BRAID
Bridging Responsible AI Divides

The Responsible AI Ecosystem:
A BRAID Landscape Study

Fabio Tollon and Shannon Vallor - June 2025

BRAID is a UK-wide programme dedicated to integrating Arts and Humanities research more fully into the Responsible AI ecosystem, as well as bridging the divides between academic, industry, policy and regulatory work on responsible AI.

Funded by the Arts and Humanities Research Council (AHRC) it represents a six-year, £15.9 million investment in enabling responsible AI in the UK. The Programme runs from 2022 to 2028.

Working in partnership with the Ada Lovelace Institute and BBC, the team brings together expertise in human-computer interaction, moral philosophy, arts, design, law, social sciences, journalism, and AI.

BRAID is extended by a network of interdisciplinary researchers and partnering organisations through the delivery of funding calls, community building events, and a series of programmed activities.

Funding reference: Arts and Humanities Research Council grant number: AH/X007146/1.

Learn more at **www.braiduk.org**

To request an alternative format of this report please email **braid@ed.ac.uk**

Abstract

This report, the first in a two-part study of the Responsible AI landscape by the UKRI's BRAID programme, provides a chronological and conceptual map of the Responsible AI (R-AI) ecosystem. We chart the role of various actors and communities in this ecosystem's emergence, especially the vital contributions from the arts and humanities, and the historical development and contestation of different meanings of 'responsibility' in the context of AI.

In the Preface, we outline the motivation and aims of the study. In Section One, we outline the different meanings and conceptions of 'Responsible AI' commonly deployed by different stakeholders. We follow in Section Two with a chronological account of how 'responsibility' for the impact of new technologies in computing came to be articulated, culminating in what today we call the Responsible AI ecosystem. We trace this history from the 1950s to the present, concluding in Section Three with a review of seven lessons learned from these 'first waves' of Responsible AI research, practice, and advocacy; lessons that can be carried forward in our collective efforts to enable and sustain responsible AI innovation, now and for the future.

Contents

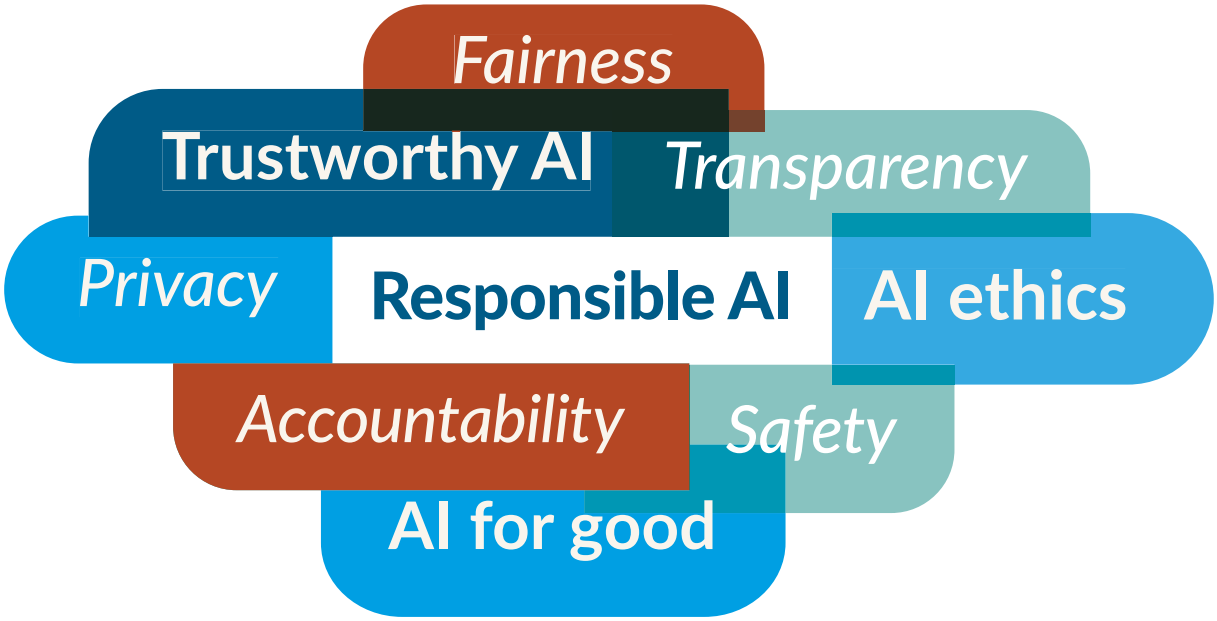
PREFACE

The UK's current AI Opportunities Action Plan, published in March 2024, proposes that the UK should "build an artificial intelligence sector that can scale and be competitive globally," while driving "growth and productivity" and "transforming citizens' experiences of interacting with the state," through increased adoption of AI "in all parts of the public sector and the wider economy." Yet none of these ambitions will come to fruition if the underlying technology, and people's experiences with it, do not reflect an equitable, transparent, safe, socially beneficial and accountable approach to AI development, deployment and adoption – that is, an approach to AI that is *responsible*.

The BRAID research programme, based at the University of Edinburgh and funded by UKRI's Arts and Humanities Research Council, was created in response to a 2022 funding call titled 'Enabling a Responsible AI Ecosystem.' BRAID stands for Bridging Responsible AI Divides, because an ecosystem cannot be divided. An ecosystem is defined and sustained by flows and connections between interdependent local ecologies and communities; there is no part of an ecosystem that is fully isolated from the rest. Yet these flows and interdependencies within the AI landscape are not well mapped or widely understood. They are also too often blocked or inhibited by the many barriers that still divide the disciplines, geographies, institutions, sectors, agencies, professions and community voices and interests that make up the AI ecosystem. Nor is it clear exactly what it would mean for the AI *ecosystem* to earn the label of 'responsible.'

> *... an ecosystem that supports the healthy development and flourishing of all the lives and communities that AI touches and shapes."*

Among the core outputs we proposed at the start of the BRAID programme, which is now in its third year, was a study that would provide a historical and conceptual mapping of the Responsible AI landscape, with an eye to better understanding the divides that need to be bridged in order to enable a healthy and sustainable Responsible AI ecosystem. That does not mean an ecosystem that supports the healthy development and flourishing of *artificial intelligence*. It means an ecosystem that supports the healthy development and flourishing of all the lives and communities that AI touches and shapes. Considering AI's energy demands alone, that will soon encompass nearly every living creature and community on the planet.



'Responsible AI' is not one thing, but a set of interweaving methods and approaches.

The scope of the study that follows, then, is by necessity restricted. While a fine -grained map of the landscape on which such an AI ecosystem could emerge would be impossible to produce in a few short years, we can benefit from a clearer view of its outlines and major topographical features, and the epochs through which it has moved. More narrowly, we are interested in the features of this landscape that pertain to its potential to host a *responsible* configuration of the AI ecosystem. What does that word, 'responsible,' mean in this context?

The label 'Responsible AI', as detailed in the study below, emerged in 2017 as a reassuring-sounding banner for programmes of research being carried out by AI companies and large consulting firms to make AI products more compliant with social expectations of fairness, transparency, accountability, safety and privacy. Yet its meaning soon became interwoven with other, older labels created by academics, policymakers and non-profits interested in related goals – 'AI Ethics,' 'Trustworthy AI,' 'AI for Good'. Today 'Responsible AI' signifies much more than an industry research agenda; it describes a sprawling, intertwined, contested and negotiated set of social expectations for what AI can and *should* be.

Unraveling and sorting out this growing tangle of meanings and expectations, and clarifying the relationships and divides among the communities that produce and shape them, is the first step toward making sense of today's Responsible AI landscape, and grasping how it might

“... a wiser choice is to first pause and look back – to collect and share the most important lessons one learned on the first rocky phase of the journey, and see how they might make our next steps more secure and fruitful.”

one day come to support a healthy AI ecosystem, in which AI enables human and planetary flourishing. This requires an *interdisciplinary* approach, one that draws heavily from the arts and humanities, as well as the social sciences. These humanistic fields enrich and broaden narrowly technical perspectives on AI that otherwise overlook the constituent social systems, values, communities, incentives and choices that give the AI ecosystem its shape and direction.

Our landscape study begins by looking through two lenses: *a conceptual lens* to understand the different ways that stakeholders in the AI ecosystem have understood the words ‘responsible’ and ‘responsibility’ in the context of AI, and a *historical lens* to see how, when and where these concepts first emerged in connection with AI, how they have changed and been contested over time, and how these changes have given rise to fruitful collaboration and consensus among stakeholders on social expectations for AI, as well as deep divides and disagreements.

This first phase of the study concludes with seven vital lessons that can be identified from the ‘first waves’ of Responsible AI research and practice from 2017 onward. These waves grew fairly steadily with AI developments until the commercial arrival of AI tools based on large language models in 2022. That transition was a seismic event in the history of Responsible AI, the reverberations and repercussions of which are still being felt. It has opened and widened new fractures within communities in the Responsible AI landscape, while diminishing the commercial incentives for Responsible AI development, use and governance. As a result, several AI companies and governments have recently walked back their prior commitments to safe and responsible AI; quixotically, this comes at the very moment when these actors urgently need to secure public trust and confidence in AI in order to encourage its growth and adoption.

This shift presents then, a critical opportunity – a moment of great instability in the Responsible AI landscape, but also mounting social pressure to move forward with the changes needed to make the AI ecosystem in the UK – and globally – more sustainable, permissible, desirable, *responsible*. At such moments of urgency and instability it can be tempting to just leap forward. But a wiser choice is to first pause and look *back* – to collect and share the most important lessons one learned on the first rocky phase of the journey, and see how they might make our next steps more secure and fruitful.

SECTION ONE

Mapping the Responsible AI Ecosystem

Introduction

What does it mean to say that a company, institution, or individual is 'responsible'? On the face of it, it seems that such a claim is normative. 'Responsible' means some tendency to behave in a socially desirable or acceptable way (Tigard, 2021, p. 114). This suggests that the label comes with a certain positively value laden endorsement: To be 'responsible' in this way is good, and something to be encouraged. In the context of assessing an individual's character, claiming that a person is 'responsible' suggests a number of positive traits: That they can be relied upon, that they have behaved well in the past, that we can expect them to behave well in the future, etc.

It also seems legitimate to ascribe this sense of responsibility to collective agents such as corporations or states. These entities can (and should!) behave responsibly, and calling them 'responsible' reflects them having done so and their commitment do so in the future. This sense of responsibility has also come to be applied in the context of research. To do research in a 'responsible' manner is virtuous: It might mean that the research is unbiased, fair, and respects human dignity. The label, then, matters.

In recent years the concept of responsibility has been applied to a number of contexts of innovation and technology, including artificial intelligence technologies (Dignum, 2019; Zhu, 2019; De Laat, 2021). On the surface this seems a good thing, as of course we want the development, deployment, and use of AI-systems to be in line with certain normative principles – such as ethical demands and considerations of justice – and it seems the 'responsible' frame should give us just that.

If we are not careful, we run the risk of 'ethics washing' where R-AI is used as a cover for unethical practices.

The worry, of course, is that the label, by having this (positive) normative meaning, can be used as a cover or screen for irresponsible, unethical (or even unlawful) practices, what is commonly known as 'ethics-washing.' That term arose from the earlier 'greenwashing,' used to challenge superficial and insincere corporate efforts to appear environmentally responsible (De Freitas Netto *et al.*, 2020). Corporations, for example, might boast of their own efforts in R-AI, but in practice that can mean very little (Microsoft Corporation, 2019; De Laat, 2021). Put another way, corporations may employ an artificially narrow and restrictive definition of R-AI that suits their own interests. Such restrictive perspectives on R-AI are often used as confidence-builders; they offer reassurance that AI-related harms are problems either solved or soon to be solved by 'the experts', and that cutting-edge tools are available to ably manage AI risks. They often stand in stark contrast with more inclusive or expansive perspectives that admit that there are things we don't know, problems with AI we don't yet know how to solve, and impacts we cannot fully or reliably predict (Ipsos, 2023). Such admissions usefully broaden both the scope and ambition of R-AI. However, even sincere, inclusive or expansive uses of the term lack definitional clarity and agreed-upon criteria, as we show below.

The Many Meanings of Responsible AI

R-AI "has now become a brand-like umbrella term for the development of principles, approaches and methods of understanding what responsible AI development means and how it can be implemented" (Drage, McInerney and Browne, 2024). And while there is no settled consensus as to what R-AI means or ought to mean (at least in the UK), (Ipsos, 2023) there are a number of shared challenges and themes that link many of the actors in the R-AI ecosystem, as we show in this report. Even so, different stakeholders often have different starting points and priorities in setting out the goals and motivations behind R-AI: The term has been wielded by industry to project an image of trustworthiness and safety; by government bodies and agencies to define the aims of AI policy, investment, and regulation; and by civil society organisations to demand accountability and redress for AI-driven harms.

Yet as noted in the Preface, many government and industry commitments to responsible AI development have recently been scrapped or substantially watered down in the new race

R-AI encompasses a spectrum of values and goals.

for AI market growth and geopolitical dominance. For example, Google DeepMind, Meta, Twitter/X and Microsoft have all disbanded or eliminated Responsible AI teams since the market race for Generative AI dominance began (Criddle and Murgia, 2023; Paresh, 2024). On 5 February 2025 Google scrapped its prior promise to not develop AI weapons or applications "likely to cause harm" (Hooker and Vallance 2025). That same week, the US and UK refused to sign the Paris AI Action Summit statement that called for AI to be *sustainable*, *open*, *inclusive*, *transparent*, *ethical*, *safe*, *secure and trustworthy* (Milmo and Courea, 2025). The US government has scrapped AI safety regulations (Wheeler, 2025), while the UK in February 2025 rebranded their AI Safety Institute to the AI Security Institute (McKeon, 2025), restricting its mission to exclude prior R-AI concerns like unfair bias and protection of free speech. Paradoxically, this leaves the entire AI ecosystem in a more fragile and endangered state, as the hoped-for beneficial consequences of AI adoption cannot be secured without responsible development and use.

The retreat from prior R-AI commitments to safety, sustainability, fairness in AI development and narrowing of its aims to something like 'national security' or 'economic competitiveness' stands in stark contrast with the wide scope of R-AI ambitions to ensure that AI is safe and beneficial not for just a lucky few, or a nation, but humanity and the planet as a whole. Indeed, R-AI has been touted as being concerned with ensuring that AI systems respect sustainable development goals, human rights and democratic values (OECD, 2019; Beckman, Hultin Rosenberg and Jebari, 2024). This shows just how large the *scope* of R-AI is, and by extension the corpus from which it can draw insights. It also illustrates the expansive range of stakeholders who can assert material, moral and political interests in the concept of 'Responsible AI' and its use, interests that in many cases stand in tension with one another.

There is a sharp break between these diffuse, colloquial and often contested uses of the terms 'responsible' and 'responsibility' in the AI context, and the more rigorous and focused scholarly analysis of the normative concept of responsibility that has historically been carried out within the humanities, and moral and legal philosophy in particular. While isolated elements of the latter occasionally make their way into policy dialogues or industry publications on Responsible AI, often filtered through the intermediary of applied research in 'AI ethics' or 'technology ethics,' there remains a wide gulf between scholarly and practical perspectives on what 'responsible' AI is or must be. The lack of a clear understanding of what R-AI means, and the different background assumptions of the multiple stakeholders involved in R-AI, demonstrate that there is an urgent need to bridge this gulf, to the extent that this is possible, in order to ensure that the concept can do the work for society that we need it to do.

What does that work need to achieve? One common way of answering that question is to invoke the ideal of a *responsible AI ecosystem* (UK Government, 2022; Stahl, 2023). This ecological framing of the goal of Responsible AI aims at a future state of affairs in which responsibility appropriately infuses and guides the complex interactions between the diverse and vast community of actors with a stake in AI and its societal and planetary impact.

We take the ecosystem metaphor to offer the best prospect for mirroring the complexity, diversity and dynamism of the responsible AI landscape, while supplying a coherent normative aim: namely, the social and moral health or *flourishing* of the system. As a precursor to this

aim, this landscape study of R-AI offers a systematic overview and analysis of the dimensions, relationships, and evolving interactions within this community of actors that need to be better understood and harmonised, and the divides that must be more effectively bridged, if a flourishing and responsible AI ecosystem is to be realised.

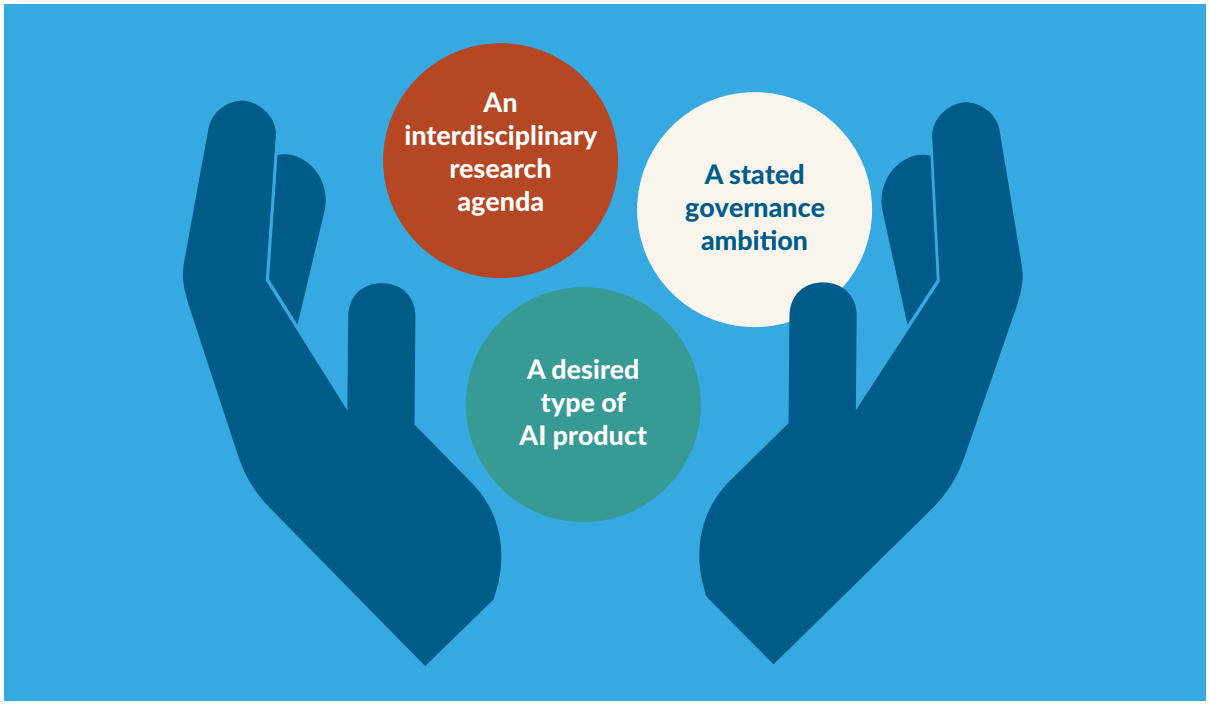
The goal of this study is to outline the contours of the existing R-AI landscape with an eye to answering some key questions that will be relevant to researchers and practitioners in Responsible AI from industry, academia, government, research institutes, and NGOs.

We will first provide a chronology of the conceptual and practical emergence of Responsible AI, which we analyse as concurrently developing and now consisting of 4 major dimensions:

1. **An interdisciplinary research agenda**
2. **A stated governance ambition**
3. **A desired type of AI product**
4. **A broad community or ecosystem of stakeholders**

Responsible AI understood as an interdisciplinary research agenda describes a *body of knowledge and skilled expertise* that we need to develop, refine and then consistently apply to the design, use and governance of AI systems to render them socially and ethically acceptable.

R-AI as a governance ambition originated from a desire (and sometimes a need) for tech companies to self-regulate. Since 2017, Meta, Google, Microsoft, IBM, PwC and Accenture have all produced internal R-AI documents which each offer a set of principles and/or core values. These Responsible AI *principles* are used to inform the development of various tools and practices, such as internal ethics reviews, risk assessments, and product testing, in the



The goals of R-AI vary, with stakeholders having distinct, and often competing objectives.

hopes that these will realise particular values within their AI business. Soon after, nations began to frame Responsible AI as a government ambition, part of their own innovation strategies. It remains to be seen how public confidence in, and affection for, AI adoption and innovation will be altered by the recent retreat by powerful states and corporate actors from their prior voluntary commitments to R-AI. Evidence suggests that public and consumer concerns about AI harms are rising, even when AI's benefits are recognised, and that publics are highly exposed to AI-generated harms (The Ada Lovelace Institute and The Alan Turing Institute, 2025). This trajectory is unsustainable if we hope to secure and distribute the benefits of AI in ways that strengthen our societies rather than diminish them.

Fortunately, not all R-AI efforts arise from corporate and government strategy. Outside of particular corporate, government, and research agendas, there is an interest in developing a single reliable system or comprehensive set of standards and techniques for ensuring that AI products and services are 'socially benign' or have responsible characteristics. This kind of work is carried out under different umbrellas by researchers, technical standards organisations, and policymakers alike, adopting labels like 'AI Standards,'[1]'AI Assurance' (Brennan *et al*., 2023), or design initiatives to promote 'AI for Good' (Taddeo and Floridi, 2018; Floridi *et al*., 2020; Umbrello and van de Poel, 2021). The common thread linking these approaches is the framing of 'Responsible AI' as a sociotechnical programme of work that cuts across organisational and national lines, and which, when executed well, can deliver more socially beneficial and trustworthy AI products or systems. Here the normative target is the *technology* rather than the developer, user or organisation. Yet even these efforts may stall in the absence of corporate or governmental cooperation. For example, in the US, it remains unclear whether deep federal funding cuts to NIST (National Institute of Standards and Technology) will end NIST's internationally touted AI Standards programme and AI Risk Management Framework (Knight, Paresh and Feiger, 2025).

The key insight from this initial survey of the landscape is that R-AI is not a singular concept or effort with a fixed meaning and clear definitional boundaries, but a complex and dynamic ecosystem pervaded by tensions and interdependencies. This correct framing is essential if we hope to enable and sustain R-AI's future, and it deeply informs this study. Research on implementing and utilising the ecosystem metaphor is widespread, and has been applied to business, innovation processes, the economy, and industrial processes (Moore, 1993; Hess, 2010; Stahl, 2022). Recently, Bernd Stahl has applied it to 'artificial intelligence' research, and shown that it has some analytical utility (Stahl, 2022). We see the ecosystem metaphor in governmental policy, such as the UK government's national AI strategy document, where reference is made to the "AI ecosystem" (UK Government, 2022). The OECD also invokes the metaphor, by recommending that governments aim at "fostering an inclusive AI-enabling ecosystem" (OECD, 2019).

---

1 For example, see: https://aistandardshub.org/

In a White Paper from 2020, the EU Commission also suggested that policy interventions be guided by building ecosystems of "trust" and "excellence" (European Commission, 2020, p. 3). More recently, the first edition of the 'Global Index on Responsible AI' identifies three "pillars of the responsible AI ecosystem" (*Global Index on Responsible AI*, 2024, p. 10) While this is of course not a conclusive argument in favour of using the ecosystem metaphor, it does suggest that doing so has been viewed favourably, and is becoming a more common approach to understanding R-AI. In a later section we will provide further justification for this approach, while noting some potential limitations.

By tracing the history of these four dimensions of R-AI, we can better map the current state of the 'ecosystem' of R-AI, answering the following questions in the process:

1. **Who** are the different communities and interests that today constitute the R-AI ecosystem?
2. **What** does 'Responsible' AI mean to these different actors and communities that make up the ecosystem, and what concept(s) of 'responsibility' anchor it?
3. **Why** do we need R-AI? What are the ends/goals/aims that a flourishing R-AI ecosystem/ community would realise?

The divides in the R-AI ecosystem require thoughtful interventions.

> *There are very real planetary costs to sustaining current levels of AI development, something that was already evident even for 'small' systems like the Amazon Echo device that Crawford and Joler analysed in their groundbreaking work 'Anatomy of an AI System'."*

The ecosystem metaphor is one we take quite seriously, and so it is worth spelling out further why we use it and what purpose it serves. Ecosystems consist of multiple (interacting) ecologies. In a similar way, R-AI is characterised by a multitude of interacting material, financial, technological, corporate, and political ecologies. As Kate Crawford and Vladan Joler documented in their widely influential use of arts and humanities methodologies for AI literacy, these ecologies cannot be understood in isolation from one another, and AI itself cannot be understood without a grasp of their complex interrelations and dependencies. There are very real planetary costs to sustaining current levels of AI development, something that was already evident even for 'small' systems like the Amazon Echo device that Crawford and Joler analysed in their groundbreaking work 'Anatomy of an AI System' (2018).

That work now stands as part of the permanent collection on display at New York City's Museum of Modern Art (MOMA) and Britain's V&A, but it is not *only* a work of art/design; it is also routinely used in classrooms and other spaces to convey important truths about the AI ecosystem. For it reveals, in a way that no bare text could, how an AI technology is not a discrete pile of code but an increasingly sprawling reality of planetary reach, interwoven with and dependent upon the multi-layered material ecology of a software device, the vast global ecology of human labour that mines, manufactures, assembles and distributes the device, as well as the technical ecology of infrastructure and expertise required to design and operate it, and the political ecology needed to govern it, even the environmental cost of disposing of it at the end of its useful life.

What Crawford and Joler's analysis revealed, and what continues to inform the ecological perspective on AI, is the striking interconnectedness of the raw material and (often invisible) human labour required to generate AI-products, and the power systems that underpin and sustain this development. From this critical intervention leveraging the combined power of the arts and humanities, and from the ecological perspective on AI that it has helped to inspire, we also gain greater insight into the challenges we must collectively meet in order to enable and sustain the entire system's health: Healthy and thriving ecosystems are balanced by the health of their composite ecologies. They develop mechanisms of resistance and resilience, they support symbiotic relations, and they are regulated by constant adjustment to the constraints

of a dynamic environment. What might a healthy, thriving, and *responsible* AI ecosystem look like? More specifically, what work does the concept of 'Responsible AI' actually do to help us move toward that condition of ecological health?

Before we can answer that question, we must better understand this ecosystem's current state, how the language of 'Responsible AI' has shaped it, and what divides or barriers stand in the way of the system's healthy ecological functioning. To do that, we need to get a handle on the different communities or 'sub-ecologies' that co-constitute a 'responsible' AI ecosystem. Following influential work on 'innovation ecosystems', we focus on the interactions between academia, industry, civil society, the natural environment, and the state, and the ways they are involved in R-AI research and practice (Carayannis *et al.*, 2021).

Understanding the current state of an ecosystem can also be advanced by grasping the material and other conditions of its early emergence and development. The initial phase of our study, then, aims to give some historical context to the emergence and development of the R-AI ecosystem, so that we may better understand its current condition. This will aid us not only in getting clear on the different meanings and uses of the R-AI concept to shape the AI ecosystem, but also in enabling its more effective use in the context of AI research, policy development, and governance.

The significant contribution of the arts and humanities to R-AI research and practice.

Before proceeding, however, there is an important caveat that we need to point out. Our study aims to tell a conceptual and historical story of Responsible AI, with the goal of showing how the arts and humanities have informed, shaped, and anchored the concept and remain vital to understanding it today. However, the role of the arts in the R-AI ecosystem has itself been marginalised. While our story highlights important, ground-breaking work that has bridged this divide and made a direct impact on R-AI practice, such as that of Joy Buolamwini, Kate Crawford, Vladan Joler, Trevor Paglen, and Mimi Onuoha, many other artists, writers, designers and performers who have engaged the social, political and ethical dimensions of advanced computing have been comparatively neglected by R-AI researchers and practitioners. Often these divides have only been bridged by the uptake of interdisciplinary perspectives and methods in R-AI. While it is beyond our scope to excavate the decades of work in the arts that have explored digital and computing innovation's impact on society, this lacuna in our landscape study underscores the need for BRAID's wider ambition: to at last fully integrate the arts and humanities in the Responsible AI ecosystem.

R-AI has been a concept 'at work' for many years already, and from an overview of its history we will be able to draw some important lessons for the future. Indeed, given that we are at what many today consider to be an 'inflection point' in AI development, it is an ideal time to take stock of what the 'first waves' of R-AI research and practice have shown us. This is a critical opportunity to show how that knowledge might be consolidated, translated and applied more effectively to ensure the health and sustainability of the AI ecosystem, nationally and globally.

We therefore aim for this study to benefit not only those interested in the conceptual and historical foundations of R-AI, but those developing current policy and practice around R-AI. In the second phase of this landscape study to begin in 2025, we will hear directly from those policymakers and practitioners how they are seeking to clear a path to a Responsible AI ecosystem in the UK and beyond.

A Complex R-AI Ecosystem

The R-AI ecosystem as we know it today has a rich and varied history, and those currently engaged in R-AI research and practice come from diverse fields such as philosophy, sociology, law, computer science, engineering, robotics, and the arts (to name but a few). Due to the multi-disciplinary nature of R-AI (both its history and current form) getting a handle on this 'history' is itself a difficult task.

The history of computer ethics, for example, seems essential to understanding R-AI today, but by itself is too broad. Additionally, the narrow focus on AI, versus broader concepts such as human-computer interaction or even 'information technology,' is a relatively new phenomenon. More than that, there is no one group of practitioners that we can call 'the' Responsible AI community. Instead, what we observe is that the R-AI ecosystem consists of many overlapping and intersecting communities, with diverse, contested, and evolving bodies of practice.

In order to understand the evolving relationships and interactions among these communities, it is important we understand where they come from. This historical view aids us in understanding R-AI as more than a concept, but as a material and social ecology of *practice*

influenced by the diversity of its elements. As noted earlier, R-AI has developed not just within diverse academic communities, but also industry, the public sector, and civil society (and this is by no means an exhaustive list).

Additionally, these different perspectives bring with them different targets and priorities for R-AI, such as fundamental research, risk management, or governance. As will become clear throughout, these different focus points result in what can at times seem like an uncoordinated struggle to get at the 'heart' of R-AI (whatever that might mean). Our aim, then, is to get at the 'hard core' of R-AI, even if it turns out that that core is dynamic, flexible, and interactive.

Not only is R-AI the result of complex and evolving interactions across different communities, disciplines and sectors, but the meaning of 'responsible' is itself contested and historically contingent, both in moral and legal philosophy and in colloquial use. What is meant by 'responsibility' in relation to technology has also broadened, from delimited professional obligations to more expansive obligations to society and affected publics.

The late 20th century philosopher Hans Jonas suggested that the increasing scale and power of human action requires a proportionate increase in our responsibilities, as well as a qualitative and profound shift in its meaning as we become responsible for the natural world, future generations and even the future of nonhuman life (Jonas, 1984, p. 1). In addition to this, we find a trend in technical communities where the role of the scientist and engineer has been extended beyond those responsibilities traditionally associated with their professional roles to include considerations of the social and moral impact of their work (Frodeman and Mitcham, 2000; Douglas, 2003), a trend greatly amplified by the power of AI technologies (Vallor, 2024).

Thus, to clarify the contemporary meaning and practical import of 'Responsible AI,' and the prospects for a flourishing and sustainable R-AI ecosystem, it will help us to first get a more solid grip on how it emerged and arrived at its current state, one still fragmented and torn by many conflicting incentives, interests and ambitions. A second stage of the analysis must then identify the conditions that must be met if we aim for the R-AI ecosystem to reach a state of healthy and sustainable equilibrium. A helpful starting point for the first analysis is to consider how the concept of responsibility first begins to shape our understanding of new technologies.

Philosophical Perspectives on Technology and Responsibility

Reflection on responsibility in relation to technology is not new, but this reflection was of course not always motivated by a particular interest in AI. The philosopher Hans Jonas, for example, argued for a thickening of our sense of responsibility in light of technological developments such as nuclear warheads (Jonas, 1973). At a similar time David Collingridge was writing about how unpredictable impacts of new technologies challenge our ability to control them (Collingridge, 1980).

In addition to reflection on changing responsibility practices in relation to technology, there is also a rich history of what we might today call 'digital ethics' (Véliz, 2023) but what was once called 'computer ethics' (Johnson, 1985; Moor, 1985). Today, the term 'computer ethics' strikes us as dated partly because of the narrow focus it suggests. When computers were chunky room-sized devices it made sense to focus on the machines as the locus of our ethics. However, over time our attention began to focus on the *invisible* aspects of computing, and in the 1990s

> *In order to have a future, we must first confront our past."*
>
> **Crawford and Joler, 2023**

theories such as Luciano Floridi's 'information ethics' become more popular (Floridi, 1999). Today, digital ethics comes to encompass far more than just 'computers', and has many sub-disciplines interested in far more than just 'information'.

While there have of course been similar kinds of work done on various historical features of normative aspects of AI, few have looked to the specific history of thinking on 'responsible' and 'responsibility' and its anchor in the humanities. One recent study, for example, looked at the ways in which technological change influences how we think about responsibility (Shanley, 2022). Specifically, Shanley traces the development of Responsible Research and Innovation (RRI), through the '60s and '70s, as an intellectual movement grounded in specific academic disciplines, but also characterised by a "shift in public attitudes towards the value of more participatory forms of engagement" (Shanley, 2021, p. 235).

There have also been historical approaches to specific case studies in AI (Cihon, Maas and Kemp, 2020; Srinivasan and Uchino, 2021) and histories of particular disciplines, such as computer science and 'digital ethics' (Bynum, 2008; Muller, 2022). We also see histories of specific pieces of legislation (Stix, 2022; Smuha and Yeung, 2024). However, with the exception of Shanley's work (which nonetheless has a broader remit than ours), more often than not the historical reflection of these works is noted rather briefly, with the main focus being a specific problem in AI discourse.

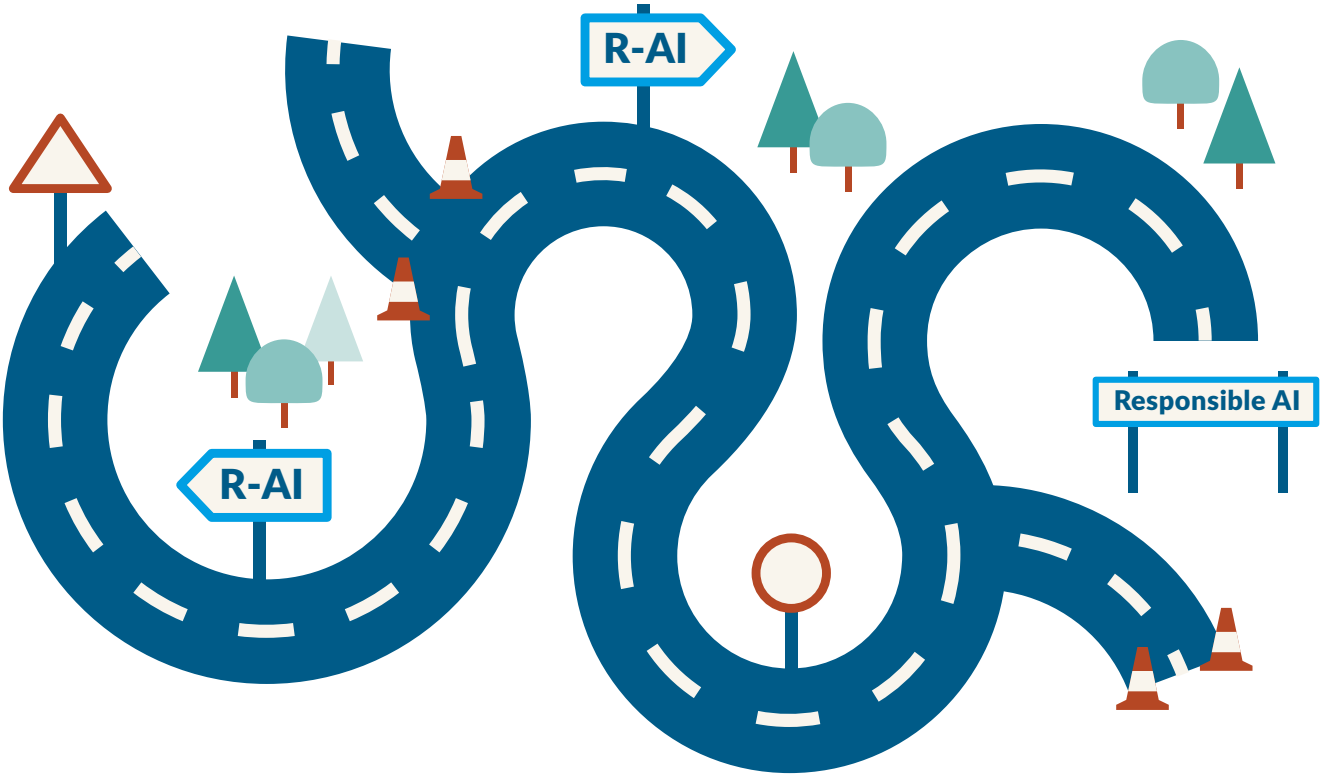
In this study we hope to make a more general contribution, by reflecting on broader historical developments in academia, industry, and governance, in the service of teasing out key lessons for the future of R-AI. More recently, Crawford and Joler have done exciting work mapping the 'AI Empire' by chronicling the historical operations of power and technology since 1500 (2023). They claim that "if we are to address the urgent challenges of the contemporary time - including technocratic fascism, climate catastrophe, colonial wars, and wealth inequality - we need to contend with the interwoven nature of their histories. In order to have a future, we must first confront our past." (Crawford and Joler, 2023). We take inspiration from this arts and humanities-led approach coming as it does from a longer and lesser-known tradition of arts and design thinking around human-computer interaction more broadly, and thus our goal is to show how the historical development of thinking around 'responsibility' as it relates to technology – its development and its deployment – can shape the responsible stewardship of the AI ecosystem.

While 20th century uses of normative ethical concepts like responsibility to apply to computing technologies were largely theory-driven or oriented toward applied professional ethics, in the first decades of the 21st century a clear 'governance' focus started to emerge, with approaches such as Responsible Research and Innovation (RRI) becoming increasingly influential. RRI takes as its object of study the responsible oversight of the innovation process, governing this process through regulatory mechanisms to ensure that the products of innovation are socially desirable (Von Schomberg, 2012, p. 50).

In parallel with developments in RRI, which maintained a focus on the technology practitioner perspective, the conceptual and empirical knowledge foundations of 'Responsible AI' were deepened by the field of 'AI ethics' that emerged during this period. AI ethics itself has worn many faces, from the moral philosophy-centered research on AI that emerged from the scholarly crucible of digital and information ethics (Johnson, 1985; Maner, 1996; Brey, 2008), to the body of sociological and empirical research that was driven by growing awareness of harmful AI and algorithmic impacts, particularly upon marginalised communities (Birhane, 2021b; Birhane *et al.*, 2022; Queer in AI *et al.*, 2023), to the advocacy and civil society voices that initially used 'AI ethics' as a frame, before turning toward more politically rich concepts like algorithmic and data justice (Fazelpour and Lipton, 2020; Irani, 2023). Tech corporations and public sector actors have also come to play an important role in shaping the direction and set of best practices associated with R-AI (Legassick and Harding, 2017; Pichai, 2018; Ada Lovelace Institute, 2021; The Alan Turing Institute, 2021).

This brief tour showcases the way that ethical reflection on technology has changed over time, and how the focus on AI is relatively new. It also highlights the twists and turns in the journey from theory to practice, starting with what has been called the 'empirical turn' in early 21st century philosophy of technology (Achterhuis 2001, Verbeek 2022), to the emergence of RRI and 'ethics by design' as sites of practitioner intervention (Dignum *et al.*, 2018; Brey and Dainow, 2024), to the industry toolkits, policy white papers and training resources that continue to populate the R-AI landscape.

The many faces of AI ethics.



Implementing R-AI is often challenging, with competing stakeholder objectives, making it a non-linear process.

Yet R-AI continues to draw heavily upon research informed by theoretical perspectives from many disciplines, whether on the philosophical conditions of moral responsibility for AI (Vallor and Vierkant, 2024), the political requirements of AI's legitimacy as a social power (Zimmermann, Vredenburgh and Lazar, 2022) or the more sociological, ethnographic and anthropological analyses by Science and Technology Studies (STS) scholars (Selbst *et al*., 2019; Beer, 2022).The practitioners who currently make up the R-AI ecosystem come from these (and many more) disciplines and sectors, and their interaction is what makes the R-AI ecosystem what it is.

The hope is that this study, with its focus on the history of R-AI, can help ground these practices in a way that both showcases the dynamism of these distinct sets of practices, and provides some normative orientation for how best to enable a flourishing ecosystem.

In what follows we outline some of the most significant moments in the history of critical reflection on 'responsibility' in relation to technology, and eventually AI. Though the literature draws heavily from the humanities, perspectives from computing, engineering, arts and design and social sciences play important parts in the story. This provides an in-depth theoretical base from which we can understand and appreciate how this notion of responsibility can and should be applied to AI research and practice. We start from the postwar period, going through each decade up until the present (currently, 2025).

1950

1950 Cybernetics pioneer **Norbert Wiener**'s *The Human Use of Human Beings*

1954 **Martin Heidegger**'s *The Question Concerning Technology* (*Die Frage nach der Technik*)

1960

1964 English translation of **Jacques Ellul**'s *The Technological Society*

1966 First ACM Guidelines for Professional Conduct in Information Processing

1970

1976 **Joseph Weizenbaum**'s *Computer Power and Human Reason*

1978 **Walter Maner** coins term 'computer ethics'

1980

1980 **Langdon Winner**'s *Do Artifacts Have Politics?* and **David Collingridge**'s *The Social Control of Technology*

1985 *Metaphilosophy* Issue on 'Computers and Ethics' with **James Moor**, **Deborah G. Johnson**, **Terry Bynum** and **Maner**; publication of **Johnson**'s *Computer Ethics* textbook

1990

1992 **Donald Gotterbarn** leads first ACM Code of Ethics

1995 Launch of **ETHICOMP** conference and journal *Science and Engineering Ethics*

1997 Launch of **CEPE** conference

1999 Launch of **Springer** journal *Ethics and Information Technology*

2000

2004 **First conference on 'roboethics',** in Sanremo, Italy; **IEEE** launches roboethics committee

2009 **Wendell Wallach** and **Colin Allen**'s *Machine Ethics*

2010

2010 **ImageNet contest launched**

2013 **René von Schomberg**, **Jack Stilgoe**, **Richard Owen** and **Phil Macnaghten** publish research on frameworks for responsible research and innovation (RRI)

2016 **Joy Buolamwini** founds the **Algorithmic Justice League** and releases *The Coded Gaze*

2017 **AI Now Institute** founded

2018 **Kate Crawford** and **Vladan Joler**'s *Anatomy of an AI System*; Responsible AI guidelines released by Google, Microsoft, OpenAI and others; first **AAAI/ACM** conferences on AI, Ethics and Society (AIES) and Fairness, Accountability and Transparency (now FAccT); launch of **Ada Lovelace Institute**

2019 **OpenAI** claims **GPT-2** too powerful for public release; OECD AI Principles adopted; IEEE-SA launches work on standards for responsible and ethically designed AI

2020

2020 Google dismisses **Timnit Gebru** and soon after, **Meg Mitchell**; publication of *The Oxford Handbook of Ethics of AI*

2021 **UNESCO** releases first global standard on the Ethics of AI

2022 **OpenAI** releases **ChatGPT**

2023 UK and US each establish an **AI Safety Institute (AISI)**; in USA, **NIST** releases **AI Risk Management Framework**; **Meta** and **Microsoft** disband Responsible/Ethical AI teams

2024 **EU AI Act** Passed

2025 **UK AISI** renamed the **AI Security Institute**; US and UK decline to sign declaration at Paris AI Action Summit; US Executive Order 'revokes certain existing AI policies and directives that act as barriers to American AI innovation'; Google drops Responsible AI pledges to not develop AI weapons and surveillance

SECTION TWO

A Chronology of the Responsible AI Ecosystem

1950s – 1960s

There has always been a strong interest in the ethical implications of new technological systems. One of the earliest and most important 20th century voices on this subject was cybernetics pioneer Norbert Wiener (Wiener, 1954, 1961). Already in the 1950s, Wiener was reflecting on the potential impacts that information technologies would have on important human values (for example, happiness, security, freedom, health). Wiener himself did not explicitly refer to his project as one concerned with the creation or articulation of a new branch of ethics, nor did he offer much in the way of metaphilosophical reflection on his project more generally (Bynum, 2008). While he did not create any unique terms for what he was doing, his work became influential in what became 'computer ethics' (in the 1970s and 1980s) and 'information ethics' (in the 1990s).

Another tentative starting point for reflections on not just technology more generally but on the implications of computing technologies comes from developments in professional ethics. We can trace the beginning of this history to computer scientist Donn Parker, who in the 1960's began reflecting on what today we might call 'computer ethics'. He remarked that "when people entered the computer centre, they left their ethics at the door" (Bynum, 2001, p. 110). Here we see the beginnings of a nascent subfield, where moral evaluations of computing practice arise from within the technical community. Parker was also involved with the US *Association for Computing Machinery* (ACM), and pushed them to adopt the 'Guidelines for Professional Conduct in Information Processing' in 1966. The current version of this document is known as the 'ACM Code of Ethics and Professional Conduct' (2018). Professional ethics thus has a long history in reflections on computing, and still plays a role in that reflection today.

In parallel to work done on professional computing ethics, there was the emergence of the interdisciplinary field of Science and Technology Studies (STS) from the 1960s onwards. STS took as a premise that science and technology were socially embedded and constructed enterprises, and used a variety of empirical and historical methodologies to tease out the implications of this thesis for society. Today, while we find STS scholars, those versed in professional ethics, and philosophers jointly participating in R-AI research, there is often a deep tension between these approaches.

For example, moral philosophers are typically comfortable making explicit normative claims, while STS scholars frequently adopt a more descriptive orientation towards their research. One of the early pioneers of this kind of research was Jacques Ellul, who in his 1964[2] text *The*

2 The original French version of the book, *La Technique ou l'enjeu du siècle* ('Technique or the stake of the century'), was published in 1945.

Technological Society, argued that we live in a society dominated by "technique". *Technique*, for Ellul, "*is the totality of methods rationally arrived at and having absolute efficiency*… in *every* field of human activity" (Ellul, 1964, p. xxiv). It is thus a totalising influence, where "nothing at all escapes *technique* today" (1964: 22). *Technique* is a method for making processes more efficient, rational, and profitable. While it seems clear that Ellul is levelling some kind of social critique, throughout the text he insists that he is not making any value judgements of his own, but rather reporting the facts as they are.

Here we see the tension: This kind of research is itself deeply political and normatively laden, but the dominant disciplinary orientation of the social sciences is descriptive. This stands in sharp contrast to the thread of overt philosophical critique of advanced industrial cultures and technological systems that took shape in Herbert Marcuse's *One Dimensional Man* (1964), and in Martin Heidegger's *The Question Concerning Technology* (1954), which by the mid 60s was already shaping philosophers' thinking about our troubled relationship to technology.

Heidegger's existential focus laid bare the essence of technology as a social force that shapes not only how humans live, but constrains how they think, perceive and value the world and one another, in ways that can be devastating. Were it not for Heidegger's eventual exposure as a craven Nazi sympathiser and his rightful loss of status in the philosophy of technology canon, his critique – highly influential at the time – might have renewed force for thinking about AI today. Yet Heidegger's position on responsibility was even more discouraging, and far less helpful than Ellul's. His perspective on modern technology as a necessarily alienating metaphysical force overriding human agency left little room for thinking about our responsibility for technosocial change.

In general, during this postwar period, reflection on technology and human responsibility had two primary foci; one being the devastating material effects these technologies could have (such as the effects of nuclear weapons or chemical pollution)[3], and the other being the troubling political and existential impacts of new technologies on human freedom, agency, and purpose. 'Responsibility' for these feared impacts, then, was perceived as growing in unison with the technological powers at our disposal. Much of the work during this time was being done in research intensive arenas, by academics, and so the focus on practice and governance that we find later had yet to take hold.

1970s – 1980s

In the 1970s and 1980s the ethical and political implications of technology were attracting an increasing amount of attention. As noted earlier, Jonas was an early pioneer in reflecting on the relationship between technology, human action, and responsibility (Jonas, 1973).

---

3 A great example of a popular science book warning about the dangers of environmental harm during this time was Silent Spring (Carson, 2002).

Meanwhile, philosophers such as Langdon Winner were interested not only in the instrumental effects of technology but in the potential for artifacts to be carriers of values themselves. The most famous example of this is Winner's analysis of New York architect Robert Moses' design of a series of bridges linking Long Island and Jones beach (Winner, 1980). Moses was in charge of designing the bridges, which were unusually low. This had the effect of prohibiting larger vehicles, such as busses, from accessing the beaches. Winner argued that this was a feature, not a bug, of the bridges' design, which served to limit the access of poor and mostly African American locals from travelling to the beach with public transport.

Winner's point here is not just that the bridges are 'used' politically, but rather that they are *themselves* political: Winner argued that Moses had in fact imbued his own (discriminatory) values into the bridges. The bridges, in his terminology, had politics[4]. Winner's work has two links to discussions on R-AI. First, he initiates what would later become a core focus of R-AI on *values in design* (Brey and Dainow 2023), with a focus on the way that designed intentions come to manifest in material artifacts[5]. Second, Winner was especially interested in the way that these technologically embedded values come to shape *power relations* (Turculet, 2023), a theme that would also become dominant in later R-AI discourses (Costanza-Chock 2018, Benjamin 2019, Crawford 2021).

This shift toward the material and political aspects of technology was influential in the development and consolidation of STS as a distinct research agenda, culminating in the establishment of the Society of the Social Studies of Science (4S) in 1975 (Jasanoff, 2000). During this time, STS differentiated itself from other disciplines with an explicitly empirical and material, even embodied approach to studying the impacts of technology (Turkle, 2005; Haraway, 2006). This was distinct from, for example, the more conceptual work of Hans Jonas, Albert Borgmann, Don Ihde, Andrew Feenberg and other late 20th century philosophers of technology.

Almost in parallel to this work on the material aspects of technology, computing pioneer Joseph Weizenbaum was doing research on the normative aspects of computing and human-machine interaction (Weizenbaum, 1976). His work on this topic was prompted by his experiences in creating and testing a computer program 'ELIZA'. Weizenbaum was especially concerned that there was a tendency among people to see humans as machines, and that various computer technologies might exacerbate this trend. In *Computer Power and Human Reason*, he argues that we need to uphold the morally important distinction between deciding and choosing. Deciding, for Weizenbaum, was an activity that could be carried out by means of computation, i.e., by *calculation*. Choosing, in contrast, is a product of *judgement*, and is thus a distinctly human activity (Weizenbaum, 1976, pp. 259–260). Thus, we might conclude, responsibility is for humans, not machines.

---

4 For influential critiques of Winner's characterization of Moses see (Joerges, 1999; Woolgar and Cooper, 1999).

5 There is also a link here to work in Value Sensitive Design, which bakes values into the design process (Friedman, Hendry and Borning, 2017; Umbrello and van de Poel, 2021).

> *It was also during this time that the idea that there should be a unique discipline devoted to computer ethics gained significant academic traction."*

It was also during this time that the idea that there should be a unique discipline devoted to computer ethics gained significant academic traction. Three philosophers stand out here: Walter Maner, Deborah G. Johnson, and James Moor. Maner was apparently the first to use the term 'computer ethics' (Maner, 1980), and used it to refer to "ethical problems aggravated, transformed, or created by computer technology" (Bynum, 2001, p. 110). During the 70's Maner developed computer ethics courses and travelled around the US giving lectures and workshops on the topic (Bynum, 2001, 2008). Johnson, in 1985, published her textbook *Computer Ethics*, which was and still is a foundational text in the history of computer ethics. In this same year James Moor published his now famous "What is Computer Ethics" article, which appeared in a computer ethics special issue of the journal *Metaphilosophy* (Moor, 1985).

Important to note is that Johnson and Maner had a significant disagreement about the role that computer technology played in ethics. For Johnson, this technology allowed us to see old problems in a new light, giving them a 'new twist', but did not necessary bring anything unique to the table (Johnson, 1985). For Maner, however, this technology generated entirely new ethical problems, implying new dimensions of responsibility to be laid at the feet of, among others, computing professionals. Maner argued that "computer ethics is an academic field in its own right with unique ethical issues that would not have existed if computer technology had not been invented" (Maner, 1996, p. 137).

Moor, however, went further than both Johnson and Maner, by asking *why* computers raised these apparently unique ethical problems (Moor, 1985). He suggested that "Computer ethics requires us to think anew about the nature of computer technology and our values" (Moor, 1985, p. 268). The reason for this, according to Moor, is that computers have *logical malleability*: they "can be shaped and moulded to do any activity that can be characterised in terms of inputs, outputs, and connecting logical operations… the limits of computers are largely the limits of our own creativity" (Moor, 1985, p. 269). For Moor, this meant that the impacts of computers could and would go far beyond the 'computer ethics' of Johnson and Maner, creating policy and conceptual 'vacuums'.

What we see during this time is distinct from the orientation towards the effects of technology that characterised the previous period. Here there is a focus on the material, but also a critical engagement with how that materiality comes to shape our social and political structures. Moreover, the worries about the decision-making power of 'computers' contributed to educational changes in computer science curriculums, once again shifting the window of responsibility and the meaning of responsible innovation.

Debates over the ethical impact that computers might have, both on societies and individuals, contributed to a growing awareness that there was increasing responsibility on those who

had power to shape the design and development of these systems. As we will see in the next section, this work became increasingly important in efforts to develop professional ethics for computing engineers and developers. While this thinking had yet to go 'mainstream', there were nascent ideas here that would come to influence the development and meanings of R-AI. Being 'responsible' now took a different meaning, and expanded from a general philosophical critique, to concrete recommendations to develop proper ethical training and education for technology professionals.

1990s

During the 1990s computer ethics came into its own. Many research programs, conferences and journals relating to the ethics of computing technology emerged. In 1995, the ETHICOMP series of conferences began at De Montfort University, and the CEPE (Computer Ethics and Philosophical Enquiry) conference series was launched at Erasmus University in 1997. The journal of *Ethics and Information Technology* was first published in 1999, and *Science and Engineering Ethics* launched in 1995.

During this time the philosopher Luciano Floridi was also writing about "Information Ethics", which was (and still is) an attempt to understand and evaluate the ethical impact of new digital technologies on individuals and society (Floridi, 1999). A significant dimension of Floridi's influence was a focus on the ethics and metaphysics of *information* rather than computers' processing power and analytical capacities. This research program was an important part of what later became 'data ethics,' and many researchers working on R-AI still describe themselves as Data Ethicists. Instead of a focus on the ethical implications of digital technologies themselves, data ethicists look at the role of data, and practices around data, and how they might be managed responsibly.

Another major thread to pick up on during this time is the work of Donald Gotterbarn and colleagues advocating for professional responsibility in the context of computer ethics. This work involved working with influential professional societies such as the ACM (Association for Computing Machinery) and the IEEE (Institute of Electrical and Electronic Engineers) by developing and licensing standards and professional codes of conduct for computing and software professionals (Gotterbarn, Miller and Rogerson, 1997).

While this work had limited impact in the long run – in part due to institutional inertia and in part due to a growing resistance to professionalisation and self-regulation within the software community – there was a concerted effort during this time to both establish the novelty of concerns raised by 'computer technology' and prepare practitioners and stakeholders to deal with these issues (Gotterbarn, 1991).

An increasingly popular discussion in debates during this time was precisely where 'computer ethics' ended or began. Did any situation that involved a computer come under the umbrella of computer ethics? Or was there a specific role that a computer needed to play in the moral scenario? Gotterbarn, for example, argued for a more circumscribed understanding of computer ethics: "the only way to make sense of 'Computer Ethics' is to narrow its focus to those actions that are within the control of the individual *moral* computer professional" (1991, p. 28).

There was also increasingly influential literature on 'engineering ethics' (Fleddermann, 2012), which investigated the important role of responsibility in engineering education (Prichard, 1998), and the engineer's interaction with society (Nichols and Weldon, 1997). This literature gradually moved away from framing engineering and engineers as being engaged in ethics as a simple exercise in cost-benefit analysis and risk management, and came to highlight that engineering was a *social* concern, and that there is a difference between the risk of a particular design and its social acceptability (Busby and Coeckelbergh, 2003, p. 364).

During this time, however, the precursors of R-AI were mostly academic researchers in philosophy and STS – there was only minimal engagement with computing disciplines themselves, and virtually no related systematic activity in the computing industry, barring a number of ground-breaking but localised efforts such as Xerox PARC and corporate artist-in-residence schemes.[6] While the work of Gotterbarn reached beyond academia, the focus was on *individuals* and their associated professional responsibilities. And while there was an emerging concern for data privacy, these concerns were mostly compliance related. Responsibility as it related to digital technologies, then, was a rather narrow and academic enterprise.

2000s – 2010

The early 2000s saw the establishment and mainstreaming of critical and ethical reflection on information technologies, with the emergence of a number of sub-fields and disciplines. Up until around 2010, however, most concerns were around issues related to 'big data', on the one hand, and the ethics of robotics, on the other. Specific interest in AI technologies was therefore still quite niche, as the disruptive impact of machine learning technologies had not yet dawned. Yet there were already early moves toward exploring it, such as the 2002 'Artificial Stupidity/ Artificial Intelligence'[7] conference at The Banff Centre, which included perspectives from the arts and game design as well as human-computer interaction.

With respect to robotics, the first conference on 'roboethics' took place in Sanremo, Italy, in 2004[8]. This focus on robotics was fed in part by optimism around the future impacts of the field. Additionally, there were also growing worries about the application of autonomous systems such as 'robot soldiers' in the military context, and what that would mean for our practices of ascribing responsibility (Sparrow, 2007).

That same year (2004), the Institute of Electrical and Electronics Engineers (IEEE) Robotics and Automation Society (RAS) established a technical committee on 'Roboethics'. One of the goals of this committee was "analysing the ethical implications of robotics research… by promoting the discussion among researchers, philosophers, ethicists, and manufacturers"[9]. Notice how

6 For an overview of the ways that artists challenge and extend our thinking about emerging technologies see (Century, 2022)

7 See https://archives.banffcentre.ca/link/descriptions53584 (accessed 7 February 2025).

8 https://web.archive.org/web/20070928113317
http://www.roboethics.org/sanremo04/index.php (accessed 31 October 2023).

9 http://www.roboethics.org/ieee_ras_tc/index.php (accessed 31 January 2024).

“these systems were based on learning from large datasets, rather than environmental or embodied learning.”

there is no explicit mention of policy makers or those involved in governance. It seems that the scope of responsible research and innovation with respect to robotics was rather limited, at least at this early stage.

Another field that developed in tandem with roboethics, and indeed was influenced by it, was ‘machine ethics’ (Anderson and Anderson, 2007; Wallach and Allen, 2009). This research domain was not interested in whether human beings *used* machines ethically, but rather with whether these machines themselves could be ‘ethical’. The focus, therefore, was on the software necessary in order to instantiate a ‘moral machine’ capable of making moral decisions. This resulted in a debate on what the best architecture would be for this kind of machine: ‘Top-down’ symbolic approaches based in formal logic and rule-based programming, or ‘bottom-up’ approaches based on neural networks and embodied learning techniques.

The machine ethics community was heavily focused on the software components of intelligent systems, and much of the fiercest debate in this field concerned the type of software architectures that would be needed for machine ethics to be successfully realised (Wallach and Allen, 2009). This approach seemed to assume that ethics was something that could (at least in principle) be ‘engineered’ or ‘implemented’ into a machine (whether ‘bottom up’ or ‘top down’). These and other assumptions would prove to be critical points of discussion in the field, which is still active (but far from uncontroversial) today (van Wynsberghe and Robbins, 2019; Formosa and Ryan, 2021).

What was becoming increasingly clear towards the end of this period was that *data* would play a role in the development of AI-systems going forward. Although the *significance* of data had not quite been grasped, the emergence of fields like ‘data ethics’ and discussions around ‘big data’ would pave the way for distinct R-AI research agendas in the future. We can see traces of this idea in the narrative around the shift from ‘Web 1.0’ to ‘Web 2.0’, where there was a focus on the role of data management as foundational to the future development of the internet (O’Reilly, 2007). However, what nobody anticipated at the time was the imminent merger of big data with machine learning developments like backpropagation, self-supervised learning and natural language understanding that would ensure the current dominance of ‘bottom-up’ strategies for AI. For the most part, these systems were based on learning from large datasets, rather than environmental or embodied learning.

What we see during this period is parallel growth in the fields of ‘roboethics’ and ‘data ethics’. What we do not see is sustained attention and focus on AI-systems specifically. The field of the ‘ethics of AI’ or ‘AI ethics’ had yet to really emerge, and the dominance of AI only really took hold after 2010. Part of the reason for this, which will become clearer in the next section, is the

lack of commercial applications for AI-systems. During this time, the field was still dominated by academic discussions that had yet to penetrate into mainstream discourse.

With reflections on data ethics, however, we start to see explicit reference to governance start to enter the picture. The 'responsible' use of data become not only a technical concern but also a legal and political one.

2011 – 2015

One of the hallmarks of this period was the solidification of Responsible Research and Innovation (RRI), especially in the European context, as a distinct governance perspective to adopt in relation to innovative products (von Schomberg, 2011, 2013, p. 51). RRI took inspiration from the work of researchers like Jonas, Collingridge, and Winner, in their focus on responsibility and emphasis on the social and political constitution of technological artifacts.

Moreover, RRI leaned heavily on work done in "Technology Assessment" (TA), itself a subdiscipline of STS, which moved away from 'technological forecasting' and instead focussed on making research and innovation processes more reflective and inclusive (Schot and Rip, 1997; Guston and Sarewitz, 2002; von Schomberg, 2013). In practice this meant baking in insights from TA (Grunwald, 2011) and STS literature more broadly so that the impacts of new technologies could be assessed "beyond their anticipated market benefits and risks" (von Schomberg, 2013, p. 51).

Stilgoe *et al*. took this idea further by explicitly articulating a forward-looking conception of RRI in their 2013 paper. They (broadly) define RRI as "taking care of the future through collective stewardship of science and innovation in the present" (Stilgoe, Owen and Macnaghten, 2013). While RRI did not specifically take aim at AI-based technologies and was a more general mechanism of governance and oversight, its legacy is nonetheless important. Part of the reason for this is that it explicitly aims to articulate a sense of 'responsible' that is compatible with emerging science and technology, which is a similar frame that we find in R-AI. The sense of 'responsible' in RRI is less about 'who' or 'what' is responsible for some outcome, but rather centres on questions of the specific technology under consideration, such as, for example, synthetic biology (Taylor and Woods, 2020). As von Schomberg notes,

> "For modern innovations, responsibility for the consequences of implementation is then primarily related to the properties and characteristics of the products or the technology and less to the privileged owners and creators of the technology"
> (von Schomberg, 2013, p. 53).

This framing is important to bear in mind going forward, as this sense of 'responsible', as understood by RRI practitioners, is instructive with respect to the 'responsible' in R-AI. RRI researchers directed their attention towards ensuring that technologies were designed in such a way that they displayed 'responsible' *characteristics*. 'Responsible', here, meant something like 'socially benign' or 'socially beneficial'. We can see similar calls in contemporary R-AI literature, with calls for AI research to be in the service of the 'social good' (Floridi *et al*., 2020). Within the RRI literature, however, one finds a tension: is it the product itself that needs to be 'responsible', or is it the implementation of that technology in a certain context that ought to be 'responsible'?

A key feature of this period was also the commercial success of AI systems, which prompted widespread recognition of their potential ethical impacts."

The shift identified by von Schomberg from 'owners to products', while useful, seems to downplay the role that social context might play in the appropriateness and responsible character of a given technological development (Carrier, 2021, p. 4764). This is a theme that we find in current discussions in R-AI, as it might seem natural to read R-AI as being more concerned with technology than with society or social context. This perception, however, has been changing in recent years, and there is an awareness that "a system can be transparent, fair and accountable whilst operating in an ethically questionable domain" (Sadek *et al*., 2024).

For example, a proposed AI intervention that satisfies the primary principles understood to be necessary conditions of R-AI -- Fairness, Accountability, and Transparency -- might still be ethically unsound, even profoundly evil (Keyes, Hutson and Durbin, 2019). Part of the reason for this is that these principles are *contested*, and this means they can be cashed out in a variety of ways that when put together might be straightforwardly unethical. Another more fundamental problem is that these principles are *insufficient* conditions – they are silent on fundamental rights and considerations of justice that, if ignored, can in principle allow for an algorithmic holocaust to be carried out in the open, while wearing 'Responsible AI' as a badge of ethical legitimacy.[10]

A key feature of this period was also the *commercial* success of AI systems, which prompted widespread recognition of their potential ethical impacts. One of the main drivers of this process was the emergence and recognition of the (commercial) powers of AI systems, and the associated ethical concerns that these systems raised. ImageNet was one of the first cases of AI-powered technologies getting the spotlight.

Released in 2009, ImageNet is a huge database of images that can be used in training models for object recognition. Each year from 2010 to 2017 there was an ImageNet contest where software developers could test their classification algorithms, hosted on the official ImageNet site[11] (in 2017 the contest was transferred to Kaggle[12], and it was closed in 2020). Termed the 'ImageNet Large Scale Visual Recognition Challenge' (ILSVRC), the contest involved participants developing models that could classify images as belonging to one of a thousand different categories.

---

10 Keyes, Hutson, and Durbin, in their sardonic A Mulching Proposal showcase how this might play out (2019).

11 https://image-net.org/challenges/LSVRC/index.php (accessed 30 January 2024).

12 https://www.kaggle.com/c/imagenet-object-localization-challenge (accessed 30 January 2024).

In 2010, the best program could label an image correctly 72% of the time. In 2012 that number jumped to 85%, and in 2015 achieved 96% (which is higher than the human average of 95%). The 2012 results were seen as a watershed moment, not only because of the high level of accuracy, but also because of the innovative method used to achieve it, artificial neural networks (ANNs). While ANNs had been around since the mid-20th century, their success up until this point had been haphazard. The winner of the 2012 competition, Supervision (created by the team at the University of Toronto), however, was a deep convolutional neural network. 'Deep' here means that the neural network had many more layers than its predecessors, and so could process the images in a more detailed manner. The network is 'convolutional' because each image is broken down into smaller parts for processing, making the final label less contingent on where certain objects happen to be in the image.

These deep neural networks were recognised as having the ability to be used in a variety of commercial contexts, and this ability to be useful in multiple domains spurred public, private, and academic interest in AI research. For example, image recognition technology is the foundation of the field of computer vision, and has been used in research on autonomous vehicles and in medical diagnostic tools. While models trained on the ImageNet dataset had immense commercial success, there were early concerns regarding bias in the training data (concerns which later led to full blown controversies) (Buolamwini and Gebru, 2018). A highly impactful intervention from the arts and humanities during this time was Kate Crawford and Trevor Paglen's *Training Humans* exhibition, which featured 'ImageNet Roulette'.[13]

ImageNet Roulette, developed by Leif Ryges, exposed the harmful stereotypes embedded in the ImageNet dataset (Statt, 2017). It did so in ways that struck the user personally; by uploading your own photo, you saw the harmful stereotypes that ImageNet would use to classify *you*. Five days after the exhibit opened, the ImageNet project announced it would remove 600,000 photos linked to 1593 "unsafe" and "offensive" or "sensitive" categories of people (Kuesel 2019). Further ethical issues, such as the data being labelled by low-wage workers, would be repeatedly highlighted over the next few years (McQuillan, 2022; Perrigo, 2023).

Significantly, the first Ethics of AI conference also took place in 2012 (Impacts and Risks of Artificial General Intelligence, AGI IMPACTS). Since this time, the field of 'AI ethics' has expanded massively, with institutes at a number of universities (Oxford, Bonn, TU Munich). These institutes are just the tip of the iceberg, as many centres that deal with issues covered in AI ethics do not necessarily describe themselves as such, or have a broader remit (such as digital technologies more generally). Yet the ethics of AI is increasingly the focus, as indicated by the establishment of a growing number of professorial chairs and postgraduate programmes in the field.

As noted earlier, this interest in AI specifically was spurred by recent advances in the abilities of these systems (not to mention their commercial impact). In barely a half-decade, 'AI Ethics'

13 For more information on ImageNet Roulette, see https://www.chiark.greenend.org.uk/~ijackson/2019/ImageNet-Roulette-cambridge-2017.html (accessed 27 November 2024)

While it was clear that the 'ethics' at the time lacked teeth, in the coming years there was a regulatory and academic push to make AI ethics stand up to its name."

or the 'Ethics of AI' went from a niche interest to a global concern, as researchers and policy makers started to realise the impacts that this technology could have on global governance and individual well-being.

It is instructive to look at that first conference in 2012 and its focus: The future of artificial *general* intelligence. Some of the themes of the conference were questions regarding a future where we are surrounded by 'super intelligent' beings. Another area of concern was the question of AI 'safety' (now a heavily contested term), and whether we can ensure that AI-systems remain 'predictable'. From these brief observations it seems that the 'ethics' of AI had yet to fully live up to its name. The 'ethical' issues the conference raised seemed to be more concerned with a speculative AI-enveloped future than with any short- or medium-term risks that AI-systems might pose to living people and communities (such as questions of bias or discrimination).

This, unfortunately, is a theme that has repeated itself in more recent debates about the future risks of AI, with increasing polarisation between those stressing 'existential' risks and the need for AI Safety (capital S) and those concerned with AI safety (small s) and real-world risks like concentration of power, labour disruption, bias, energy consumption and discrimination. However, it was still difficult to predict the mainstream success that a focus on the ethical impacts of AI would receive in the coming years. While it was clear that the 'ethics' at the time lacked teeth, in the coming years there was a regulatory and academic push to make AI ethics stand up to its name.

2016 – 2020

During this time increasing attention was drawn to AI-based systems due to several high-profile incidents involving AI used in recruiting (Dastin, 2018), facial recognition (Buolamwini and Gebru, 2018), criminal sentencing (Angwin *et al*., 2016), and voter profiling (Susser, Roessler and Nissenbaum, 2019). 'Poet of Code' Joy Buolamwini's work was particularly influential during this period, amplified by her use of the arts as a medium. Her 2016 mini-documentary *The Coded Gaze* debuted at the Boston Museum of Fine Art, exploring and critiquing racial prejudice in facial recognition technologies.[14] Her 2018 spoken word poem 'AI, Ain't I a Woman?' conveyed the message of the influential Gender Shades paper to an audience

14 The video can be found here: https://www.youtube.com/watch?v=162VzSzzoPs

beyond researchers, as did the award-winning feature documentary film based on her work, *Coded Bias*, directed by Shalini Kantayya, which premiered at the Sundance Film Festival in 2020.

From 2015 we started to see a significant increase in the number of academic publications explicitly linking AI to 'ethics'. For example, in 2015 there were 10 articles published that mention 'AI' or 'artificial intelligence' and 'ethics' or 'ethical' in the title. By 2020 that number had grown to 342 (Borenstein *et al*., 2021). One of the key themes during this time was the attempt to mobilise principles into action (Peters *et al*., 2020).

The academic work during this time focussed on a wide range of issues, and importantly for the purposes of this study, questions concerning responsibility were gaining a significant degree of attention. More specifically, the question and existence of 'responsibility gaps' became a popular theme of investigation (Danaher, 2016; Köhler, Roughley and Sauer, 2017; Nyholm, 2018; Santoni De Sio and Van Den Hoven, 2018; Oimann and Tollon, 2024). These discissions, however, are fairly high level and concerned with the conceptual work of clarifying the meaning of 'responsibility' and whether there is a technological bar to the appropriate attribution of moral responsibility for highly autonomous machine actions.

In addition to this academic work, advances were also being made in the policy domain. In 2016, across 25 countries, there was a single bill passed into law mentioning 'AI'. By 2021 that number had jumped up to 18 (Zhang *et al*., 2022). Additionally, different regions published their own reports: The 'Report on the Future of Artificial Intelligence' (2016, United States), and the 'OECD Recommendation of the Council on Artificial Intelligence' (OECD, 2019) (OECD countries). The EU was a leader during this time, adopting the General Data Protection Regulation in 2016 (European Parliament, 2016), and in 2019 the AI High-Level Expert Group (AI HLEG) presented their findings to the European Commission (EC) in the form of the Ethics Guidelines for Trustworthy AI (AI HLEG, 2019).

Organisations were also developing ethical guidelines for AI research: The Advancement of Artificial Intelligence (AAAI) adopted a Code of Professional Ethics and Conduct in 2019 (and later, in 2021, released a Diversity Statement). The AAAI Code sees 'acting responsibly' as reflecting on the wider impacts of one's work and consistently supporting the public good.[15] Although directed at individual practitioners engaged in AI development, there is constant reference to the wider social context.

Private companies also started to publish their own AI guidelines, and in 2016, a "Partnership on AI" was formed. This Partnership was a coalition between a number of large AI companies (Amazon, Apple, Facebook, Google, Microsoft, and IBM), and had the goal of developing a set of best practices for AI research, development, and deployment[16]. Additionally, individual companies started developing their own guidelines, such as the 'Microsoft AI Principles'

15 https://aaai.org/about-aaai/ (accessed 30 January 2024)

16 https://partnershiponai.org/about/ (accessed 30 January 2024)

These principles are fairness, reliability and safety, privacy and security inclusiveness, transparency, and accountability."

(Microsoft Corporation, 2019), 'OpenAI Charter' (OpenAI, 2018), and Google's 'AI Principles' (Pichai, 2018). A number of high-profile academic publications were also released, for example two edited collections by Oxford University Press (Dubber, Pasquale and Das, 2020; Liao, 2020).

An important theme here is the choice within industry to adopt the 'Responsible AI' frame, as opposed to using the term 'ethical'. This is especially true in the case of Microsoft, Google, and Accenture, who all explicitly mention some notion of responsibility. In the case of Microsoft, 'responsible AI' is about defining principles and putting them into practice through "governance, policy, and research" (Microsoft Corporation, 2019). These principles are fairness, reliability and safety, privacy and security inclusiveness, transparency, and accountability.[17]At Google, R-AI is similarly cashed out in terms of specific principles (fairness, interpretability, privacy, safety and security), and there is a similar call for these principles to be translated into "best practices".[18]Meta also set up a R-AI team in 2019, and the company has a list of "pillars of responsible AI": privacy and security, fairness and inclusion, robustness and safety, transparency and control, and accountability and governance.[19]

As noted above, there was already a plethora of academic literature on AI ethics by the time these and other companies started to develop their own internal R-AI teams and guidelines, so it seems it was a choice to go with the 'R-AI' label. One possible reason for why the R-AI frame made sense from a business perspective is, as mentioned in the introduction, its inherent, *positively valenced* normativity. To do work on AI in a 'responsible' manner is good, and further questioning regarding the *nature* of that good, its specific content, how it might relate to practice, etc. can be left undefined.

'Ethics', by contrast, suggests an open-endedness and undecidedness that 'responsible' does not. When we do ethics, we are, for example, trying to figure out the difference between right and wrong, how to live a good life, and what we mean by 'justice'. These are not questions that are easy to answer, and sometimes we would do well to admit that our questions themselves might be misguided in some important way. 'AI ethics', therefore, has a far broader remit than R-AI, and is concerned with more foundational questions, such as what it means to be human in the face of AI. These are questions that private companies might not have an interest in answering (or asking).

---

17 Microsoft have also recently partnered with UNESCO, and committed to promoting UNESCO's Recommendation on the Ethics of AI

18 While having these guidelines in place is better than having none at all, there is little point to them if they do not result in any meaningful change. Out of 24 large technology firms who had AI guidelines in place, only half had introduced any concrete steps to put these 'guidelines' into practice (De Laat, 2021, p. 1147). Worries about 'ethics washing' seem to therefore be warranted (Bietti, 2020).

19 https://ai.meta.com/responsible-ai/ (accessed 30 January 2024).

> *We can therefore see the emergence of a R-AI frame that extends beyond a narrow concept of responsibility in technical products and starts to include more socially informed understandings of AI-systems."*

Additionally, the sense of 'responsible' that was adopted seemed to draw on understandings of the term that were more comfortable to employ in purely *technical* environments. While applying 'ethics' to AI might be thought to imply the need for social and humane expertise in value theory, and even political expertise in the requirements of justice and fundamental rights, 'responsibility' might be taken to imply something far less demanding, expensive, and contested – namely, the ability to follow the right processes and use the right tools, like checklists, impact assessments, fairness benchmarks and explainability kits (Lima *et al*., 2022). The assumption here is that only (or mostly) technical expertise, i.e., expertise in creating AI-systems, is sufficient. The 'responsible' part could be picked up and layered in along the way, so long as the guidelines and principles were there and could be referred to.

With the publication of the many competing R-AI principles, toolkits and documents from 2018 onward, we can see the story start to become a bit messier. Not only because of the different perspectives being represented in all these proposals, but also due to the sometimes overlapping and ambiguous language used. One example of this comes from the European AI HLEG mentioned earlier. One of their key contributions was that AI should be 'trustworthy'[20], and this framing proved to be incredibly popular with policymakers and academics. However, there are worries that the term 'trustworthy AI' might conflate the meanings of 'trusted' and 'trustworthy' by inappropriately anthropomorphising AI-systems (Stix, 2022), and that 'trustworthy' and 'responsible' may be conflated in policy guidelines (Reinhardt, 2023).

Nevertheless, the 'trustworthy' frame remains a popular one, both in and outside of policy (Liao and Sundar, 2022). Significantly, it is in dialogue with debates in this literature that the R-AI frame also started to emerge. Part of the reason for this is that in order for AI to be 'trustworthy' in any meaningful sense, it is important that its development and implementation be 'responsible'. Trust has to be earned, and one way to do so is to ensure that the development and deployment of one's products is in line with various normative principles and practices.

In the context of emerging technology, it could mean that the development and deployment of a technology are fair, non-biased, explainable, transparent, accountable, etc. To achieve these values, however, requires a combination of technical, moral, and social expertise.

---

20 If I can 'trust' a product, service, or company, it means I can rely on them in some important sense. Trusting someone's testimony, for example, means that I can rely on what they have said (Hills, 2023, p. 744). Although many accounts of trust in the philosophical literature deal with reliability, it is widely acknowledged that trust is 'reliance plus X' (and X is usually cashed out as some kind of motivation) (Hills, 2023), whether X is a commitment (Hawley, 2014), good will (Baier, 1986), or that I be moved positively by the idea that you depend on me (Jones, 1996).

We can therefore see the emergence of a R-AI frame that extends beyond a narrow concept of responsibility in technical products and starts to include more socially informed understandings of AI-systems. In the HLEG report, responsible AI development "aims to benefit, empower and protect both individual human flourishing and the common good of society" (AI HLEG, 2019). 'Responsible', here, is not only about the technology, but also about the social compatibility of the technological innovation.

There was also continued growth in the 'ethics of AI', 'ethical AI' and 'AI ethics' in this period. The focus of this scholarship was no longer on abstract worries about machine intelligence, but on various ethical concerns that were being generated by commercially released AI systems. Parts of this literature remained quite 'academic', by, for example, engaging with questions already raised by Johnson and Maner in the 1980s: Do these new technologies introduce 'unique' moral problems, or merely new variants of old ones? This query was applied to a number of concepts: surveillance and manipulation (Klenk and Hancock, 2019; Susser, Roessler and Nissenbaum, 2019), opacity and bias (Obermeyer *et al*., 2019), automation and employment, and the potential for autonomous systems (especially in the military context) (Santoni De Sio and Van Den Hoven, 2018).

However, it was also during this time that we see more collaborations between industry, academics, and policymakers. As we have already seen, from 2016 onwards there was a rise in industry contributions to this discourse. Companies like Google, Microsoft, and DeepMind were doing both fundamental research and looking to influence policy (De Laat, 2021). DeepMind (a Google subsidiary) set up the 'DeepMind Ethics & Society' research unit. The goal of this unit was to "conduct interdisciplinary research that brings together experts from the humanities, social sciences and beyond, along with voices from civil society and technical insights from our team at DeepMind" (Legassick and Harding, 2017).

We can also see the broadening of R-AI to include more interdisciplinary and multidisciplinary perspectives, with DeepMind explicitly referencing the humanities as a source of legitimate expertise in the domain of AI. 'Sociotechnical' approaches to AI were also gaining traction, with an emphasis on the idea that these systems are composed of both technical and social properties (Selbst *et al*., 2019; Sartori and Theodorou, 2022; Weidinger *et al*., 2023; Kudina and Van De Poel, 2024).

> *One way of tracking this interdisciplinarity is to look at some of the major conferences that started during this period."*

One way of tracking this interdisciplinarity is to look at some of the major conferences that started during this period. One example is the ACM Conference on Fairness, Accountability and Transparency (ACM FAccT). Running since 2018, FAccT is labelled as a computer science conference, but from the start has been far more than that. Just looking at that inaugural conference gives us a sense of the diversification beyond the computing field. The first keynote, Latanya Sweeney, had expertise in the *governance* of technology, while the second keynote, Deborah Hellman, was a law professor. At that 2018 FAccT conference Joy Buolamwini and Timnit Gebru presented their ground-breaking 'Gender Shades' paper on the underperformance of facial recognition technologies on women, darker skin tones, and (intersectionally) on darker-skinned women (Buolamwini and Gebru, 2018).[21] Since then, Buolamwini has gone to produce multi- and inter-disciplinary exhibitions, using artistic modalities to mobilise a sharp critique of AI systems.

Mimi Onuoha, in a similar vein, has explored a number of artistic mediums (data, video, print, code, and archival media) to question and expose the underlying assumptions and operations of technological progress through algorithms and datafication. For example, her 2016 'The Library of Missing Datasets,' with its metal cabinets of empty file folders with labels like 'employment statistics that include those in federal prisons' or 'publicly available gun trace data,' is a powerful interactive reminder of those community interests, perspectives and values that datafication routinely neglects. Like Buolamwini, Crawford, Joler and Paglen's work, Onuoha's influence on the Responsible AI ecosystem extends far beyond the art world, and has been referenced in research presented at FAccT and other venues.

FAccT has matured over time into the premier venue for cutting edge research in R-AI, bringing both technical and social expertise together. Many papers presented at FAccT have been collaborations across and between industries, sectors, and academic disciplines. Conferences such as FAccT and the equally interdisciplinary AIES (Artificial Intelligence, Ethics, and Society, run by AAAI/ACM) are therefore important hubs not only for the research they produce, but also for understanding the composition of the research community that feeds into the R-AI ecosystem, and the vital role of the arts, humanities and social sciences in that composition.

An additional piece to this puzzle is the increased involvement of non-profits in publishing and doing high impact work related to R-AI. Organisations such as the Ada Lovelace Institute (funded by the Nuffield Foundation, started in 2018) and The Alan Turing Institute (founded in 2015)[22] in the UK and the AI Now Institute in the US (founded in 2017) were (and still are) addressing real, practical issues arising from AI development. For example, AI Now advised the US Federal Trade Commission on AI in 2021, and some of these institutes are working together.

---

21 For an account of how surveillance technologies are informed by the history of racial formation, see (Browne, 2015).

22 Originally a centre for data science, the Institute added AI to their remit in 2017.

In 2021, AI Now and Ada Lovelace Institute (ALI) partnered with Open Government Partnership (OGP) to produce a report that evaluated the "first wave" of algorithmic accountability policy in the public sector (AI Now Institute, Ada Lovelace Institute, and Open Government Partnership, 2021). For an organisation like ALI (a core delivery partner of the BRAID programme, with the University of Edinburgh), R-AI has a strong participatory component. Much of their work and research goes into understanding who is affected by developments in AI research, and ensuring that "data and AI work for people and society" (Ada Lovelace Institute, 2021, 2023a). That is, their work is concerned with influencing policy and practice around AI.

The Alan Turing Institute (ATI), which has a core AI research focus, also has a strong focus on Equality, Diversity, and Inclusivity (EDI), and asserts that EDI is essential to responsible AI research (The Alan Turing Institute, 2021). For example, ATI aims to assess the "impacts of data and AI on society by placing diversity, inclusion, human rights and the law at the core of responsible research, innovation and governance" (The Alan Turing Institute, 2021, p. 14). These guidelines are not unicorns: since 2016 there has been a concerted effort to integrate and operationalise EDI principles in R-AI research (Cachat-Rosset and Klarsfeld, 2023). Whether this has been successful in practice is of course a separate question.[23] Towards the end of this period we also observe private interest in academic research in a more material sense: Facebook (now Meta) announced in 2019 that they would be funding a new Institute for Ethics in Artificial Intelligence at the Technical University of Munich (TUM) in a deal worth 7.5 million USD over 5 years.[24]The increase in funding opportunities around R-AI would accelerate over the next few years.

Towards the end of this period we also observe private interest in academic research in a more material sense: Facebook (now Meta) announced in 2019 that they would be funding a new Institute for Ethics in Artificial Intelligence at the Technical University of Munich (TUM) in a deal worth 7.5 million USD over 5 years"

23 Another issue here is who is writing up these guidelines: Cachat-Rosset and Klarsfeld found that of "the 120 identified authors for the 46 selected guidelines, we found that 60.8% are male, 74.2% are white and 78.3% are nationals of Western countries (from North America, Europe or Australia)" (2023: 737).

24 https://about.fb.com/news/2019/01/tum-institute-for-ethics-in-ai/ (accessed 20 February 2024)

This has been a tumultuous period in R-AI's history, characterized by the dismantling of R-AI teams at private companies, increased government and private funding for R-AI research, and the simultaneous fracturing of that research into newly polarized communities."

At their best, these endeavours resulted in the rich intellectual history outlined here being mobilised in the service of just technological development. At its worst, however, the inherently normative nature of R-AI can be used as a cover for disruptive and exploitative practices by private companies.[25] By 2019 more than 80 R-AI guidelines had been made public, and so one of the key tests going forward for these principles is their ability to result in meaningful changes on the ground and in practice (Jobin, Ienca and Vayena, 2019). The subsequent years would see this tension come to the fore, as public, private, and academic interest in the regulation and responsible governance of AI started to merge, with multi-sector and stakeholder approaches becoming increasingly popular.

2020 – Present

This has been a tumultuous period in R-AI's history, characterised by the dismantling of R-AI teams at private companies, increased government and private funding for R-AI research, and the simultaneous fracturing of that research into newly polarised communities. The start of the tumult was marked by the very public dismissal by Google in late 2021 of Timnit Gebru[26], co-leader of their ethical AI team.

Her dismissal was in part due to conflict over a paper she had co-authored (Bender *et al*., 2021), where Google executives apparently determined it did not meet internal scholarly standards, though Gebru and her co-authors challenged that narrative and claimed that it was a shield for Google's discomfort with the paper's conclusions about the ethical risks and social impacts of large language models, the technology that would shortly become Google's primary AI product focus. It has been noted by many scholars and observers that each of the ethical concerns anticipated by that paper, colloquially known by the shortened title 'Stochastic Parrots,' have now materialised in commercial LLMs and their applications.

Following Gebru's dismissal, Google carried out a restructuring of their R-AI teams[27], and ended up firing the other lead of their ethical AI team, Margaret Mitchell (another co-author

25 For example, OpenAI, in their charter, explicitly state that they want to "ensure it [AI] is used for the benefit of all, and to avoid enabling uses of AI or AGI that harm humanity or unduly concentrate power". They say this while simultaneously exploiting labourers in Kenya (Perrigo, 2023). Perhaps this is due to the framing: the 'AI' might be responsible even if those who develop and deploy it are not.

26 https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/ (accessed 20 February 2024)

27 https://www.theverge.com/2021/2/18/22289264/google-restructuring-ethical-ai-team-timnit-gebru-firing (accessed 20 February 2024)

on the Stochastic Parrots paper, who had vigorously and publicly defended the work and Gebru). Additionally, both Microsoft[28] and Meta[29] dismantled their R-AI teams. At Meta most of the R-AI expertise was moved into the generative AI product team, while at Microsoft the entire 'Ethics and Society' team was laid off (Microsoft did, however, maintain a dedicated 'Office of Responsible AI').[30]

These shifts in industry to deprioritise R-AI investment were in contrast to massive financial investment in AI research from philanthropists and private industry, which has included some notable commitments of R-AI related funding. For example, the Schwarzman Centre for the Humanities, where the Institute for Ethics in AI is now hosted at the University of Oxford, is the product of a 185 million GBP gift from Stephen A. Schwarzman.[31] Schmidt Futures, led by the former CEO of Google Eric Schmidt and his wife, Wendy, have committed a total of 400 million USD to enable the development of AI, including funding for 160 postdoctoral fellowships at nine universities.[32]

Academic research on R-AI has continued, with the field addressing conceptual questions at the heart of 'responsibility' (Constantinescu *et al*., 2021; Himmelreich and Köhler, 2022; Gogoshin, 2024; Oimann and Tollon, 2024) and 'moral status' (Gunkel, 2018; Friedman, 2022), but also taking a 'practical turn' investigating the ways that principles and theories can be put to work in AI development and deployment (Hagendorff, 2022).[33]

These shifts in industry to deprioritize R-AI investment were in contrast to massive financial investment in AI research from philanthropists and private industry, which has included some notable commitments of R-AI related funding."

28 https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs (accessed 20 February 2024).

29 https://www.theverge.com/2023/11/18/23966980/meta-disbanded-responsible-ai-team-artificial-intelligence (accessed 20 February 2024).

30 Despite this, companies were still developing R-AI guidelines. For example, Accenture define R-AI as "the practice of designing, developing, and deploying AI with good intention to empower employees and businesses, and fairly impact customers and society—allowing companies to engender trust and scale AI with confidence" (Accenture, 2024).

31 https://www.schwarzmancentre.ox.ac.uk/ (accessed 10 October 2024)

32 https://www.forbes.com/sites/michaeltnietzel/2022/10/26/schmidt-futures-will-invest-additional-148-million-in-artificial-intelligence-research/ (accessed 10 October 2024)

33 There has also been work in 'sustainable AI' and 'AI for sustainability' (Van Wynsberghe, 2021), with Bolte and van Wynsberghe even suggesting that sustainable AI presents a 'structural shift' in how we do AI ethics (Bolte and Van Wynsberghe, 2024).

There has also been increased public funding for research into AI, including R-AI, with UK Research and Innovation releasing funding of £117 million to 12 Centres for Doctoral Training in AI in 2023, including the Responsible NLP CDT at the University of Edinburgh, where UKRI's £15.9 million BRAID programme (which funded this study) is also based.[34] Additionally, there have been increasing efforts by governments to establish national 'AI strategies' that emphasise voluntary industry commitments to responsible AI development and self-regulation, such as a 'pro-innovation approach' in the UK and the 'risk management framework' in the US (National Institute of Standards and Technology, 2023; Secretary of State for Science, Innovation and Technology, 2023).

A technical breakthrough during this period that drove this explosion of public interest and investment in AI was the release and adoption of 'Generative AI' systems, such as OpenAI's GPT-3 (Generative Pre-trained Transformer) in early 2020, and ChatGPT in 2022 (Floridi and Chiriatti, 2020, p. 3). These are built upon Large Language Models (LLMs) and are capable of generating 'novel' content (from unvetted training data) in response to user prompts. By means of LLMs as well as new 'diffusion' models capable of handling other modalities than text, generative AI outputs can now range from text (ChatGPT, Copilot, Bard, LLaMA) to images (DALL-E, Midjourney, Stable Diffusion), audio (AudioCraft) and video (Sora) as well as multimodal input and output.

While 'Generative AI' is not that new, with attempts to use AI-systems to create art going back to the 1950's[35], what makes these systems disruptive is both the technical breakthroughs[36] that led to their new performance capabilities, now widely viewed as comparable to skilled humans, and their expansive commercial adoption across industries such as marketing, software development, entertainment, education, art, and journalism. However, in addition to open questions about the reliability and true capabilities of these models, there have been serious copyright concerns regarding the training data used to create them, as there is a lack of attribution and compensation for those who created the original works in the corpus that form the basis of the training data (Samuelson, 2023).

There are many other concerns with this kind of technology, including skyrocketing environmental impacts from LLMs heavy compute demands; the enabling of deepfake porn and political disinformation; the automation and devaluation of labour in creative industries and higher education; wide dissemination of fabricated falsehoods or dangerous advice polluting the information environment; manipulation and deceptive chatbots; and outputs that perpetuate or amplify toxic content and harmful cultural, racial, class, gender and disability stereotypes and bias. For a more exhaustive list see (Bird, Ungless and Kasirzadeh, 2023).

With the massive investment, labour market disruption, and publicity associated with AI (and especially generative AI), narratives around R-AI have now taken centre stage. One of the key public fault lines in the field to emerge in recent years has been the question of whether

34 https://www.ukri.org/news/ukri-invests-in-the-next-generation-of-ai-innovators/ (accessed 21 November 2024)

35 Such as Harold Cohen's 'AARON' in 1972, which consisted of a number of computer programs that Cohen developed throughout his life.

36 Such as general adversarial networks and advances in transformer techniques (Vaswani *et al.*, 2017).

> *With the massive investment, labour market disruption, and publicity associated with AI (and especially generative AI), narratives around R-AI have now taken centre stage."*

to focus R-AI efforts on the present and near-term harms that preoccupy most AI ethicists, or the long-term, speculative AI risks that are central to work in what is now called 'AI safety.' A key issue in the AI safety debates has been worries about so called 'existential risk' (x-risk) associated with AI. Such worries were brought into the media and political spotlight by those who argue that AI systems might cause catastrophic events which result in the unrecoverable collapse of human society (Bostrom, 2014; Ord, 2020).[37] This perspective heavily influenced work on 'AI Safety', a research program that aims to reduce the probability of (certain) catastrophic outcomes due to AI by focussing on issues like the hypothetical prospect of artificial superintelligence (ASI) that escapes human control (Vold and Harris, 2021; Kasirzadeh, 2024).

In contrast to this kind of approach, most R-AI and AI ethics researchers still emphasise the more immediate and shorter-term risks associated with AI (failures of accountability, harmful bias, discrimination, deepfakes, fraud and disinformation) and focus on making vulnerable groups and publics safer from AI's already evident and growing harms (Obermeyer *et al*., 2019; Abebe *et al*., 2020; Birhane, 2021a; Crawford, 2021; Gebru *et al*., 2021). While both of the aforementioned camps can be included under the umbrella of R-AI, as both seek to reduce AI's harms and secure its benefits, what has been observed in practice are deep divides between these two groups, especially in terms of which issues to prioritise and the methods and resources that should be drawn on to deal with risks (Lazar and Nelson, 2023).

While the UK in 2023 established an 'AI Safety Institute' (AISI), focussed on "developing sociotechnical infrastructure" for safe AI, there is widespread disagreement as to what 'safety' means, and what the most urgent risks from AI really are (Lazar and Nelson, 2023; 'Introducing the AI Safety Institute', 2023; Kasirzadeh, 2024; Vallor and Luger, 2023). The 2025 renaming of AISI as the AI Security Institute, and its rebranding as focussed on threats to national security rather than social health and human wellbeing, has only deepened the ambiguity of its link to R-AI. These and other moves by the US and UK in 2025 to deprioritise AI safety research *and* mitigation of near-term risks have arguably given members of the wider R-AI community renewed reason to work together for the common cause: AI that is used to secure, rather than destroy, the chances for human flourishing – now and in the future.

---

37 For a detailed philosophical analysis of 'existential risks' see (Torres, 2023).

SECTION THREE

Looking Backward to Go Forward: Seven Lessons from the First Waves of Responsible AI

The historical review of the Responsible AI landscape outlined above is a story of how we arrived at this critical moment for the Responsible AI ecosystem, and the diverse bodies of knowledge and communities of practice that have been built along the way. Yet this ecosystem is still immature, riven with divisions and imbalances, and far from stable or flourishing. This is an ideal time to take stock of what we have learned thus far, and consolidate and disseminate that knowledge more widely across the ecosystem so that it can inform policy and practice. What lessons can those of us working toward a thriving Responsible AI ecosystem take from these 'first waves' of Responsible AI research, practice, and advocacy, to carry forward into the future?

Responsible AI since 2017 has matured and developed. It has also, however, changed and become entwined with various other labels. We have outlined the different communities and practices that have shaped and steered this process, and R-AI today is a broad and rich frame for a number of distinct practices and methods of research. However, as our study notes, this richness and plurality also brings ambiguity with regards to the appropriate use of the term, and the meanings and values that are commonly associated with it.

Our analysis employed an overarching ecological metaphor to bring some conceptual unity to this contested ground, yet deep divides remain. Given today's unprecedented proliferation and impact of AI-driven technologies across political, social, economic, and environmental spheres, it is essential that we consider how a contested and divided R-AI ecosystem can develop and mature in a healthy way, so that it may meet the growing social need for a wiser and more considered appraisal of how these technologies ought to be developed and deployed.

This moment of uncertainty and tension, therefore, offers those of us who advocate for a responsible AI ecosystem in the UK a unique opportunity to reflect on what lessons we have learned from the 'first waves' of R-AI research and practice, and how to carry those lessons further. These lessons must be informed not only by the history of R-AI since it was expressly named in 2017, but also by the broader historical reflections we have offered on the past century of human attempts to infuse new technologies with an *ethos* of responsibility.

Seven lessons in particular stand out:

1. The 'AI' in R-AI is an elusive and rapidly moving target

AI is not one thing, but many: a diverse and sprawling body of products, services, techniques and knowledge that ranges from merely speculative philosophical possibilities like AGI/ASI, to bespoke tools for narrow research tasks, to commercial systems already deployed at global scale. Virtually all AI products and services today rely upon machine learning techniques that employ statistical methods, but beyond this, AI's capabilities, limitations, and risk/benefit profiles vary widely across different AI products, applications, and models.

Different system architectures will give rise to both different testing methods and different ethical concerns. For example, the ethical concerns that arise from a large-language-model used to enable a generative AI chatbot look very different from those arising from a machine learning tool trained to predict the changing migration patterns of local wildlife.

Moreover, the path taken by AI developments is full of rapid, surprising leaps and turns, as well as unexpected lulls and lags; even 'expert' predictions of AI development timescales and impacts are historically poor. Uncertainty, ambiguity and opacity haunt any attempt to reliably model the AI ecosystem; due to the complexity of AI supply chains and deployment models, it is not even consistently possible to identify who are the 'developers, 'users' and 'stakeholders' in AI-systems. All of these features combine to make it hard to draw generalisable lessons about roles, practices, and responsibilities for all 'AI'.

Experienced R-AI practitioners have learned that we need to be very clear and explicit in how we define and distinguish between (different) users and developers for a given AI tool or system, without assuming that these categories are straightforward or static. We have also learned to consider the non-linear development of particular AI products, services, and models, and how this influences our ability to hold different stakeholders in various AI supply-chains accountable.

We have learned that complex feedback loops between AI research, AI products and services, and other social systems mean that how the technology is perceived, used and experienced

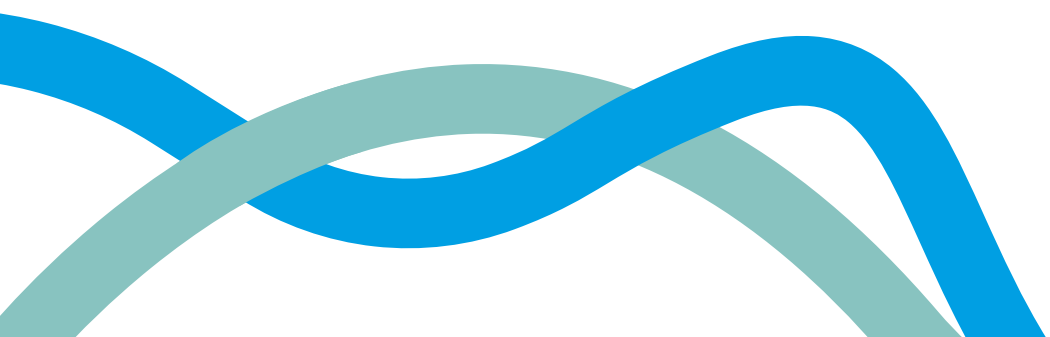
by different communities will shape its success or failure, and its impacts, in ways that were not apparent at the design or deployment stage, and that can in turn rapidly alter any social context in which AI is used. This also highlights the powerful role that social and cultural forces can play in AI development, for example by setting up sites of resistance, or constructing alternative and influential narratives of what the future with AI might be like. These same forces can drive the technology to be reconfigured for destructive *or* constructive social ends.

Two practical strategies for R-AI can be driven by this lesson. First, *R-AI must grow beyond fixed interventions such as ethical 'frameworks' and technical 'toolkits.'* While these will remain of some use to researchers and practitioners, we need to invest in more flexible, context-embedded and responsive approaches to R-AI, such as post-deployment incident reporting and monitoring, that build in capacity to course-correct.

A narrow upstream focus on the system's specifications, intended use cases, and performance capabilities will often leave developers and users unprepared for the actual impact of the technology, and unable to anticipate and mitigate its harms in time. BRAID's funding of scoping calls and demonstrator projects that emphasise the need to develop, test and evaluate R-AI interventions in well-defined application contexts is one reflection of this strategy.

Second, we must recognise that *R-AI practitioners who can draw on broad communities of social, cultural and political expertise will be better placed than narrow, siloed disciplinary experts to understand what any particular manifestation of AI is doing in the world and what it might become*. In order to become sufficiently sensitive and responsive to the wider social dynamics that shape the perception of AI more generally, and the development and impact of specific AI products and services, the R-AI ecosystem needs to build in new kinds of expertise to development pipelines and governance processes – expertise not only in ethics, but in complex social and cultural dynamics and systems thinking.

For example, post-deployment reporting of discrete incidents of harm is not enough; we also need better ways to measure the more diffuse impact of AI deployments on wider social dynamics, communities and institutions, from the creative industries and education to social care and democratic health. Incorporating insights from the arts and humanities, as well as the social sciences, is an essential mechanism to make these interventions robust, as reflected in BRAID's strategy and funding priorities from AHRC. However, this is no trivial task; bridging traditional disciplinary and experiential divides between different knowledge communities remains an immense challenge for enabling a healthy R-AI ecosystem.



2. R-AI must expand stakeholder reach to include impacted communities

When considering the 'who' of R-AI, that is, the stakeholders to whom we need the AI ecosystem to be responsible, we need to think further than the 'end user,' further even than companies, governments, and civil society organisations. The first wave of Responsible AI work, through both its successes and its failures, made clear that vulnerable groups and affected communities – who are often heavily impacted by AI even when they are not 'users' of the technology or involved at any stage of its development – must be more directly and fully brought into the fold of R-AI work. Heeding their experiences and perspectives at every stage, from ideation, design and prototyping to the marketing, deployment, evaluation, post-deployment monitoring and governance of AI products and services, is often the difference between impactful and meaningful R-AI efforts and those that collapse into ethics-washing.

For example, no set of corporate Responsible AI principles or toolkits succeeded in aligning the rollout of generative AI tools with the interests of artists, designers and writers; indeed, our early scoping research (by CREAATIF) suggests that these tools may already be harming creatives in the UK more than they are helping them; a finding reinforced by recent public protests by UK artists who fear the destruction of their livelihoods and surrender of their creative rights (Glynn, 2025). Yet many in the arts and culture sector are enthusiastic and highly skilled users or even developers of digital tools, and would welcome the opportunity to shape AI that supports their practice. For example, the BRAID-funded project *Performance, Participation, Provenance and Reward (P3R) in responsible AI*, led from the University of London, will work with UK musicians to create tools that give them new ways to receive remuneration and credit for live performances. Of course, it might not be feasible or reasonable to have the same level of participatory engagement for all types of AI products and services, but for those that are likely to have a foreseeably large impact on people, it is essential that affected communities and vulnerable groups are brought into the fold; otherwise, R-AI risks being a wasted exercise.

Going forward, the challenge for UK advocates of a R-AI ecosystem is to *take this lesson to heart in ways that go beyond mere lip service to equitable and participatory AI futures, and identify the bridges and interventions needed to align the incentives of powerful actors in the AI ecosystem with the interests of those groups and communities whose lives are starkly affected by their work.*

3. Narrowly technical approaches to R-AI do not work

Narrow, technical approaches to AI ethics and safety tend to fall short of the ambitions for a responsible AI ecosystem for several reasons. First, as noted under the first lesson, siloed technical perspectives are detached from a view of the wider social, political and environmental contexts, and local contexts of application, that powerfully shape what any given AI artifact does in the world, and how safe or beneficial it is or can be. By focusing narrowly on system capabilities and design choices, technical experts routinely overlook how an AI tool that works well for the imaginary 'average user' can have devastatingly unsafe or unjust consequences when it lands in the real world and touches real human beings, especially members of marginalised and vulnerable communities. R-AI researchers and practitioners have acquired a vast trove of evidence for this pattern, which has recurred again and again with AI tools deployed for facial recognition, predictive policing, fraud detection, healthcare triaging, judicial risk assessment, hiring and countless other applications.

Second, we have learned that narrowly technical approaches often focus exclusively on quantitative methods of modelling and assessing the safety and performance of AI tools, which by design omits qualitative evidence of risk and harm, such as the testimony of those who currently experience it. Narrowly technical approaches disproportionately centre harms that can easily be counted – deaths, physical injuries, property damage – and decentre or ignore harms not easily given a numerical value, such as psychological damage, reputational harm, or injustice. They likewise favour artificial measures of social harms, like fairness or toxicity 'benchmarks' for datasets, which often do a poor job of tracking the actual harms in the world. They also may delay recognition of emerging safety and responsibility issues until there is a large enough sum of incidents to measure and track by quantitative methods.

Third, R-AI practitioners have grappled with the dominance among technical experts of deterministic thinking about technology. 'Techno-determinism' is a term used to characterise the (false) belief among many technologists that advanced technologies lock us in to certain futures to which we must adapt, and that the technologies themselves have a kind of agency and direction that humans are powerless to change. This way of thinking leads to the erroneous conclusion that certain developments in AI are 'inevitable'.

It also encourages a dangerously constricted approach to addressing AI harms, since the power of human agency and social choice to drive change is largely discounted. By framing only model

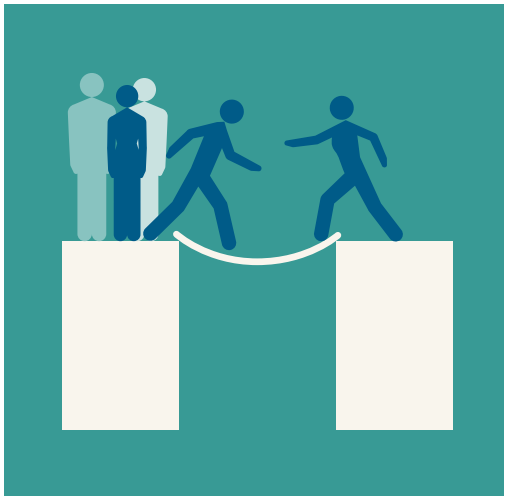
sizes, capabilities, features and datasets as possible sites of intervention, narrow technical safety methods systematically omit the wider role of human agency and choice – notably including regulatory options – in shaping AI's trajectory and impacts. Moreover, the growing political sway of techno determinist framings fosters passive deference and 'learned helplessness' in leaders and institutions that might otherwise use their cultural and political power to effectively govern and steer technological innovation toward lasting societal benefit.

We can both enrich and challenge technical approaches to AI safety and responsible AI with vital perspectives from the arts and humanities, which exemplify and articulate the ways in which human reflection, narrative understanding, and cultural critique can be forces for social good that powerfully shape and even steer technological change. 21st century advances in clean energy tech, sustainable agriculture and environmental sensing were all kicked into gear by the force of the late 20th century environmental movement to value and protect planetary health, a cultural movement led not just by scientists but equally by filmmakers, writers, artists and activists.

Likewise, artistic and humanistic perspectives can instigate change and highlight the spaces in which human agency can most effectively meet the challenges of AI. But the arts and humanities need to be integrated as far *more* than just critical voices from the 'outside' of the AI ecosystem's centres of power. The BRAID-funded 'Anticipating Today' fellowship, by positioning the arts and humanities as central to R-AI, alongside and directly engaged with technical expertise, will develop tools to aid policymakers in grasping the full scope of societal impacts that follow from specific AI technologies. An even broader range of integrated technical, moral, and cultural expertise drives the BRAID-funded project 'Responsible use of AI in the creation, archiving, reactivation and conservation of artworks and their archives', led from the University of Nottingham. To address the thorny challenges around using AI for the preservation of artistic heritage, the project will unite experts from the fields of philosophy, computer science, dance performance, creative writing, English, performance and new media, media art and archives, archival studies and emerging technologies.

Moving forward, then, *it has become clear that sociotechnical approaches to Responsible AI are required. This means going beyond looking only at the technical specifics or capabilities of an AI product or service: AI evaluation needs to include social context, human interaction, cultural values and systemic impacts if it is to be truly responsible* (Sartori and Theodorou, 2022; Weidinger *et al*., 2023; Kudina and Van De Poel, 2024). The power of human agency to shape and drive AI development can only be harnessed if we create spaces for a richer chorus of voices to be heard, and this will involve broadening the scope of how we evaluate AI products and services.

4. Public trust is essential to a sustainable R-AI ecosystem



Recent years have brought into sharp relief what can happen when technological 'progress' is viewed with increasing scepticism by wide swathes of society. Consider, for example, the effect of a tide of public health disinformation and distrust of medical experts on the effectiveness and uptake of vaccines, and the recent appointment of anti-vaccination voices to US federal leadership. Or consider the political effect of climate denial and skepticism on delaying the necessary investments in infrastructure and technology needed to safeguard billions of lives.

For AI development to be sustainable (not only ecologically, but democratically), it needs to reckon with this staggering decline in public trust in the science and technology we require to flourish, which has indeed recently begun to turn upon AI (Ada Lovelace Institute, 2023b; Centre for Data Ethics and Innovation, 2024; Tyson and Kikuchi, 2028). If we want AI to be a widely adopted tool for meeting challenges in healthcare, energy and other sectors critical to human welfare, then we need to work to earn this trust back before it is too late.

We cannot rely solely on the visions of tech entrepreneurs to attain this goal; their unbridled enthusiasm and optimism for AI cannot replace the need for democratic forms of meaningful and effective public engagement with AI development, deployment and governance. This is not to say that there is a simple relationship between making a technology more transparent and a resultant increase in public trust. 'Democratising AI' is not a matter of making AI available to all, or putting its power under majority rule. It's about legitimising the power of AI in society, a legitimacy which in democratic societies must be earned from and granted by the wider publics who are subject to that power.

The BRAID funded 'Inclusive Futures' fellowship aims to amplify the voices of marginalised communities through the use of participatory methods to shape future visions of AI. These future visions are important when we reflect on the wide economic and social gulf between those who benefit the most from AI systems in the Global North, and those who will bear the brunt of the environmental and political costs of AI systems in the Majority World.

To move forward toward a healthy and sustainable R-AI ecosystem, then, it is essential that we provide alternative narratives as to what a trustworthy configuration of AI in society looks like. This will involve interrogating the democratic legitimacy of AI, and making sure that there is effective public engagement in the design and development of AI-systems.

5. Good intentions are not enough for R-AI

We cannot only rely on the motivation of virtuous individuals to build and sustain a responsible AI ecosystem. Responsible AI practitioners have learned the hard way that strong organisational incentives and systems of accountability must be in place to enable, motivate and reward R-AI work and decision-making. While individual moral values and moral judgment are vital to responsible AI development and governance, at the system and organisational level, we know that incentives and structural forces have larger effects that can swamp and negate the social impact of individual ethical goals and decisions.

For example, in a company where the leader of an AI product team can be promoted only by rushing an AI system to market over her team's vocal objections and concerns about its safety risks, and the executives and board members above her can only be rewarded by a short-term jump in the stock price, the shared ethical intentions of the team to produce a safe and beneficial tool will not prevent a harmful outcome. A regulatory agency or oversight body which does not provide effective incentives for integrity and rigor in AI auditing practices may find itself captured by powerful interests even if that is no one's intention or desire.

Moreover, due to the non-linear nature of AI-development and deployment outlined earlier, merely having good intentions is not enough to ensure that a particular AI-system will not create or contribute to real-world harms once deployed. Responsible AI is not about individual morality, but about creating systemic alignment of AI innovation with sustainable and shared human flourishing. That ambition demands moving beyond individual efforts to soften the harmful impacts of particular systems, to look at the collective impact of AI on our societies and the planet. For example, 'Muted Registers' is a BRAID fellowship that reframes 'red-teaming' practices in R-AI away from a narrow focus on harm reduction to include more 'hopeful' ways of responding to the challenges raised by AI – responses that ask: 'what other worlds are possible?'. Another example of this broader R-AI ambition is the BRAID-funded Sustainable AI Futures project led at Bath Spa University, which aims to pioneer innovative approaches to responsibly governing AI's rapidly growing environmental impacts.

Going forward, therefore, we need to ensure that the structures and incentives that govern the R-AI ecosystem not only reward individual efforts to act responsibly, but sustain broad cultures of responsible practice that can build worlds with AI worth wanting for everyone. The conditions under which practitioners and researchers in R-AI operate have to be understood and mapped so that we can ensure the long-term sustainability of the entire field, and allow it to infuse the entire AI ecosystem rather than remaining a corrective patch. This means being responsive and attentive to what we take to be the longer-term ambitions of R-AI, and building communities capable of nurturing this ambition.

6. R-AI must address questions wider than ethics and legality

Despite nearly a decade of sustained efforts to bring ethics into the centre of AI development, use and governance, we remain a long way from a responsible AI ecosystem. Nor has the force of laws – even existing ones, like copyright protections – yet been enough to turn the tide of AI development in a responsible direction. We have learned from these failures that R-AI must be built to address the wider political, economic, environmental and cultural forces and dynamics that drive the AI ecosystem.

While legal compliance remains an important part of R-AI, it alone is not enough. The law, while certainly informed by ethics in some cases, is nonetheless distinct from it and thus can only take us so far. Our understanding of ethics itself is shaped by wider social and environmental changes, which affect our needs and vulnerabilities and hence alter our judgments of what we owe to one another. Our values are driven not merely by existing legal or ethical norms but by cultural powers of aesthetic and moral imagination that envision new or expanded human capacities and ambitions, and give rise to creative acts of design that enable new ways of living together.

Moreover, AI *itself* can steer these social forces, norms and values, in constructive or destructive ways. It can be used to expand and boost our creative and moral power to imagine and build better futures, or it can be used to devalue, automate and displace those powers. It can add to our scientific and democratic understanding of what is possible, or it can pollute the information and media landscape with falsehoods that prevent us from accurately perceiving where we stand, where we are heading, or what we can do together. These are powers that other risky technologies like nuclear power and geoengineering do not have.

Responsible AI, then, must be a field that brings together knowledge of a more diverse, dynamic and expansive sort than any other form of technology governance. More than any other, it must incorporate and make use of insights from social science to tell us where we are with AI, and the arts and humanities to help us determine where and how far we want to go with it. This doesn't just mean that those working in academic disciplines housed in the arts and humanities have input into AI development and governance (although this is of course vital). What it also means is that those involved in the creation of cultural artefacts and the expression of humanistic insights and ambitions are given platforms to shape broader human values and attitudes toward AI.

Works of art, for example, can be vehicles through which cultural and ethical values are expressed, transformed, or subverted, and thus offer powerful platforms for public critique of the status quo, as well as tools of moral and social imagination for illuminating new and better paths to human flourishing. Innovation, when it is not mindless but intelligent, must be guided by that illumination. The BRAID-funded fellowship 'AI Art Beyond the Gallery', delivered in partnership with Serpentine Galleries, is a case in point. This project will investigate the ways in which AI art can have influence beyond the gallery, and more specifically, the impact it can have on government policy. It is a poignant irony, then, that it is precisely the irresponsible deployment of AI systems today that undermines and endangers this power of the arts and humanities to guide innovation onto sustainable and humane paths. Responsible AI must not only use, but safeguard that power.

Our BRAID-funded fellowship, 'Centering Creativity and Responsibility', investigates what artists want from AI, and how AI tools can be made to serve those in the creative industries rather than the other way around. And our new artist commissions give artists new ways for their work to reach directly into the R-AI ecosystem, without requiring intermediation by other kinds of R-AI 'experts.'

To build and sustain a responsible AI ecosystem, therefore, we have to understand the place of AI within the broader cultural, economic, and political value system that supports it. While ethics and the law are important parts of this system, there are myriad other social and material powers that shape the possibilities presented and the values amplified by AI, and we must engage them all to understand and steer AI in ways that serve the interests of our planet and its people.

> *To build and sustain a responsible AI ecosystem, therefore, we have to understand the place of AI within the broader cultural, economic, and political value system that supports it"*

7. R-AI is not a problem to be solved but an ecosystem to be tended

It has been almost a decade since the ambition of Responsible AI began to be explicitly articulated, and many more decades since we started thinking about how to develop technologies ethically, responsibly and sustainably. In this time a great deal has been accomplished, both intellectually and practically; and yet we are no closer to a 'solution'. Indeed, it has become clear that Responsible AI does not name a problem to be solved, but is part of a wider social project of keeping the built world aligned with and supportive of human flourishing. For this reason, it cannot be *finished*. We cannot 'solve' R-AI like a math or engineering problem, and so it is always a work-in-progress.

This means that there is no set of guidelines or policies that will eradicate the need for the careful cultivation and stewardship of the R-AI ecosystem, keeping both the short and the long-term state of human flourishing in mind. While policy proposals and guidelines have their place and use, we need more high-level reflection on the ultimate goals of AI and the particular forms it should take in our societies.

Additionally, R-AI practitioners have learned that we also need bottom-up approaches that seek to build and sustain responsible communities of AI practice. High-level perspectives, while useful for outlining the general normative orientation of R-AI, need to be buttressed with everyday norms for ordinary work that are necessary to sustain and normalise a responsible configuration of AI in society.

BRAID will have impact by helping the UK's AI ecosystem absorb all 7 of the lessons outlined above – not by providing 'solutions' to 'problems' but by intervening in ways that weave together social, moral, creative and technical expertise to inform and guide responsible innovation. What has been learned within and across the many diverse R-AI communities of research, practice, governance and advocacy, is better understood as wisdom. It is what gives us the tools, strategies, and visions we need to bridge the divides that currently characterise the R-AI ecosystem. Only the sharing of practical wisdom, not the empty promise of solutions, can facilitate the collective decision-making we need in order to bring that ecosystem into balance and sustain its health over the long term.

Conclusion

The history of R-AI contains many lessons, some rooted in the early origins of 20th century thinking about ethics and technology, others only recently made evident by the shape of our fragile, immature, but rapidly growing ecosystem of research, practice, advocacy and governance that seeks to earn wider public trust in the development and use of AI technologies.

That ecosystem must be safely shepherded onto a trajectory of sociotechnical maturity and sustainability. Given the speed at which AI is transforming the very social fabric that supports it, the diverse and sprawling global community of R-AI researchers, practitioners, creators and leaders must widely and effectively share the lessons they have learned already, and apply them to the new challenges for R-AI already on our doorstep.

This first part of our BRAID landscape study is a preliminary step toward that end. In Part 2 of the study, to be completed in 2025, we'll hear from members of the R-AI community in their own voices, sharing their own wisdom. We'll hear their best ideas for how to bridge the divides and barriers in the R-AI ecosystem that stand in the way of us collectively implementing its lessons, and learning new ones.

References

Abebe, R. *et al*. (2020) **'Roles for Computing in Social Change'**, *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 252–260. Available at: https://doi.org/10.1145/3351095.3372871.

Accenture (2024) **'Responsible AI: Scale AI with Confidence'**. Available at: https://www.accenture.com/gb-en/services/applied-intelligence/ai-ethics-governance (Accessed: 15 January 2024).

Ada Lovelace Institute (2021) **'Ada Lovelace Strategy 2021-24'**. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2021/10/Ada-Strategy-2021-24-single-pages-for-print.pdf (Accessed: 20 January 2024).

Ada Lovelace Institute (2023a) **'Inclusive AI governance Civil Society Participation in Standards Development'**. Available at: https://www.adalovelaceinstitute.org/report/inclusive-ai-governance

Ada Lovelace Institute (2023b) **'What do the Public Think About AI?'** Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/evidence-review/what-do-the-public-think-about-ai/ (Accessed: 28 November 2024).

AI HLEG (2019) **'Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence'**. Available at: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf

AI Now Institute, Ada Lovelace Institute, and Open Government Partnership (2021) **Algorithmic Accountability for the Public Sector – Report**. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/ (Accessed: 24 January 2024).

Anderson, M. and Anderson, S.L. (2007) **'Machine Ethics: Creating an Ethical Intelligent Agent'**, AI Magazine, 28(4), pp. 15–26.

Angwin, J. *et al*. (2016) **'Machine Bias'**, ProPublica, 23 May. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Baier, A. (1986) **'Trust and Antitrust'**, Ethics, 96(2).

Beckman, L., Hultin Rosenberg, J. and Jebari, K. (2024) **'Artificial Intelligence and Democratic Legitimacy. The Problem of Publicity in Public Authority'**, AI & SOCIETY, 39(3), pp. 975–984. Available at: https://doi.org/10.1007/s00146-022-01493-0

Beer, D. (2022) **'The Problem of Researching a Recursive Society: Algorithms, data coils and the looping of the social'**, Big Data & Society, 9(2), p. 205395172211049. Available at: https://doi.org/10.1177/20539517221104997

Bender, E.M. *et al*. (2021) **'On the Dangers of Stochastic Parrots: Can Language Models be too Big?'**, FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922

Bietti, E. (2020) 'From Ethics Washing to Ethics Bashing', Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), pp. 210–219.

Bird, C., Ungless, E. and Kasirzadeh, A. (2023) 'Typology of Risks of Generative Text-to-Image Models', in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. AIES '23: AAAI/ACM Conference on AI, Ethics, and Society, Montr\'{e}al QC Canada: ACM, pp. 396–410. Available at: https://doi.org/10.1145/3600211.3604722.

Birhane, A. (2021a) **'Algorithmic Injustice: a Relational Ethics Approach'**, Patterns, 2(1), pp. 1–9. Available at: https://doi.org/10.1016/j.patter.2021.100205

Birhane, A. (2021b) **'The Impossibility of Automating Ambiguity'**, Artificial Life, 27, pp. 1–18.

Birhane, A. *et al*. (2022) **'The Forgotten Margins of AI Ethics'**, in 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea: ACM, pp. 948–958. Available at: https://doi.org/10.1145/3531146.3533157

Bolte, L. and Van Wynsberghe, A. (2024) **'Sustainable AI and the Third Wave of AI Ethics: a Structural Turn'**, AI and Ethics. Available at: https://doi.org/10.1007/s43681-024-00522-6

Borenstein, J. *et al*. (2021) **'AI Ethics: A Long History and a Recent Burst of Attention'**, Computer, 54(1), pp. 96–102. Available at: https://ieeexplore.ieee.org/document/9321834

Bostrom, N. (2014) Superintelligence. Oxford: Oxford University Press.

Brennan, J. *et al*. (2023) **AI assurance? Assessing and Mitigating Risk Across the AI Lifecycle**. Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/risks-ai-systems/

Brey, P. (2008) **'Do We Have Moral Duties Towards Information Objects?'**, Ethics and Information Technology, 10(2–3), pp. 109–114. Available at: https://doi.org/10.1007/s10676-008-9170-x

Brey, P. and Dainow, B. (2024) **'Ethics by Design for Artificial Intelligence'**, AI and Ethics, 4(4), pp. 1265–1277. Available at: https://doi.org/10.1007/s43681-023-00330-4

Browne, S. (2015) **Dark Matters: On the Surveillance of Blackness**. Durham: Duke University Press.

Buolamwini, J. and Gebru, T. (2018) **'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification'**, Proceedings of Machine Learning Research, 81(Conference on Fairness, Accountability, and Transparency), pp. 1–15. Available at: https://doi.org/10.2147/OTT.S126905

Busby, J.S. and Coeckelbergh, M. (2003) **'The Social Ascription of Obligations to Engineers'**, Science and Engineering Ethics, 9(3), pp. 363–376. Available at: https://doi.org/10.1007/s11948-003-0033-x

Bynum, T.W. (2001) **'Computer Ethics: Its birth and its Future'**, Ethics and Information Technology, 3, pp. 109–112.

Bynum, T.W. (2008) **'Milestones in the History of Information and Computer Ethics'**, in K.E. Himma and H.T. Tavani (eds) The Handbook of Information and Computer Ethics. 1st edn. Wiley, pp. 25–48. Available at: https://doi.org/10.1002/9780470281819.ch2

Cachat-Rosset, G. and Klarsfeld, A. (2023) **'Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines'**, Applied Artificial Intelligence, 37(1), p. 2176618.
Available at: https://doi.org/10.1080/08839514.2023.2176618

Carayannis, E.G. *et al*. (2021) **'Social Business Model Innovation: A Quadruple/Quintuple Helix-Based Social Innovation Ecosystem'**, IEEE Transactions on Engineering Management, 68(1), pp. 235–248. Available at: https://ieeexplore.ieee.org/document/8720229

Carrier, M. (2021) **'How to Conceive of Science for the Benefit of Society: Prospects of Responsible Research and Innovation'**, Synthese, 198(S19), pp. 4749–4768. Available at: https://doi.org/10.1007/s11229-019-02254-1

Carson, R. (2002) **Silent Spring**. 50. anniversary ed., 1. Mariner Books ed. Boston: Mariner Books, Houghton Mifflin Harcourt.

Centre for Data Ethics and Innovation (2024) **'Public attitudes to data and AI: Tracker survey (Wave 3)'**. UK Government. Available at: https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-3/public-attitudes-to-data-and-ai-tracker-survey-wave-3#foreword

Century, M. (2022) **Northern Sparks**: Innovation, Technology Policy, and the Arts in Canada from Expo 67 to the Internet age. Cambridge, Massachusetts: The MIT Press (Leonardo).

Cihon, P., Maas, M.M. and Kemp, L. (2020) **'Should Artificial Intelligence Governance be Centralised? Design Lessons from History'**. arXiv. Available at: http://arXiv.org/abs/2001.03573 (Accessed: 1 July 2024).

Collingridge, D. (1980) The Social Control of Technology. London: Frances Pinter Limited. Available at: https://doi.org/10.2307/1960465

Constantinescu, M. *et al*. (2021) **'Understanding Responsibility in Responsible AI. Dianoetic Virtues and the Hard Problem of Context'**, Ethics and Information Technology, 23(4), pp. 803–814.
Available at: https://doi.org/10.1007/s10676-021-09616-9

Crawford, K. (2021) Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press.

Crawford, K. and Joler, V. (2018) **'Anatomy of an AI System'**. Available at: https://anatomyof.ai/ (Accessed: 17 July 2024).

Crawford, K. and Joler, V. (2023) **'Calculating Empires'**. Available at: https://calculatingempires.net (Accessed: 10 June 2024).

Criddle and Murgia (2023) **'Big Tech Compnanies Cut AI Ethics Staff, Raising Safety Concerns'**, Financial Times. Available at: https://www.ft.com/content/26372287-6fb3-457b-9e9c-f722027f36b3 (Accessed: 3 March 2025).

Danaher, J. (2016) **'Robots, Law and the Retribution Gap'**, Ethics and Information Technology, 18(4), pp. 299–309. Available at: https://doi.org/10.1007/s10676-016-9403-3

Dastin, J. (2018) **'Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women'**, Reuters, 10 October. Available at: https://doi.org/10.1201/9781003278290-44

De Freitas Netto, S.V. *et al*. (2020) **'Concepts and Forms of Greenwashing: a Systematic Review'**, Environmental Sciences Europe, 32(1), p. 19. Available at: https://doi.org/10.1186/s12302-020-0300-3

De Laat, P.B. (2021) **'Companies Committed to Responsible AI: From Principles Towards Implementation and Regulation?'**, Philosophy & Technology, 34(4), pp. 1135–1193. Available at: https://doi.org/10.1007/s13347-021-00474-3

Dignum, V. *et al*. (2018) **'Ethics by Design: Necessity or Curse?',** in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18: AAAI/ACM Conference on AI, Ethics, and Society, New Orleans LA USA: ACM, pp. 60–66. Available at: https://doi.org/10.1145/3278721.3278745

Dignum, V. (2019) Responsible Artificial Intelligence. Cham: Springer Nature Switzerland.
Available at: https://doi.org/10.1007/978-3-030-30371-6

Douglas, H. (2003) **'The Moral Responsibilities of Scientists (Tensions between Autonomy and Responsibility)'**, American Philosophical Quarterly, 40(1).

Drage, E., McInerney, K. and Browne, J. (2024) **'Engineers on Responsibility: Feminist Approaches to who's Responsible for Ethical AI'**, Ethics and Information Technology, 26(1), p. 4. Available at: https://doi.org/10.1007/s10676-023-09739-1

Dubber, M.D., Pasquale, F. and Das, S. (2020) **The Oxford Handbook of Ethics of AI**. New York: Oxford university press (Oxford handbooks).

Eitel-Porter, R. and Grosskopf, U. (2022) **'From AI Compliance to Competitive Advantage: Becoming Responsible by Design'**. Accenture.

Ellul, J. (1964) The Technological Society. Toronto: Random House of Canada.

European Commission (2020) **'European Commission White Paper on Artificial Intelligence: A European Approach to Excellence and Trust'**. European Commission.
Available at: https://commission.europa.eu/document/d2ec4039-c5be-423a-81ef-b9e44e79825b_en
(Accessed: 6 August 2024).

European Parliament (2016) General Data Protection Regulation. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04

Fazelpour, S. and Lipton, Z.C. (2020) **'Algorithmic Fairness from a Non-ideal Perspective'**, AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, (i), pp. 57–63. Available at: https://doi.org/10.1145/3375627.3375828

Fleddermann, C.B. (2012) Engineering Ethics. 4th ed. Upper Saddle River: Prentice Hall (ESource).

Floridi, L. (1999) **'Information Ethics: On the Philosophical Foundation of Computer Ethics'**, Ethics and Information Technology, 1, pp. 37–56. Available at: https://doi.org/10.1023/A:1010018611096

Floridi, L. *et al.* (2020) **'How to Design AI for Social Good: Seven Essential Factors'**, Science and Engineering Ethics, 26(3), pp. 1771–1796. Available at: https://doi.org/10.1007/s11948-020-00213-5

Floridi, L. and Chiriatti, M. (2020) **'GPT-3 : Its Nature, Scope, Limits, and Consequences'**, Minds and Machines, 30, pp. 681–694. Available at: https://doi.org/10.1007/s11023-020-09548-1

Formosa, P. and Ryan, M. (2021) **'Making Moral Machines: Why we Need Artificial Moral Agents'**, AI & SOCIETY, 36(3), pp. 839–851. Available at: https://doi.org/10.1007/s00146-020-01089-6

Friedman, B., Hendry, D.G. and Borning, A. (2017) **'A Survey of Value Sensitive Design Methods'**, Foundations and Trends® in Human–Computer Interaction, 11(2), pp. 63–125. Available at: https://doi.org/10.1561/1100000015

Friedman, C. (2022) **'Ethical Concerns with Replacing Human Relations with Humanoid Robots : An Ubuntu Perspective'**, AI and Ethics , (0123456789). Available at: https://doi.org/10.1007/s43681-022-00186-0

Frodeman, R. and Mitcham, C. (2000) **'Beyond the Social Contract Myth'**, Issues in Science and Technology, 16(4).

Gebru, T. *et al.* (2021) **'Datasheets for datasets'**, Communications of the ACM, 64(12), pp. 86–92. Available at: https://doi.org/10.1145/3458723

Global Index on Responsible AI (2024). **Global Center on AI Governance**. Available at: https://www.global-index.ai/ (Accessed: 7 August 2024).

Glynn, P. (2025) **'Artists Release Silent Album in Protest Against AI Using Their Work'**, BBC, 25 February. Available at: https://www.bbc.co.uk/news/articles/cwyd3r62kp5o (Accessed: 3 March 2025).

Gogoshin, D.L. (2024) **'A Way Forward For Responsibility In The Age Of AI'**, Inquiry, Pp. 1–34. Available at: https://Doi.Org/10.1080/0020174X.2024.2312455

Gotterbarn, D. (1991) **'Computer Ethics: Responsibility Regained'**, National Forum: The Phi Beta Kappa Journal, 71, pp. 26–31.

Gotterbarn, D., Miller, K. And Rogerson, S. (1997) **'Software Engineering Code Of Ethics'**, Communications Of The ACM, 40(11), Pp. 110–118. Available at: https://Doi.Org/10.1145/265684.265699

Grunwald, A. (2011) **'Responsible Innovation: Bringing together Technology Assessment, Applied Ethics, and STS research'**, Responsible Innovation.

Gunkel, D.J. (2018) **Robot Rights**. Cambridge, Massachusetts: MIT Press. Available at: https://doi.org/10.1126/science.323.5916.876a

Guston, D.H. and Sarewitz, D. (2002) **'Real-time technology assessment'**, Technology in Society .

Gyevnár, B. And Kasirzadeh, A. (2025) **'AI Safety For Everyone'**, Nature Machine Intelligence, 7: Pp. 531-542. April 2025. Available at: https://Doi.Org/10.1038/S42256-025-01020-Y

Hagendorff, T. (2022) **'A Virtue-Based Framework To Support Putting AI Ethics Into Practice'**, Philosophy & Technology, 35(3), P. 55. Available at: https://Doi.Org/10.1007/S13347-022-00553-Z

Haraway, D. (2006) **'A Cyborg Manifesto: Science, Technology, And Socialist-Feminism In The Late 20th Century'**, In J. Weiss *et al.* (Eds) The International Handbook Of Virtual Learning Environments. Dordrecht: Springer Netherlands, Pp. 117–158. Available at: https://Doi.Org/10.1007/978-1-4020-3803-7_4

Hawley, K. (2014) **'Trust, Distrust And Commitment'**, Nous, 48(1).

Hess, G. (2010) **'The Ecosystem: Model Or Metaphor? Epistemological Difficulties In Industrial Ecology'**, Journal Of Industrial Ecology, 14(2), Pp. 270–285. Available at: https://Doi.Org/10.1111/J.1530-9290.2010.00226.X

Hills, A. (2023) **'Trustworthiness, Responsibility And Virtue'**, The Philosophical Quarterly, 73(3), Pp. 743–761. Available at: https://Doi.Org/10.1093/Pq/Pqad036

Himmelreich, J. And Köhler, S. (2022) **'Responsible AI Through Conceptual Engineering'**, Philosophy & Technology, Pp. 1–30. Available at: https://Doi.Org/10.1007/S13347-022-00542-2

'Introducing The AI Safety Institute' (2023). GOV.UK. Available at: https://www.aisi.gov.uk https://www.Gov.Uk/Government/Publications/Ai-Safety-Institute-Overview/Introducing-The-Ai-Safety-Institute (Accessed: 13 February 2024).

Ipsos (2023) **'Debating Responsible AI: The UK Expert View'**. Available at: https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-11/debating-responsible-ai-the-uk-expert-view-ipsos.pdf (Accessed: 4 March 2025).

Irani, L. (2023) **'Algorithms Of Suspicion: Authentication And Distrust On The Amazon Mechanical Turk Platform'**, SSRN Electronic Journal . Available at: https://Doi.Org/10.2139/Ssrn.4482508

Jasanoff, S. (2000) **'Reconstructing The Past, Constructing The Present: Can Science Studies And The History Of Science Live Happily Ever After?'**, Social Studies Of Science, 30(4), Pp. 621–631.

Jobin, A., Ienca, M. And Vayena, E. (2019) **'The Global Landscape Of AI Ethics Guidelines'**, Nature Machine Intelligence, 1(9), Pp. 389–399. Available at: https://Doi.Org/10.1038/S42256-019-0088-2

Joerges, B. (1999) **'Do Politics Have Artefacts?'**, Social Studies Of Science, 29(3), Pp. 411–431. Available at: https://Doi.Org/10.1177/030631299029003004

Johnson, D.G. (1985) **Computer Ethics**. 2nd Edn. One Lake Street Upper Saddle River, NJ, United Sta: Prentice-Hall, Inc.

Jonas, H. (1973) **'Technology And Responsibility: Reflections On The New Tasks Of Ethics'**, Social Research, 40(1), Pp. 37–47. Available at: https://Doi.Org/10.1057/9781137349088_3

Jonas, H. (1984) **The Imperative Of Responsibility**. Chicago: University Of Chicago Press.

Jones, K. (1996) **'Trust As An Affective Attitude'**, Ethics, 107(1), Pp. 4–25.

Kasirzadeh, A. (2024) **'Two Types Of AI Existential Risk: Decisive And Accumulative'**. arXiv. Available at: http://arxiv.org/abs/2401.07836 (Accessed: 12 February 2024).

Keyes, O., Hutson, J. And Durbin, M. (2019) **'A Mulching Proposal: Analysing And Improving An Algorithmic System For Turning The Elderly Into High-Nutrient Slurry'**, In Extended Abstracts Of The 2019 CHI Conference On Human Factors In Computing Systems. CHI '19: CHI Conference On Human Factors In Computing Systems, Glasgow Scotland Uk: ACM, Pp. 1–11. Available at: https://doi.org/10.1145/3290607.3310433

Klenk, M. And Hancock, J. (2019) **'Autonomy And Online Manipulation'**, Internet Policy Review, Pp. 2–5.

Knight, W., Paresh, D. And Feiger, L. (2025) **'The National Institute Of Standards And Technology Braces For Mass Firings'**, Wired, 20 February. Available at: https://www.wired.com/story/the-national-institute-of-standards-and-technology-braces-for-mass-firings/ (Accessed: 3 March 2025).

Köhler, S., Roughley, N. And Sauer, H. (2017) **'Technologically Blurred Accountability?'**, In C. Ulbert *et al.* (Eds) Moral Agency And The Politics Of Responsibility. London: Routledge, Pp. 1–19.

Kudina, O. And Van De Poel, I. (2024) **'A Sociotechnical System Perspective On AI'**, Minds And Machines, 34(3), Pp. 21, S11023-024-09680–2. Available at: https://doi.org/10.1007/s11023-024-09680-2

Lazar, S. And Nelson, A. (2023) **'AI Safety On Whose Terms?'**, Science, 381(6654), Pp. 138–138. Available at: https://doi.org/10.1126/science.adi8982

Legassick, S. And Harding, V. (2017) **'Why We Launched Deepmind Ethics & Society'**, Google Deepmind, 3 October. Available at: https://deepmind.google/discover/blog/why-we-launched-deepmind-ethics-society/ (Accessed: 15 January 2024).

Liao, Q.V. And Sundar, S.S. (2022) **'Designing For Responsible Trust In AI Systems: A Communication Perspective'**, In 2022 ACM Conference On Fairness, Accountability, And Transparency. FAccT '22: 2022 ACM Conference On Fairness, Accountability, And Transparency, Seoul Republic Of Korea: ACM, Pp. 1257–1268. Available at: https://doi.org/10.1145/3531146.3533182

Liao, S.M. (Ed.) (2020) **Ethics Of Artificial Intelligence**. New York, NY: Oxford University Press.

Lima, G. *et al.* (2022) **'The Conflict Between Explainable And Accountable Decision-Making Algorithms'**, In 2022 ACM Conference On Fairness, Accountability, And Transparency. FAccT '22: 2022 ACM Conference On Fairness, Accountability, And Transparency, Seoul Republic Of Korea: ACM, Pp. 2103–2113. Available at: https://doi.org/10.1145/3531146.3534628

Maner, M. (1980) **Starter Kit In Computer Ethics**. Hyde Park, NY: Helvetia Press And The National Information And Resource Center For Teaching Philosophy (3).

Maner, W. (1996) **'Unique Ethical Problems In Information Technology'**, Science And Engineering Ethics, 2(2), Pp. 137–154. Available at: https://doi.org/10.1007/bf02583549

Mckeon, C. (2025) **'Rebranded AI Security Institute To Drop Focus On Bias And Free Speech'**, The Standard, 14 February. Available at: https://www.standard.co.uk/business/business-news/rebranded-ai-security-institute-to-drop-focus-on-bias-and-free-speech-b1211109.html (Accessed: 3 March 2025).

McQuillan, D. (2022) **Resisting AI: An Anti- Fascist Approach To Artificial Intelligence**. Great Britain: Bristol University Press.

Menon, A.K. And Williamson, R.C. (2018) **'The Cost Of Fairness In Binary Classification'**, Proceedings Of Machine Learning Research, 81(Conference On Fairness, Accountability, And Transparency), Pp. 1–12.

Microsoft Corporation (2019) **'Empowering Responsible AI Practices'**. Available at: https://www.microsoft.com/en-us/ai/responsible-ai (Accessed: 11 October 2023).

Milmo, D. And Courea, E. (2025) **'US And UK Refuse To Sign Paris Summit Declaration On "Inclusive" AI'**, The Guardian, 11 February. Available at: https://www.theguardian.com/technology/2025/feb/11/us-uk-paris-ai-summit-artificial-intelligence-declaration (Accessed: 3 March 2025).

Moor, J.H. (1985) **'What Is Computer Ethics?'**, Metaphilosophy, 16(4), Pp. 266–275. Available at: https://doi.org/10.1111/j.1467-9973.1985.tb00173.x

Moore, J.F. (1993) **'Predators And Prey: A New Ecology Of Competition'**, Harvard Business Review, Pp. 75–86.

Muller, V.C. (2022) **'The History Of Digital Ethics'**, In Carissa Veliz (Ed.) Oxford Handbook Of Digital Ethics. Oxford University Press.

National Institute Of Standards And Technology (2023) **'Artificial Intelligence Risk Management Framework'**. Available at: https://doi.org/10.6028/nist.ai.100-1

Nichols, S. And Weldon, W. (1997) **'Professional Responsibility: The Role Of The Engineer In Society'**, Science And Engineering Ethics, 3(3), Pp. 327–337.

Nyholm, S. (2018) **'Attributing Agency To Automated Systems: Reflections On Human–Robot Collaborations And Responsibility-Loci'**, Science And Engineering Ethics, Pp. 1–19. Available at: https://doi.org/10.1007/s11948-017-9943-x

Obermeyer, Z. *et al.* (2019) **'Dissecting Racial Bias In An Algorithm Used To Manage The Health Of Populations'**, Science, 366(6464), Pp. 447–453. Available at: https://doi.org/10.1126/science.aax2342

OECD (2019) **'Recommendation Of The Council On Artificial Intelligence'**. OECD Legal Instruments. Available at: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449 (Accessed: 8 August 2024).

Oimann, A.-K. And Tollon, F. (2024) **'Responsibility Gaps And Technology: Old Wine In New Bottles?'**, Journal Of Applied Philosophy .

OpenAI (2018) **'OpenAI Charter'**, 9 April. Available at: https://OpenAI.com/charter (Accessed: 11 October 2023).

Ord, T. (2020) **The Precipice: Existential Risk And The Future Of Humanity**. London New York (N.Y.): Bloomsbury Academic.

O'Reilly, T. (2007) **'What Is Web 2.0: Design Patterns And Business Models For The Next Generation Of Software'**, COMMUNICATIONS & STRATEGIES. Edited By M. Mandiberg, 65, Pp. 32–52. Available at: https://doi.org/10.18574/nyu/9780814763025.003.0008

Paresh, D. (2024) **'Google Splits Up A Key AI Ethics Watchdog'**, Wired, 31 January. Available at: https://www.wired.com/story/google-splits-up-responsible-innovation-ai-team/ (Accessed: 3 March 2025).

Perrigo, B. (2023) **'Exclusive: OpenAI Used Kenyan Workers On Less Than $2 Per Hour To Make Chatgpt Less Toxic'**, Time, 18 January. Available at: https://time.com/6247678/OpenAI-chatgpt-kenya-workers/ (Accessed: 9 November 2023).

Peters, D. *et al*. (2020) **'Responsible AI-Two Frameworks For Ethical Design Practice'**, Ieee Transactions On Technology And Society, 1(1), Pp. 34–47. Available at: https://doi.org/10.1109/tts.2020.2974991

Pichai, S. (2018) **'AI At Google: Our Principles'**, 7 June. Available at: https://blog.google/technology/ai/ai-principles/ (Accessed: 11 October 2023).

Prichard, M. (1998) **'Professional Responsibility: Focusing On The Exemplary'**, Science And Engineering Ethics, 4(2).

Queer in AI, O.O. *et al*. (2023) **'Queer In AI: A Case Study In Community-Led Participatory AI'**, In 2023 ACM Conference On Fairness, Accountability, And Transparency. FAccT '23: The 2023 ACM Conference On Fairness, Accountability, And Transparency, Chicago IL USA: ACM, Pp. 1882–1895. Available at: https://doi.org/10.1145/3593013.3594134

Reinhardt, K. (2023) **'Trust And Trustworthiness In AI Ethics'**, AI And Ethics, 3(3), Pp. 735–744. Available at: https://doi.org/10.1007/s43681-022-00200-5

Sadek, M. *et al*. (2024) **'Challenges Of Responsible AI In Practice: Scoping Review And Recommended Actions'**, AI & SOCIETY . Available at: https://doi.org/10.1007/s00146-024-01880-9

Samuelson, P. (2023) **'Generative AI Meets Copyright'**, Science, 14 July.

Santoni De Sio, F. And Van Den Hoven, J. (2018) **'Meaningful Human Control Over Autonomous Systems: A Philosophical Account'**, Frontiers In Robotics And AI, 5, P. 15. Available at: https://doi.org/10.3389/frobt.2018.00015

Sartori, L. And Theodorou, A. (2022) **'A Sociotechnical Perspective For The Future Of AI: Narratives, Inequalities, And Human Control'**, Ethics And Information Technology, 24(1), Pp. 102–104. Available at: https://doi.org/10.1007/s10676-022-09624-3

Von Schomberg, R. (2011) **'Introduction: Towards Responsible Research And Innovation In The Information And Communication Technologies And Security Technologies Fields'**, In R. Von Schomberg (Ed.) Towards Responsible Research And Innovation In The Information And Communication Technologies And Security Technologies Fields. Luxembourg: Publications Office Of The European Union. Available at: https://doi.org/10.1007/978-94-015-8143-1_1

Von Schomberg, R. (2013) **'A Vision Of Responsible Research And Innovation'**, In R. Owen, J. Bessant, And M. Heintz (Eds) Responsible Innovation. 1st Edn. Wiley, Pp. 51–74. Available at: https://doi.org/10.1002/9781118551424.ch3

Schot, J. And Rip, A. (1997) **'The Past And Future Of Constructive Technology Assessment'**, Technological Forecasting And Social Change, 54(2–3), Pp. 251–268. Available at: https://doi.org/10.1016/s0040-1625(96)00180-1

Secretary Of State For Science, Innovation And Technology (2023) **A Pro-Innovation Approach To AI Regulation**. London, UK: Department For Science, Innovation & Technology.

Selbst, A.D. *et al*. (2019) **'Fairness And Abstraction In Sociotechnical Systems'**, In Proceedings Of The Conference On Fairness, Accountability, And Transparency. FAT* '19: Conference On Fairness, Accountability, And Transparency, Atlanta GA USA: ACM, Pp. 59–68. Available at: https://doi.org/10.1145/3287560.3287598

Shanley, D. (2021) **'Imagining The Future Through Revisiting The Past: The Value Of History In Thinking About R(R)I's Possible Future(S)'**, Journal Of Responsible Innovation, 8(2), Pp. 234–253. Available at: https://doi.org/10.1080/23299460.2021.1882748

Shanley, D. (Danielle) (2022) **Making Responsibility Matter: The Emergence Of Responsible Innovation As An Intellectual Movement**. Maastricht University. Available at: https://doi.org/10.26481/dis.20221208ds

Smuha, N.A. And Yeung, K. (2024) **'The European Union's AI Act: Beyond Motherhood And Apple Pie?'** Available at: https://doi.org/10.2139/ssrn.4874852

Sparrow, R. (2007) **'Killer Robots'**, Journal Of Applied Philosophy, 24(1), Pp. 62–78. Available at: https://doi.org/10.1111/j.1468-5930.2007.00346.x

Srinivasan, R. And Uchino, K. (2021) **'Biases In Generative Art: A Causal Look From The Lens Of Art History'**, In Proceedings Of The 2021 ACM Conference On Fairness, Accountability, And Transparency. FAccT '21: 2021 ACM Conference On Fairness, Accountability, And Transparency, Virtual Event Canada: ACM, Pp. 41–51. Available at: https://doi.org/10.1145/3442188.3445869

Stahl, B.C. (2022) **'Responsible Innovation Ecosystems: Ethical Implications Of The Application Of The Ecosystem Concept To Artificial Intelligence'**, International Journal Of Information Management, 62, P. 102441. Available at: https://doi.org/10.1016/j.ijinfomgt.2021.102441

Stahl, B.C. (2023) **'Embedding Responsibility In Intelligent Systems: From AI Ethics To Responsible AI Ecosystems'**, Scientific Reports, 13(1), P. 7586. Available at: https://doi.org/10.1038/s41598-023-34622-w

Statt, N. (2017) **'See How An AI System Classifies You Based On Your Selfie'**, The Verge, 17 September. Available at: https://www.theverge.com/tldr/2019/9/16/20869538/imagenet-roulette-ai-classifier-web-tool-object-image-recognition

Stilgoe, J., Owen, R. And Macnaghten, P. (2013) **'Developing A Framework For Responsible Innovation'**, Research Policy, 42, Pp. 1568–1580. Available at: https://doi.org/10.1002/9781118551424.ch2

Stix, C. (2022) **'Artificial Intelligence By Any Other Name: A Brief History Of The Conceptualization Of "Trustworthy Artificial Intelligence"'**, Discover Artificial Intelligence, 2(1), P. 26. Available at: https://doi.org/10.1007/s44163-022-00041-5.

Susser, D., Roessler, B. And Nissenbaum, H. (2019) **'Technology, Autonomy, And Manipulation'**, Internet Policy Review, 8(2), Pp. 1–22. Available at: https://doi.org/10.14763/2019.2.1410

Taddeo, M. And Floridi, L. (2018) **'How AI Can Be A Force For Good'**, Science, 361(6404), Pp. 751–752. Available at: https://doi.org/10.1126/science.aat5991

Taylor, K. And Woods, S. (2020) **'Reflections On The Practice Of Responsible (Research And) Innovation In Synthetic Biology'**, New Genetics And Society, 39(2), Pp. 127–147. Available at: https://doi.org/10.1080/14636778.2019.1709431.

The Ada Lovelace Institute And The Alan Turing Institute (2025) **'How Do People Feel About AI? A Nationally Representative Survey Of Public Attitudes To Artificial Intelligence In Britain'**. Available at: https://attitudestoai.uk/findings-2025 (accessed: 26 march 2025).

The Alan Turing Institute (2021) **'Equality, Diversity And Inclusion Strategy 2021-2024'**. Available at: https://www.turing.ac.uk/sites/default/files/2021-09/edi-strategy-report_final_0.pdf.

Tigard, D.W. (2021) **'Responsible AI And Moral Responsibility: A Common Appreciation'**, AI And Ethics, 1(2), Pp. 113–117. Available at: https://doi.org/10.1007/s43681-020-00009-0

Torres, P. (2023) **'Existential Risks: A Philosophical Analysis'**, Inquiry, 66(4), Pp. 614–639. Available at: https://doi.org/10.1080/0020174x.2019.1658626

Turculet, G. (2023) **'Data Feminism And Border Ethics: Power, Invisibility And Indeterminacy'**, Journal Of Global Ethics, 19(3), Pp. 323–334. Available at: https://doi.org/10.1080/17449626.2023.2278533

Turkle, S. (2005) **The Second Self: Computers And The Human Spirit**. 20th Anniversary Ed., 1st MIT Press Ed. Cambridge, Massachusetts: MIT Press.

Tyson, A. And Kikuchi, E. (2028) **'Growing Public Concern About The Role Of Artificial Intelligence In Daily Life'**, Pew Research Center, 28 August. Available at: https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/

UK Government (2022) **'National AI Strategy'**, 18 December. Available at: https://www.gov.uk/government/publications/national-ai-strategy/national-ai-strategy-html-version#pillar-1-investing-in-the-long-term-needs-of-the-ai-ecosystem (Accessed: 8 August 2024).

Umbrello, S. And Van De Poel, I. (2021) **'Mapping Value Sensitive Design Onto AI For Social Good Principles'**, AI And Ethics, 1(3), Pp. 283–296. Available at: https://doi.org/10.1007/s43681-021-00038-3

Vallor, S. And Luger, E. (2024) **'A Shrinking Path To Safety: How A Narrowly Technical Approach To Align AI With The Public Good Could Fail'**, BRAID UK, 13 October. Available at: https://braiduk.org/a-shrinking-path-to-safety-how-a-narrowly-technical-approach-to-align-ai-with-the-public-good-could-fail (Accessed: 15 October 2023).

Vallor, S. And Vierkant, T. (2024) **'Find The Gap: AI, Responsible Agency And Vulnerability'**, Minds And Machines, 34(3), P. 20. Available at: https://doi.org/10.1007/s11023-024-09674-0

Van Wynsberghe, A. (2021) **'Sustainable AI: AI For Sustainability And The Sustainability Of AI'**, AI And Ethics, 1(3), Pp. 213–218. Available at: https://doi.org/10.1007/s43681-021-00043-6

Vaswani, A. *et al.* (2017) **'Attention Is All You Need'**, 31st Conference On Neural Information Processing Systems .

Véliz, C. (Ed.) (2023) Oxford Handbook Of Digital Ethics. 1st Edn. Oxford University Press. Available at: https://doi.org/10.1093/oxfordhb/9780198857815.001.0001

Vold, K. And Harris, D.R. (2021) **'How Does Artificial Intelligence Pose An Existential Risk?'**, In C. Véliz (Ed.) Oxford Handbook Of Digital Ethics. 1st Edn. Oxford University Press, Pp. 724–747. Available at: https://doi.org/10.1093/oxfordhb/9780198857815.013.36

Von Schomberg, R. (2012) **'Prospects For Technology Assessment In A Framework Of Responsible Research And Innovation'**, In M. Dusseldorp And R. Beecroft (Eds) Technikfolgen Abschätzen Lehren. Wiesbaden: VS Verlag Für Sozialwissenschaften, Pp. 39–61. Available at: https://doi.org/10.1007/978-3-531-93468-6_2

Wallach, W. And Allen, C. (2009) **Moral Machines**. New York: Oxford University Press.

Weidinger, L. *et al.* (2023) **'Sociotechnical Safety Evaluation Of Generative AI Systems'**. arXiv. Available at: http://arXiv.org/abs/2310.11986 (Accessed: 18 November 2024).

Weizenbaum, J. (1976) Computer Power And Human Reason. New York: W.H. Freeman And Company.

Wheeler, K. (2025) **'How Trump Scrapping AI Safety Regulations Impacts Global AI'**, AI Magazine, 23 January. Available at: https://aimagazine.com/articles/trump-scraps-ai-risk-rules-what-you-need-to-know (Accessed: 3 March 2025).

Wiener, N. (1954) **The Human Use Of Human Beings**. 2nd Edn. Garden City, New York: Doubleday.

Wiener, N. (1961) **Cybernetics: Or Control And Communication In The Animal And The Machine**. Reissue Of The 1961 Second Edition. Cambridge, Massachusetts London, England: The MIT Press.

Winner, L. (1980) **'Do Artifacts Have Politics?'**, Daedalus, 109(1), Pp. 121–136.

Woolgar, S. And Cooper, G. (1999) **'Do Artifacts Have Ambivalence?'**, Social Studies Of Science, 29(3).

Van Wynsberghe, A. And Robbins, S. (2019) **'Critiquing The Reasons For Making Artificial Moral Agents'**, Science And Engineering Ethics, 25(3), Pp. 719–735. Available at: https://doi.org/10.1007/s11948-018-0030-8

Zhang, D. *et al.* (2022) **'The AI Index 2022 Annual Report'**. arXiv. Available at: http://arxiv.org/abs/2205.03468 (Accessed: 9 November 2023).

Zhu, W. (2019) **'4 Steps To Developing Responsible AI. World Economic Forum'**. Available at: https://www.weforum.org/agenda/2019/06/4-steps -to-developing-responsible-ai/ (Accessed: 16 November 2023).

Zimmermann, A., Vredenburgh, K. And Lazar, S. (2022) **'The Political Philosophy Of Data And AI'**, Canadian Journal Of Philosophy, 52(1), Pp. 1–5. Available at: https://doi.org/10.1017/can.2022.28

Acknowledgements

This work was supported by the Arts and Humanities Research Council grant number AH/X007146/1. We acknowledge that BRAID has partnerships with, or advisory board members from, the following organisations mentioned or cited in this study: Google DeepMind, Microsoft Research, Accenture, and the Ada Lovelace Institute.

Preferred Citation

Tollon, Fabio and Vallor, Shannon (2025). The Responsible AI Ecosystem: A BRAID Landscape Study. Bridging Responsible AI Divides (**www.braiduk.org**), Edinburgh.
DOI: 10.5281/zenodo.15195686

Publication Credits

Illustrations: Ian Vickers, **eureka.co.uk**
Creative Artwork: Cate Sutton, **bebedesign.co.uk**

About BRAID

BRAID is a UK-wide programme dedicated to integrating Arts and Humanities research more fully into the Responsible AI ecosystem, as well as bridging the divides between academic, industry, policy and regulatory work on responsible AI.

Funded by the Arts and Humanities Research Council (AHRC) it represents a six-year, £15.9 million investment in enabling responsible AI in the UK. The Programme runs from 2022 to 2028.

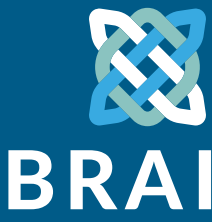
Working in partnership with the Ada Lovelace Institute and BBC, the team brings together expertise in human-computer interaction, moral philosophy, arts, design, law, social sciences, journalism, and AI.

BRAID is extended by a network of interdisciplinary researchers and partnering organisations through the delivery of funding calls, community building events, and a series of programmed activities.

Funding reference: Arts and Humanities Research Council grant number: AH/X007146/1.

Learn more at **www.braiduk.org**

To request an alternative format of this report please email **braid@ed.ac.uk**



BRAID


Arts and
Humanities
Research Council


THE UNIVERSITY
of EDINBURGH


Ada
Lovelace
Institute


BBC
R&D