible use and development of AI.

Amongst these configurations, RAI is used as a variation, synonym or component of ethical and trustworthy AI (Lu et al. 2024; Barletta et al. 2023; Xia et al. 2023). For instance, the search strategy employed for the multivocal literature review was expanded to include 'responsible' as well as "*its variations, such as ethics, ethical, responsibility, trust, trusted, trustworthiness, and trustworthy*" (Lu et al. 2024). On some occasions, the responsible use of AI is framed alongside or as part of ethical guidelines or principles, for instance a critical examination of ethical dimensions being a matter of personal and professional responsibility (Moloi et al. 2023), sitting next to transparency in data usage, bias mitigation techniques, etc (Wach et al. 2023).

In other articles, RAI is used alongside other principles (beyond only ethical and trustworthy AI) on an equal level. These include prudence, equity, privacy, democratic participation and solidarity (Mörch et al. 2021); trust, legal restrictions, ethical concerns, security (Tiron-Tudor and Deliu 2022); socially aware AI, AI for social good and beneficial AI (Foffano, Scantamburlo, and Cortés 2023); and digital rights (Baldassarre et al. 2024).

RAI is also employed as an umbrella term for various principles or challenges. For instance, comprising transparency, justice and fairness, non-maleficence, and privacy (Barletta et al. 2023); the quality metrics for operationalising RAI in risk assessments (mentioned in Theme 2) (Xia et al. 2023), the FATE principles (Singhal et al. 2024), and bias misinformation, hate speech, privacy and cybersecurity (which should be addressed to ensure responsible industry practices) (Baldassarre et al. 2024).

While one of the articles acknowledges the uncertainties and nuances around the definition of RAI and its principles (Barletta et al. 2023), another article underlines that some of these terms largely mean the same thing or cover the same principles (Lu et al. 2024), and a last one remarks that despite their nuances, these terms all share a common goal: "*to promote the development, deployment, and use of AI systems that have a positive impact on individuals, groups, and society while minimizing associated risks*" (Xia et al. 2023).

In a similar vein to the diverse terminology used for RAI, there are various meanings surrounding 'responsible' and 'responsibility. We elaborate on those next.

Different meanings of 'responsible' and 'responsibility' in the context of AI. The first notion of responsibility encountered in this umbrella review refers to accountability, which in turn can have various connotations (legal, ethical, technical, and societal) (Singhal et al. 2024). Six articles discuss the pitfalls and challenges of AI decision-making, as opposed to and in relation to human decision-making. Some raise the question of who is accountable for the consequences, especially when things go wrong in Human-out-of-the-loop approaches (e.g. autonomous AI operating independently without human intervention)? End-users, institutions or developers? (Tiron-Tudor and Deliu 2022). The development and use of AI must not contribute to lessening the responsibility of humans (Mörch et al. 2021), and therefore a clear responsibility identification for AI and its actions across stages must be ensured, by including developers, organisations, decision-makers, regulatory bodies, service providers and authoritative figures (Wach et al. 2023; Baldassarre et al. 2024; Xia et al. 2023). For instance, role-level accountability could be established through formal

contracts in order to determine responsibility boundaries (Lu et al. 2024).

Secondly, we found a notion referring to organisations' commitment to social responsibility. For instance, organisations must ensure that "*the developed AI systems are responsible throughout the entire software development lifecycle*" (Lu et al. 2024). For instance, by considering FATE perspectives through computational and methodological approaches, which "*ensure their responsible application*" (Singhal et al. 2024). Organisations must continuously monitor AI systems and implement strategies for risk mitigation (Singhal et al. 2024), as well as reporting their reliance on AI, ensuring the quality of data they use and its governance, and providing the appropriate training to stakeholders (e.g. employees, end-users) (Wach et al. 2023). For some, AI companies ultimately have the responsibility to weigh the different benefits of AI against the potential impact on society and human rights (Baldassarre et al. 2024).

The third notion refers to a set of RAI practices or principles that guide the development and deployment of AI as per the purposes established in Theme 3. However, clear guidance is only one aspect; ensuring that it is followed *responsibly* is equally important and should involve compliance with laws, policies and standards (Lu et al. 2024; Foffano, Scantamburlo, and Cortés 2023). Some highlight that ensuring stakeholder engagement and human oversight throughout the lifecycle is necessary to build and deploy AI responsibly (Tiron-Tudor and Deliu 2022; Lu et al. 2024; Gennari et al. 2023). Nonetheless, it has been noted that there is a problem of principle proliferation regarding RAI principles, which has also been the case for Ethical AI (Barletta et al. 2023).

The fourth notion relates to the (responsible) use and education about AI. Whereas the other concepts described above mostly refer to the development stage and the various actors involved in the process, this meaning largely focuses on direct and indirect end users, which is considered a crucial area of discussion (Wach et al. 2023; Xia et al. 2023; Samman 2024; Gennari et al. 2023; Moloi et al. 2023). Nonetheless, we often found the term 'using AI responsibly' somewhat ambiguous with little or no explanation of what it means.

Lastly, the fifth concept of responsibility refers to personal, professional or external motivations and incentives to follow the recommended principles in a determined manner (Moloi et al. 2023). These can be closely knitted to corporate social responsibility (Baldassarre et al. 2024) and governmental or global commitments (Foffano, Scantamburlo, and Cortés 2023).

5. What are the Current Challenges and Future Directions of RAI?

Despite AI popularity and the promises it holds, we found further calls for a continuous and more targeted focus on the ethical, societal, and environmental concerns of AI, through computational and human approaches (Mörch et al. 2021; Wach et al. 2023). Whilst some highlight the need for further public and private partnerships to address societal and environmental issues, for instance through hackathons and competitions (Foffano, Scantamburlo, and Cortés 2023), others advocate for the creation of computational methods merged with ethical evaluations that can quantitatively assess ethical components in AI systems (Singhal et al. 2024). Specific future directions in response to current challenges are described next.

Actionable tools, guidelines, practices and processes throughout the AI lifecycle. Four articles reviewed stress that AI ethical frameworks (which in turn are closely related to RAI, as explained in the previous sections) by various actors such as companies, universities, non-profit organisations, communities, and government entities *proliferate* (Lu et al. 2024; Barletta et al. 2023; Xia et al. 2023; Baldassarre et al. 2024). Some point out that these frameworks have shortcomings in how they are operationalised; in other words, that they are often too high-level, offering more theoretical than practical support, and with no or limited actionable practices or concrete implementation strategies (Baldassarre et al. 2024; Xia et al. 2023; Lu et al. 2024). Furthermore, the sheer number of frameworks makes it harder for stakeholders to keep themselves continuously aware of them and determine the most suitable option for their context (Xia et al. 2023). Another perceived limitation is that frameworks are mainly provided by private companies (Barletta et al. 2023). One article notes that substantial efforts are placed at the algorithm level during development (e.g. embedding fairness) and that RAI should be operationalised instead at the system level and across the lifecycle (Lu et al. 2024). Likewise, other articles remark that most ethical frameworks focus on the requirement elicitation phase, and that there is no comprehensive framework covering all RAI principles and software development life cycle phases whilst supporting various technical and non-technical stakeholders throughout (Barletta et al. 2023; Singhal et al. 2024; Xia et al. 2023).

Some initial solutions to this problematic are proposed, such as an RAI Pattern[2] Catalogue that stakeholders can use across products, processes and governance levels, by selecting from a range of existing tools, guidelines and processes (Lu et al. 2024). Likewise, 18 actionable recommendations for AI companies are suggested (Baldassarre et al. 2024), and a question bank with risk assessment questions labelled with different RAI principles, stakeholders and lifecycle stages is also proposed for future work (Xia et al. 2023). However, as there is still a lack of clear consensus about RAI standards and tools, further research is considered necessary. Interestingly, some advocate for establishing clearer and better guidelines, standards and frameworks for achieving the purposes of RAI (Wach et al. 2023; Barletta et al. 2023; Moloi et al. 2023), including those that relate to the use of AI, as described next.

More empirical evidence of AI effects in society. Four articles note the importance of empirically testing the relationship and material practices between humans and AI, and how the latter will affect (or is already affecting) professions

[2]In Software Engineering, a pattern is a reusable solution to a recurrent problem.

and activities across domains (Tiron-Tudor and Deliu 2022). The impact of AI tools over time, in particular on the development of children, is a necessary but overlooked aspect (Gennari et al. 2023). Longitudinal studies, relative analyses and qualitative methods (Samman 2024) coupled with standardised instruments that help evaluate contextual factors (e.g. learning and engagement) (Gennari et al. 2023) could help to unpack the complex effects of AI in sectors such as education. Moreover, further investigation is needed into the ethical and responsible trade-offs when developing and deploying AI. Inherent conflicts between different components within RAI and related principles are a reality that is also often overlooked (Singhal et al. 2024). For instance, by attending to the needs of some stakeholders, others' needs may get neglected or undermined. More exploration is needed into how to mitigate these frictions in real-world scenarios.

Formal training on AI skills and the ethical and social impacts of AI. Seven articles urge for providing the necessary AI skills and awareness of the social and ethical complexities that AI introduces across domains and society as a whole. An ongoing and reflective AI literacy in relation to the context and evolution of AI should be sought (Gennari et al. 2023), through comprehensive training and support in the proper use of AI tools and platforms (Samman 2024). Innovative and playful tools for teaching AI are already being proposed; future work in this area should consider targeting even younger groups (e.g. pre-school children) (Gennari et al. 2023). Academic training for future professionals is also crucial (Mörch et al. 2021), for which efforts to update the educational curriculums are currently undergoing (Tiron-Tudor and Deliu 2022), including equipping them with critical thinking, problem-solving and communication skills (Moloi et al. 2023). Moreover, workers should continuously acquire new skills through further upskilling and retraining (Wach et al. 2023). Last but not least, a better integration of ethics and humanities in AI, fostering multidisciplinary and improving diversity (e.g. gender balance) across disciplines is also considered necessary to face the challenges of the AI landscape (Foffano, Scantamburlo, and Cortés 2023).

Domain-specific standards and professional codes. It is unclear how prepared for AI integration various domains are. What becomes more evident as AI begins to permeate society, is that professions are reconfigured, with the creation and shifting of roles, responsibilities, tasks, algorithms, tools, competencies and distribution of knowledge and expertise (Tiron-Tudor and Deliu 2022; Samman 2024). Therefore, new professional standards, regulations and codes of ethics should be collaboratively created, in order to promote the responsible use of AI for each specific domain (Mörch et al. 2021; Tiron-Tudor and Deliu 2022; Wach et al. 2023; Moloi et al. 2023). In turn, governance structures must be established at different levels (e.g. developers, communities, institutions, industries, etc.) (Wach et al. 2023), as well as non-supervisory structures (Samman 2024) to help regulate the AI market (Wach et al. 2023).

Discussion

Unclear and Varied RAI Terminology

The definitions and terms associated with RAI vary, and several articles included in the review do not provide a definition of RAI. This finding is consistent with that past research in which, of 254 papers examined, only five specifically addressed the definition of RAI (Göllner, Tropmann-Frick, and Brumen 2024). As highlighted in our findings, RAI is also frequently conflated or used interchangeably with terms like Ethical AI and Trustworthy AI (Göllner, Tropmann-Frick, and Brumen 2024). This reflects the broad and multi-faceted nature of responsibility in the context of advanced technologies, a characteristic shared with the expansive scope of RRI, where concepts like fairness, accountability, transparency, and ethics (FATE) are central (Memarian and Doleck 2023).

Interestingly, we note that computational techniques to achieve goals related to RAI, such as fairness, explanability and transparency (Memarian and Doleck 2023) were not present in the reviewed papers. This could be due to RAI in the literature being framed as a governance framework for AI rather than specific techniques that help with making the systems themselves amenable to be deployed and used more responsibly. The focus on governance and fairness may overlook AI's unforeseen and unknown implications of AI (Tahaei et al. 2023). Another potential reason is that our search keywords were limited to 'AI' and 'Artificial Intelligence'. The team thoroughly discussed the search queary in early stages of the umbrella review. Having the option to include other relevant keyworkds such as 'machine learning', 'deep learning', 'large language models' and more, we opted for first investigating RAI as a general term without specifying any particular approaches. Of course there are many computational techniques that align with the described principles and purposes of RAI, but do not mention the words 'AI' or 'Artificial Intelligence' at all. This reflects what others have pointed out regarding how these terms have regained popularity in the recent years and the politics behind this shift (Crawford 2021; Stokel-Walker 2024).

Lastly, there are fundamental questions regarding the underlying understanding of goals such as 'ensuring AI systems are responsible throughout their lifecycle'. Particularly, do these efforts aim to *embed* responsibility within the systems, *delegating* accountabilty to them, or design and deploy them *responsibly*? Even if the design and development address ethical concerns and issues such as bias towards certain populations (e.g. racial bias), to what extent can the AI systems themselves be considered 'responsible'? As others have noted, a significant shift is needed towards '*seeing responsibility as relational over seeing responsibility as an agent property*" (Vallor 2023). We note there are two layers in which RAI is understood: firstly, by considering the RAI principles themselves, and secondly, by implementing or adhering to them 'responsibly' (i.e. by following established procedures or recommended practices). Given the autonomous capabilities of AI systems, it could be more useful and clearer to shift from terminology that could be misleading (especially to the general public) such as 'responsible AI' to expressions like 'AI made responsibly or following

responsible practices'.

The RAI Gap: Aspirations vs Implementation

The literature appears to indicate that we are currently at a transition point where we clearly see the need for more responsible approaches to AI, have developed frameworks for this purpose, but have yet to generate the evidence needed to clearly show that RAI can bring about the changes that the literature suggests it will. An interesting paradox exists between the proliferation of numerous ethical and RAI frameworks and the need for more actionable guidelines. However, a follow-up matter to addres is what constitutes sufficiently actionable steps. At the moment, we are relying on 'I'll know it when I see it' rather than a specific definition which impacts on the evidence that we are able to collect in support of RAI.

Some concepts within the literature point towards ways that we might understand how RAI is better than just AI but the lack of clarity around purpose and accountability mean there is limited motivation amongst certain stakeholders and in certain domains. It is telling that so much of the discussion around RAI is driven by the arrival AI into new sectors and disciplines, the ambiguity of its impact, and the concerns that arise. RRI principles highlight the need for involvement of diverse stakeholders and active collaboration but it is clear from the literature we include within this review that the focus is less on conceptual understanding and more on development and deployment. To better understand the potential impact of more responsible approaches in AI, we suggest the need for consensus on not only definitions but how these concepts change across the different lifestages of the technology. We also recommend that more evidence is generated to clearly show how more responsible approaches bring benefits and help mitigate risks.

The concept of RAI could appear aspirational; that is, conceptually appealing but difficult to achieve in practice. Some argue that RAI remains highly abstract (Baldassarre et al. 2024; Xia et al. 2023; Lu et al. 2024), and whilst there are concerns about its operationalisation, a key challenge is how to effectively assess it and prevent it from becoming a mere checklist exercise (Dignum 2019). Despite the ongoing efforts and clear good intentions demonstrated by various stakeholders, the rapid evolution of AI and its consequences (e.g. job displacement, copyright concerns, gender, racial and cultural biases), whether intended or not, have led to the term RAI becoming somewhat of a buzzword (Baeza-Yates 2023). It is essential to establish mechanisms that prevent superficial adoption of RAI principles, incorporating a mix of incentives and penalties to support its implementation. It would appear as though we do not learn from the past (i.e. social media and the ineffective Online Safety Act). We need, as others have pointed out, systems that are thoroughly tested *before* they are deployed to the public or real-world scenarios (Liu et al. 2024). But to what extent is this feasible given the ongoing AI race? Is the reflective nature of RRI in direct conflict with the rapid pace of AI development? Furthermore, the current political and economic landscape will pose obstacles to RAI adoption, for instance in light of ongoing reductions in research funding within the US academic sector affecting a range of important areas relevant to RAI such as diversity, marginalised populations, and bias, to name only a few.

RAI Awareness, Education and Training

Our results reflect those of (Kiemde and Kora 2022) who state that education has a prominent role to play in developing Responsible AI through the democratization of technology. The authors suggest that by introducing ethics courses into academic training, developing the skills of AI developers, and conducting research on RAI, education will promote the integration of ethical values and support the development of RAI through the diversification of AI teams.

The current state of RAI literature indicates that there is a clear demand for practical and actionable guidelines that are sufficiently detailed without being overly high-level. This suggests that there may be a gap in RAI skills that require more than technical proficiency and cover a broader understanding of AI's societal implications and ethical dimensions. An approach that merely asks for specific instructions on what to build may not be ideal. Formal training in AI literacy must extend to skills such as ethical reasoning, critical thinking and social awareness. These competencies must enable professionals not only to build AI systems but to question their appropriateness, potential harm, and long-term impact. Developing these sensitivities could help to ascertain that RAI principles may be intentionally broad or open-ended to reflect the reality that questions surrounding ethics in AI development and deployment often do not have a definitive or universally correct answers, especially as some RAI principles aim to attend to vulnerability, which could take different forms (Vallor 2023).

Most guidelines and principles that have emerged in the last couple of years, from governments, policy organisations, social agencies, or tech companies, largely lack concrete proposals for education, even though most recognise that education will play a crucial role in the future of AI (Dignum 2021). Education and training efforts must be aligned with domain-specific standards and professional codes of conduct, many of which are yet to be developed and iterated as the adoption of AI increasingly disrupts and transforms domains.

Future work must continue exploring and documenting direct and indirect use of AI to help elucidate on what 'using AI responsibly' means. Empirical evidence about AI's effects on society is urgently needed to ground these debates in reality and ensure that education and policy are guided by observed outcomes rather than marketing narratives or dystopian fears (Brown et al. 2025). Likewise, AI literacy must enable the general public to understand AI harms and impacts affecting their own everyday lives.

Recommendations and Limitations

This umbrella review aimed to explore the various ways in which prior academic literature has referred to RAI. Figure 2 presents an initial taxonomy of the different understandings of RAI as found and discussed in this article. Given the evidence that RAI can have different connotations, we recommend not assuming a shared understanding of what RAI

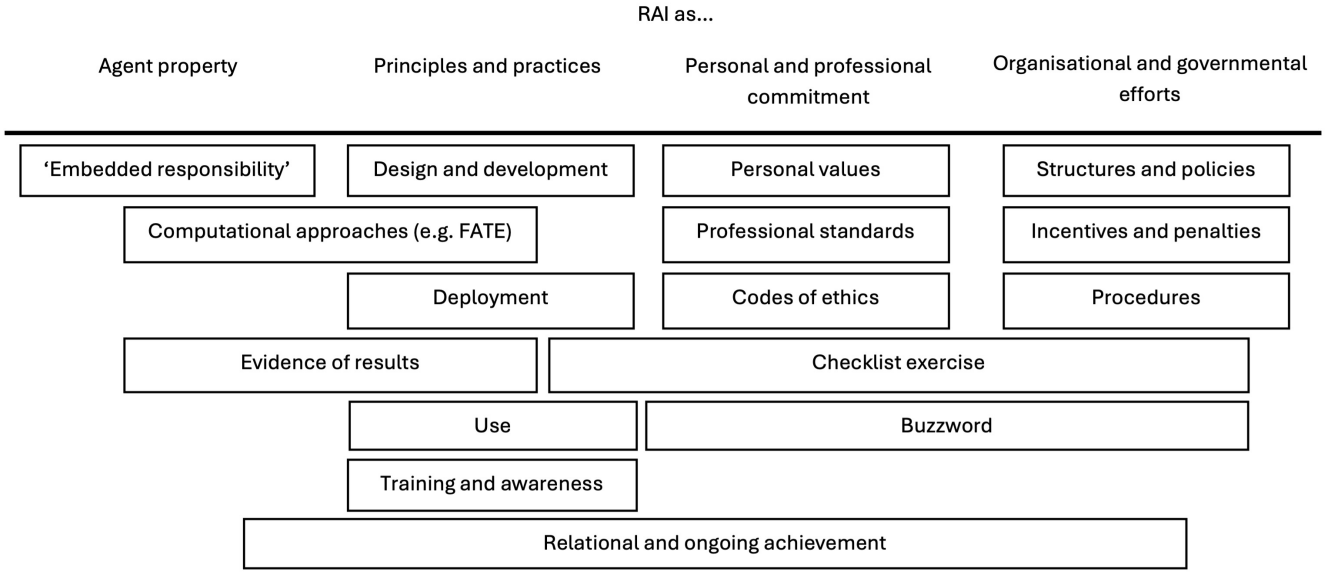

Figure 2: A taxonomy of Responsible AI

comprises or entails. The proposed RAI taxonomy includes four main facets in which RAI is understood: as an agent property, as principles and practices, as personal and profesional commitment, and as organisational and governmental efforts. This of course is merely an initial (and incomplete) attempt to unravel the complexities and implications of RAI.

When possible, we recommend to provide a definition or an intended meaning for RAI, as well as for the terms responsible, and responsibility. However, beyond advocating for consensus or providing a new unified definition, we aim to shed light on the range of perspectives that conform RAI, especially as different stakeholders prioritise RAI values differently (Jakesch et al. 2022). RAI and responsibility are nuanced and cannot be considered binary attributes. We advocate for moving away from an ultimate goal in which 'AI is responsible' (agent property) towards approaching RAI as a relational and ongoing achievement which is comprised by the principles, practices, comittments and efforts at individual, organisational and national/international levels (Deshpande and Sharp 2022). After all, the work of RAI is never over. Even after measures are implemented and potential risks are identified, harms may still occur due to the evolving nature of AI and its face-paced landscape. Moreover, RAI practitioners are advocates, but they need organisational support (Rismani and Moon 2023).

Other relevant work also discusses the concepts of RAI; our research should be seen as complementary as it only covers systematic reviews. Moreover, this review considered articles published up to May 2024, when the search was conducted, and therefore reviews related to RAI published after that point were not considered in these findings.

Conclusion

An umbrella review (i.e. a review of reviews) of the academic literature between 2013 and early 2024 related to RAI was conducted. We found that RAI is motivated by the fast paced advancement of AI and subsequent disruption created across domains such as healthcare, education, policymaking, and professional practice. Despite the potential benefits of AI, there continue to be risks and concerns surrounding it. This in turn calls for computational and human measures and principles that span from risk mitigation to ensuring beneficence and addressing social problems. However, the term RAI is irremediably conflated and used interchangeably with other terms such as ethical and trustworthy AI. Likewise, 'responsible' and 'responsibility' in the context of AI are used with related but ultimately different connotations (i.e. accountability, social responsibility, an umbrella term for a set of principles or guidance, acting with integrity, and the personal, professional or external motivations and incentives for those actions). The current challenges and future directions of RAI include devising a way around or through the proliferation of tools, guidelines and frameworks that are meant to guide the development, deployment and use of AI. This may entail not only devising more specific and tailored approaches, but promoting the sensitivity to ascertain that RAI principles are sometimes intentionally broad to acknowledge that ethical questions in AI lack definitive or universal answers. Likewise, collaborative efforts should emphasise more empirical evidence of the impacts of AI in society, in particular over time, further formal training on AI literacy (skills, ethical and social impacts), and the creation of standards and professional codes across specific domains. Ultimately, it is essential to establish mechanisms that prevent superficial adoption of RAI principles, incorporating a mix of incentives and penalties to support its implementation. If AI-related harms persist without accountability or consequences, despite harms being well-documented, AI will continue to be designed, developed, deployed and used irresponsibly.

Positionality Statement

We are a team of academics and researchers at various career stages from various countries (in Europe, the Americas and Africa) affiliated with a university in the UK, and as such we have considerable collective experience working with Responsible Research and Innovation (RRI) principles, which are highly encouraged and sometimes required by our national research funding councils. We appreciate that RRI may be a less familiar concept for some international audiences, especially outside Europe. Therefore, it is important to highlight that the local conditions and research environment in our specific geographical context have influenced the development of this paper, from the formulation of guiding questions, to the analysis and presentations of results. Our views, however, comprise a mix of positive, hopeful, ambivalent, sceptical and cynical perceptions and feelings towards AI, RAI and RRI, which are also reflected in the contents of this paper.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Online; accessed 7 May 2025.

Baeza-Yates, R. 2023. Lecture Held At The Academia Europaea Building Bridges Conference 2022: An Introduction to Responsible AI. *European Review*, 31(4): 406–421.

Baldassarre, M. T.; Caivano, D.; Fernández Nieto, B.; Gigante, D.; and Ragone, A. 2024. Fostering Human Rights in Responsible AI: A Systematic Review for Best Practices in Industry. *IEEE Transactions on Artificial Intelligence*, 6(2): 416–431.

Barletta, V. S.; Caivano, D.; Gigante, D.; and Ragone, A. 2023. A Rapid Review of Responsible AI frameworks: How to guide the development of ethical AI. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, EASE '23, 358–367. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700446.

Braun, V.; and Clarke, V. 2021. *Thematic analysis: A practical guide*. SAGE publications Ltd.

Brown, V.; Larasati, R.; Third, A.; and Farrell, T. 2025. *A Qualitative Study on Cultural Hegemony and the Impacts of AI*, 226–238. AAAI Press.

Cornwall, A.; and Jewkes, R. 1995. What is participatory research? *Social Science & Medicine*, 41(12): 1667–1676.

Crawford, K. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Deshpande, A.; and Sharp, H. 2022. Responsible AI Systems: Who are the Stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 227–236. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.

Dignum, V. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 2156. Springer.

Dignum, V. 2021. The role and challenges of education for responsible AI. *London Review of Education*, 19(1): 1–11.

EC. 2014. Horizon 2020. https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en. Online; accessed 19 May 2025.

European Commission and Directorate-General for Communication and Directorate-General for Research and Innovation. 2013. *Responsible research and innovation (RRI), science and technology – Report*. Publications Office.

Foffano, F.; Scantamburlo, T.; and Cortés, A. 2023. Investing in AI for social good: an analysis of European national strategies. *AI & society*, 38(2): 479–500.

Gennari, R.; Melonio, A.; Pellegrino, M. A.; and D'Angelo, M. 2023. How to Playfully Teach AI to Young Learners: a Systematic Literature Review. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter*, CHItaly '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400708060.

Göllner, S.; Tropmann-Frick, M.; and Brumen, B. 2024. Towards a definition of a responsible artificial intelligence. In *Information modelling and knowledge bases XXXV*, 40–56. IOS Press.

Jakesch, M.; Buçinca, Z.; Amershi, S.; and Olteanu, A. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 310–323. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Jirotka, M.; Grimpe, B.; Stahl, B.; Eden, G.; and Hartswood, M. 2017. Responsible research and innovation in the digital age. *Commun. ACM*, 60(5): 62–68.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9): 389–399.

Jung, G. 2018. *Our AI overlord: The cultural persistence of Isaac Asimov's three laws of robotics in understanding artificial intelligence*. Master's thesis, University of California.

Kiemde, S. M. A.; and Kora, A. D. 2022. Towards an ethics of AI in Africa: rule of education. *AI and Ethics*, 2(1): 35–40.

Liu, E. J.; So, W.; Hosoi, P.; and D'Ignazio, C. 2024. Racial Steering by Large Language Models: A Prospective Audit of GPT-4 on Housing Recommendations. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712227.

Lu, Q.; Zhu, L.; Xu, X.; Whittle, J.; Zowghi, D.; and Jacquet, A. 2024. Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *ACM Comput. Surv.*, 56(7).

Memarian, B.; and Doleck, T. 2023. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence*, 5: 100152.

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, 1(11): 501–507.

Moloi, K.; Maladzhi, R. W.; Nemavhola, F. J.; Mthombeni, N. H.; Tsoeu, M. S.; and Mashifana, T. 2023. The Risks Associated with AI Chatbots in Teaching Future Engineering Graduates: A Systematic Review. In *2023 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*, 1–6.

Mörch, C.; Atsu, S.; Cai, W.; Li, X.; Madathil, S.; Liu, X.; Mai, V.; Tamimi, F.; Dilhac, M.; and Ducret, M. 2021. Artificial intelligence and ethics in dentistry: a scoping review. *Journal of dental research*, 100(13): 1452–1460.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Owen, R.; Stilgoe, J.; Macnaghten, P.; Gorman, M.; Fisher, E.; and Guston, D. 2013. *A Framework for Responsible Innovation*, chapter 2, 27–50. John Wiley & Sons, Ltd. ISBN 9781118551424.

Portillo, V.; Craigon, P.; Dowthwaite, L.; Greenhalgh, C.; and Pérez-Vallejos, E. 2022. Supporting responsible research and innovation within a university-based digital research programme: Reflections from the "hoRRIzon" project. *Journal of Responsible Technology*, 12: 100045.

Portillo, V.; Greenhalgh, C.; Craigon, P. J.; and Ten Holter, C. 2023. Responsible Research and Innovation (RRI) Prompts and Practice Cards: a Tool to Support Responsible Practice. In *Proceedings of the First International Symposium on Trustworthy Autonomous Systems*, TAS '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400707346.

Rismani, S.; and Moon, A. 2023. What does it mean to be a responsible AI practitioner: An ontology of roles and skills. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 584–595. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.

Samman, A. M. A. 2024. Harnessing Potential: Meta-Analysis of AI Integration in Higher Education. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, 1–7.

Singhal, A.; Neveditsin, N.; Tanveer, H.; and Mago, V. 2024. Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review. *JMIR Med Inform*, 12: e50048.

Stahl, B. C.; and Eke, D. 2024. The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74: 102700.

Stahl, B. C.; Portillo, V.; Wagner, H.; Craigon, P. J.; Darzentas, D.; Garcia, S. D. O.; Dowthwaite, L.; Greenhalgh, C.; Middleton, S. E.; Nichele, E.; Wagner, C.; and and, H. W. 2024. Implementing responsible innovation: the role of the meso-level(s) between project and organisation. *Journal of Responsible Innovation*, 11(1): 2370934.

Stefanidi, E.; Bentvelzen, M.; Woźniak, P. W.; Kosch, T.; Woźniak, M. P.; Mildner, T.; Schneegass, S.; Müller, H.; and Niess, J. 2023. Literature Reviews in HCI: A Review of Reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Stokel-Walker, C. 2024. *How AI Ate the World: A Brief History of Artificial Intelligence–and Its Long Future*. Canbury Press.

Tahaei, M.; Constantinides, M.; Quercia, D.; and Muller, M. 2023. A Systematic Literature Review of Human-Centered, Ethical, and Responsible AI. arXiv:2302.05284.

Tiron-Tudor, A.; and Deliu, D. 2022. Reflections on the human-algorithm complex duality perspectives in the auditing process. *Qualitative Research in Accounting & Management*, 19(3): 255–285.

Ulnicane, I.; Eke, D. O.; Knight, W.; Ogoh, G.; and Stahl, B. C. 2021. Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies. *Interdisciplinary Science Reviews*, 46(1-2): 71–93.

Vallor, S. 2023. Edinburgh Declaration on Responsibility for Responsible AI.

Von Schomberg, R. 2013. A vision of responsible research and innovation. *Responsible innovation: Managing the responsible emergence of science and innovation in society*, 51–74.

Wach, K.; Duong, C. D.; Ejdys, J.; Kazlauskaitė, R.; Korzynski, P.; Mazurek, G.; Paliszkiewicz, J.; and Ziemba, E. 2023. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2): 7–30.

Wakunuma, K.; Ogoh, G.; Akintoye, S.; and Eke, D. O. 2025. *Decoloniality as an Essential Trustworthy AI Requirement*, 255–276. Cham: Springer Nature Switzerland. ISBN 978-3-031-75674-0.

Xia, B.; Lu, Q.; Perera, H.; Zhu, L.; Xing, Z.; Liu, Y.; and Whittle, J. 2023. Towards Concrete and Connected AI Risk Assessment (C2AIRA): A Systematic Mapping Study. In *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)*, 104–116.