European Review, Vol. 31, No. 4, 406–421

doi:10.1017/S1062798723000145

LECTURE HELD AT THE ACADEMIA EUROPAEA BUILDING BRIDGES CONFERENCE 2022
An Introduction to Responsible AI

RICARDO BAEZA-YATES

Institute for Experiential Artificial Intelligence (EAI), Northeastern University, Boston, Massachusetts 02136, USA Email rbaeza@acm.org

Artificial intelligence (AI) has finally reached most people on our planet thanks to generative AI tools for text and other media. This has started a controversy about the possible benefits and risks, where responsible AI is key. Hence, we introduce the concept of responsible AI, its relation to AI ethics, and why the terms ethical or trustworthy AI should not be used. We then cover the three main relevant aspects of this new field: principles governance and regulation.

Introduction

Artificial intelligence (AI) is increasingly the power behind who we are and who we can be if you are connected to the internet. AI observes how we move, the things we say and the decisions we make. Human beings are frail and susceptible to rash decisions or illness. Our digital selves are what can be improved by AI and, conversely, the ones that could be most at risk from entities that ignore the requirements for responsible AI.

If we ignore the ethical considerations that should be the basis for a functioning society, then there is every chance that AI becomes a cluster bomb as it can reach 5 billion people through the internet: a next-generation tech solution that can be both angel and demon. Ethics, justice and trust are human traits, so it is irrational to assume that the tools we use can either be designed to exhibit the same characteristics or, increasingly, the ethical and fairness considerations that form part of our humanity.

Responsible AI isn't a buzzword. It's a commitment to making the world a better place. From chatbots and voice–user interfaces through autonomous vehicles and the Internet of Things, the ethical questions that AI asks need to be answered, both now and in the future. Responsibility is also a human trait, but our legal system has extended it to institutions, and we use it in that way: an institution should be responsible for bad uses of AI.

In this work, we will discuss the different facets of and our expectations towards responsible AI. This will include but not be limited to what the term means, its history and evolution, the principles behind it and the ways to implement it.

We will also consider the different forms of discrimination and bias in AI systems, the challenges that developers face, the current regulatory landscape and how responsible AI benefits society. Finally, we will look at the current state of AI (including the explosion in generative AI), and the key questions that stakeholders should be asking before they use this technology in their processes.

What is Responsible AI?

With so much media attention and noise currently swirling around the AI sector, it can be easy to forget that the technology hasn't just magically appeared. It has now been 80 years since it first appeared and it became really mainstream a bit more than 10 years ago. But now ChatGPT, launched by OpenAI into the public consciousness in November 2022, has started the equivalent of an AI arms race.

Future historians will likely point to this moment as the catalyst for what came next, but the questions, challenges and implications for widespread AI system usage have been associated with the tech for years. AI has been embedded in our lives for so long that it is often hard to remember when we weren't interacting with it. Amazon's ecommerce dominance was built on AI algorithms, Netflix killed Blockbuster with its state-of-the-art recommendation engines, social media relies on AI to decide what we need to see and when, and big tech wouldn't be investing billions of dollars in pushing AI to the next level if it didn't see a favourable impact on the bottom line.

Before we can discuss responsible AI in detail, we should remind ourselves as to what the term artificial intelligence means. Pop culture fans will always conjure images of sentient computers from a dystopian future, but the reality is less SkyNet and more about automated solutions for mundane tasks.

For the purposes of this article, we are not assessing the state of AI per se. Instead, we are focused on how responsible AI can evolve and, importantly, what its future will (in theory) look like. In addition, we must factor in the concept that machine intelligence is often a mirror of humanity. That means that we need to accept that it can be flawed and prone to biases or learned behaviours.

Simply put, AI is the science and engineering behind the concept of making 'intelligent' machines. The term itself is very broad and covers multiple techniques, including the most popular one, machine learning. That includes neural networks, which when they have many layers, are called deep learning.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

In addition, although there is a temptation to think that the technology is actually intelligent, it is unlikely – in its current form, at least – to be plotting the downfall of humanity. Rather, AI tools are developed to perform tasks that seemingly require human intelligence.

The key word here is human. AI relies on data to work effectively, an integral component of every use case that its developers can dream up. In addition, AI is tied to enormous amounts of computational power and financial backing. When you look at AI under these terms, it becomes clear that the technology itself is more akin to a child being taught the difference between right and wrong as opposed to a machine that sees humanity as a threat.

It is very easy to attach convenient terms such as ethical, trustworthy and responsible to a bright and shiny new tech toy, but there are real-world implications of attaching human traits to what is essentially a machine. All three of these words have been associated with AI in the last few years, and there is a danger that the average person will assume they are interchangeable.

Ethical AI implies a degree of moral agency, a human condition that requires value judgements and intended behaviour. However, AI, in its current incarnation, is not a moral agent. Ethics is thereby something that only a human being can demonstrate, so transferring this to a machine is not a good idea.

Trustworthy AI is even more problematic. AI is not human, so we should steer away from attaching humanistic capabilities to not only an algorithm but also a machine that is designed to operate in a vacuum.

We already know that AI does not work all the time, so asking users to trust it is misleading. If, 100 years ago, someone wanted to sell people an airplane ticket calling it trustworthy aviation then the buyer should have been worried, because if something works, why do we need to add 'trustworthy' to it? That is the difference between engineering and alchemy.

Simply put, AI should not be put in either of these buckets. Both 'ethical' and 'trustworthy' are capabilities that can be applied to the human element in the process, but we run into all manner of concerns if they become part of the discourse.

For that reason alone, responsible AI is a better fit. Granted, responsibility is also another human trait, but the concept has the benefit of including ethical considerations and accountability under its umbrella. Institutions are also held responsible for the actions they take, imposing laws and regulations that prevent unethical or illegal behaviour that can harm individuals and the institution itself.

When we think about how wide the responsibility net can be cast, then attaching the right word to a nascent technology is critical. And, from an integration and implementation perspective, the need to define what responsible AI is (and isn't) becomes increasingly important in a society that has been recently upended by the explosion in generative AI as a business optimization and productivity tool.

In a recent interview with ODBMS Industry Watch (Zicari 2022), I said that responsible AI aims to create systems that benefit individuals, societies and the environment. It encompasses all the ethical, legal and technical aspects of developing

and deploying beneficial AI technologies. It includes making sure your AI system does not interfere with human agency, cause harm, discriminate or waste resources.

The caveat is that the words 'responsible AI' have taken on a life of their own. Type them into a search engine and you get 500 million hits in less than a second. Apply a filter for actual news and there are more than 30,000 results generated in the same time period. For context, it takes ChatGPT around three seconds to write a 1000-word blog post on the topic.

There is an increasing number of companies that are trying their hardest to nail down what responsible AI actually is. All of them are likely to be focusing their efforts on the principles behind the design and implementation of the system itself. We will go into these principles in more detail in the third section, but the main ones can be identified as legitimacy, fairness, transparency, accountability, robustness, privacy, security and sustainability.

All of these tick the boxes of what responsible AI can be, but we should think about what that word means. For a human being, responsibility is the moral, legal or mental accountability that we have to each other or an entity or object (Merriam-Webster dictionary). Then, by creating systems that benefit individuals and society as a whole, we must ensure that the system does not interfere with human agency, cause harm, discriminate or waste valuable resources.

Solutions should also be both technologically and ethically robust, a minimum requirement that covers everything from data to algorithms through to design and end user interfaces. In addition, there should always be accountability, not with the system per se but the humans who have control over or input into the AI itself.

The caveat to acknowledging the need for responsible AI is that we must consider what irresponsible AI looks like. Design and implementation are significant elements within the sphere of responsible AI, but even a well-designed system could incorporate illegal or unethical practices.

For instance, developers need to address the consequences of unintended harm, especially when it comes to bias and discrimination. No company is going to admit that their AI is irresponsible, especially if that impacts the brand. But you don't need to be part of the global AI community to understand that machines – much like human beings – can make mistakes. The difference here is that we often judge machines to a higher standard than humans.

Take autonomous vehicles. These cars-of-the-future are packed with AI tools and capabilities. But if one harms or kills somebody, then it is the company that sells the AI that is at fault and not the individual or team that designed that AI to, say, not recognize the pace of an elderly person as they crossed the road or a person of colour in low light. Human beings are involved in thousands of accidents every single day, and it is unlikely that there will ever be a call to ban human drivers from the roads.

AI is also part of our financial system. A bank may decide to run a person's credit through an automated assessment tool to decide whether they should be given a loan. The tool is not looking for reasons to deny that application, but it will draw on every bit of available data to make its decision. If the numbers don't add up, the loan is likely to be denied, irrespective of context and, critically, without a human being

having the final say. Again, we don't apportion blame to the institution but to the automated system that merely did its job.

These are just two examples, but there are literally thousands of others where AI has taken the critical hit, and not the human. If you want to take a deeper dive into why responsible AI matters, the AI Incident Database (https://incidentdatabase.ai/) was set up to index 'the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems'. And while that was obviously not the intent of the AI developer, the questions that the widespread adoption of potentially irresponsible AI will bring to the surface are just the tip of the iceberg. AI is currently at a critical inflection point, and the questions that are being asked are not going away.

We will cover the regulatory proposals that are currently being discussed later but it doesn't take much of a stretch to understand that the implementation of responsible AI principles should be at the top of the list. Again, it is people that are building the structures, designing the systems, and training the algorithms, so the onus shifts away from the machine and back to the humans in the loop, although I prefer to say that humans need to be in charge and the computers in the loop.

With that in mind, let's shift our attention to the principles that underlie the concept of responsible AI and the ethics that are intrinsically linked to its required adoption.

Principles and Governance

Before we address the principles and governance required for responsible AI, we should consider the here and now. Algorithmic systems are the backbone of both the digital economy and engrained automated processes, with the public face of AI – ecommerce, streaming services, chatbots and automated decision making – all widely used. This integration is likely to become even more widespread before the end of the decade.

A recent forecast by the International Data Corporation (IDC 2023) said that global spending on AI systems would hit $154 billion in 2023 alone, a compound annual growth rate (CAGR) of 26.9% from the year before. Much of that spending would be on software, hardware and services, with two industries – banking and retail – expected to account for at least 25% of total spend. In addition, IDC analysts noted that only one of the 36 use cases for AI identified within the forecast would have a CAGR of less than 24% over the next five years.

If we unpack this prediction and drill down into what it means for not only AI but also the developers, researchers and, critically, the end stakeholders in this brave new world, then it is clear that we must be very careful about how we both integrate these systems and measure their societal impact. Companies that are slow to adopt AI will be left behind, IDC said, and data-driven decisions would become the norm. That scenario raises the bar on several levels.

Let's not forget that the complexity and opacity of AI systems makes them difficult to understand. When you throw the pace of innovation into the mix,

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

ensuring that these human-trained models are used in a way that is fair, ethical, and beneficial to society is paramount.

That means that we must adhere to not only a defined set of principles, but also to a form of governance that fits with the concept of responsible AI. In a perfect world, these would align with AI-specific regulations and oversight. We will discuss regulatory activity in more detail in the fourth section, but it will not come as a shock that discussions over how AI can be regulated – a set of rules that will take on added importance in the post-ChatGPT era – can best be described as ongoing.

To fully understand why establishing a set of principles and an effective form of governance is a critical path to take, we should first ask what we want from an algorithmic system. Once we have established our level of expectation, the challenges to responsible AI that these models present can (in theory) become easier to both understand and alleviate.

An algorithm is a self-contained step-by-step set of operations used to perform calculations, data processing, and automated reasoning tasks. The key component is data. For example, most AI algorithms are based on statistical models that are 'learned' or 'trained' from datasets through machine learning (ML). This is called supervised ML. Those that don't fall under this umbrella are driven by analytics, unsupervised ML: the discovery, interpretation, and communication of meaningful patterns in data. The picture is completed with reinforcement learning – learning online – and symbolic AI, among others.

Algorithmic systems have been part of the private and public sector for many years. Companies integrate these tools into their workflow processes to make decisions about potential hires and staff assessment. Banks and financial institutions use them to assess customer spending or credit-based loans, while the digitization of healthcare has allowed algorithms to make judgement calls about patient wellness. Algorithmic decisions can be found in the criminal justice system, often without substantive review by human beings.

From an AI perspective, the increased use of and adoption into society is not breaking news. It should also not come as a surprise that efficiency and decision-based standardization are often cited as the reasons why the machine, and not the human, is the one who passes judgment.

The problem that we face is that while these systems have been designed with the utopian ideal of making society more equitable, inclusive and efficient, automation can also be subject to the same biases and discriminatory practices that are part of the human condition. Much like their human designers, algorithms can fail to respect the rights of individuals and be prone to harmful discrimination or other negative effects.

That makes it even more critical that we establish both guidelines for algorithmic systems to follow and rules that comply with established legal, ethical, and scientific norms. In addition, there is a requirement that the risks associated with their use be proportional to the specific problem that these systems have been designed to address. And while there has been a significant increase in the conversations around what these rules or guidelines should look like, the core requirements of what we should want from an algorithmic system have not changed.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

These requirements are not extensive. Common wisdom tells us that they are likely to mirror what we expect from human beings, albeit that a person has agency over decision-making, and a machine is merely adhering to what it has been taught. An algorithmic system, thereby, must have a level of transparency and accountability, an awareness of bias, access and redress, an explanation for its decisions, and date provenance. Additionally, it must display auditability, validation, and be tested in real-world scenarios. However, the algorithms and other underlying mechanisms that artificial intelligence or machine learning systems utilize to make their predictive decisions can be opaque. This has the effect of not only making these decisions less understandable but also difficult to work out whether these outputs are biased or erroneous.

This lack of transparency is an oft-cited challenge in the scientific community, especially when it comes to assessing what impact AI will have on society as a whole. Factors that contribute to these uncertain outcomes can be informational, technical, economic, competitive, and social. AI can produce unexplainable results, and it is an accepted fact that even the most well-engineered algorithmic systems will have bugs or training data that might not align with the intended use.

For instance, the informational data used to train models and create analytics could have been obtained without the data subject's knowledge or explicit consent. The cost of being transparent might not be feasible (from an economic standpoint) or the algorithm itself could be too technical to allow for easy interpretation. In addition, transparency has the potential to hand an advantage to third parties in the form of trade secrets or the ability to game the system itself. All these factors can be influential in how the system is deployed and, crucially, any problems with bias or inaccuracy in the automated decision-making process.

According to the Association for Computing Machinery (ACM 2022), there are nine core principles that developers, system builders and AI policymakers must adhere to. These can be defined as follows:

- Legitimacy and competency
- Minimizing harm
- Security and privacy
- Transparency
- Interpretability and explainability
- Maintainability
- Contestability and auditability
- Accountability and responsibility
- Limiting environmental impacts

All of these core principles were introduced by the ACM Technology Policy Council (TPC) in October 2022, and I was one of the two lead authors. These principles are designed to recognize that algorithmic systems are used by governments, private companies, and other entities to 'make or recommend decisions that have far-reaching effects on individuals, organizations, and society'.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

Additionally, the TPC has stated that a 'one size fits all' approach to responsible AI systems is unlikely to be effective, noting that the context of that system will play an integral role in its development and implementation. However, TPC also believes that the establishment of these principles will lead to further discussions among stakeholders, researchers and policymakers. These conversations will both help to chart the course of AI governance by focusing attention on what that will look like and promoting the 'reliability, safety, and responsibility of algorithmic systems'.

The caveat to the establishment of these instrumental principles is that the current state of AI has uncovered significant (and potentially game-changing) challenges. If we take the generic issues that underpin the integration of AI systems into the public sphere out of the equation – principles versus techniques, regulation, our own cognitive biases, and cultural differences – then we find that there are other roadblocks on the path to responsible AI.

These include questions over responsible versus trustworthy AI (see the Introduction), discrimination (facial recognition, justice, the sharing economy, large language models), phrenology (pseudo-scientific biometric-based predictions), unfair digital commerce (end user exposure and popularity bias), and the stupidity of the models themselves due to human incompetence.

The latter is arguably a contributing factor to how we define what it means to be responsible, especially when we are thinking about how systems are used in such a way to be fair and ethical. End users should be able to understand how AI systems work and how they make decisions. At the same time, there is an acknowledgement that these tools are intended to augment and not to replace or harm human beings.

Taking all the above into account, we need to shift our attention to AI governance. The principles set out by the TCP are a good platform to build upon, but the deploying entity must adhere to a clear governance structure when deciding to design and launch an algorithmic system. Applying the principles will lead to necessary trade-offs, and it is expected that these will be dependent on the context of the deployment, the influence of public policy and, increasingly, the market forces that incentivize the development of the systems themselves. In Figure 1, we depict how the ACM principles should be instantiated during the system lifecycle as well as the main tools that need to be used at each stage.

So, what does effective governance entail? And how can we apply a comprehensive and integrated approach to the thorny question of responsible AI itself? The answer to both to these questions can be found in how we approach the subject of ethics in an algorithmic system.

An ethical approach to problem-solving centres on one core concept; what is the right thing to do? The key to solving the issue of responsible AI, according to the puzzle-solving in ethics (PiE) model, comes from not getting lost in the rules, regulations and approvals that come with asking ethical questions. Making the right decision from day one is paramount, as is the need to avoid the problems associated with an undefined strategy or lack of audit processes. These become more integral to the conversation when we factor in the increased power of algorithmic systems and the technical challenges that they present.

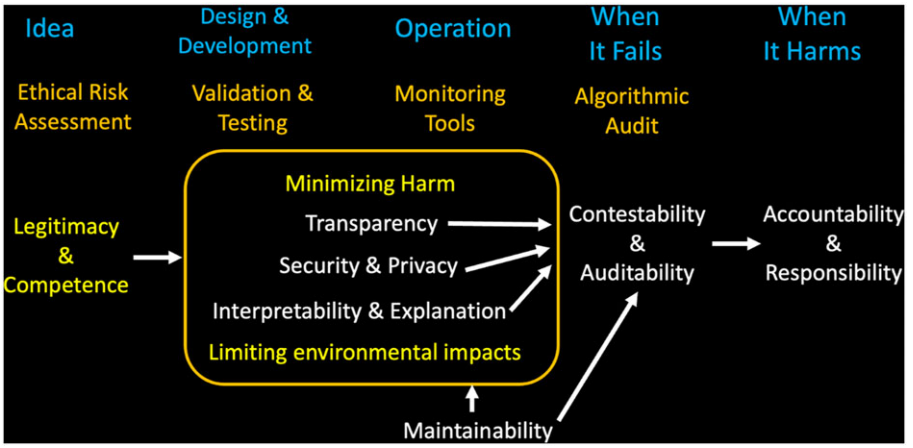

Figure 1. Principles and tools across the lifecycle of an algorithmic system.

At the Institute for Experiential AI (EAI), for example, we use the PiE Model that was developed by the AI Ethics Lab (Canca 2019). This model provides a roadmap to responsible AI that includes governance, the product and the skills required for effective integration. In addition, the institute has created a high-level AI Ethics Advisory Board (Northeastern Global News 2022) and provides deep technical audits of an AI system to determine if there is any bias, appropriate use of data, security and, importantly, the right quality. The last aspect is of vital importance, even more so when you look at how generative AI has been released into the wild.

OpenAI – ChatGPT's developers and owners – has publicly stated that it wants to keep AI 'safe and broadly beneficial', and reportedly spent more than six months working with a more powerful version (GPT-4) to ensure that it ticked as many ethical boxes as possible (Open AI 2023). Real-world use cases are far more difficult to predict than those tested in a laboratory setting, and the company has said that society will take time to adjust to how AI evolves and the potential for harm.

As we have noted before, AI systems are everywhere, but that merely raises the bar in terms of how we promote responsible AI and ensure that everyone is playing by the rules. That means we need to not only regulate AI but also agree on what those regulations are. In the next section, we will discuss why that is not as easy as it sounds.

Regulation

Investment in and integration of AI is top-of-mind for companies of all sizes, but it is the reference points and expectations that have been attached to this next-generation tech that are important to implement in any AI future. Regulation, an essential part of the puzzle, remains a work-in-progress.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

As noted in the previous section, acknowledging the principles and intended governance of AI is only the first step. What matters is how we regulate the use of AI in the real world. At time of writing, the regulatory landscape is akin to the Wild West. This situation has not been helped by the emergence of generative AI tools such as ChatGPT and the rush to integrate these large language models into workflows across a wide range of industries.

For the wider AI community, this very public awareness of what AI systems are capable of has raised not only questions about how much people know about the technology but also what level of regulation will be required to keep it under control.

ChatGPT, for example, became the fastest-growing consumer application ever, attracting more than 100 million Monthly Active Users (MAU) in the initial eight weeks following its release in November 2022. Other generative AI systems were quickly released by big tech companies, with an almost inevitable level of regulatory interest from global government agencies following soon after.

Reuters reported (Shepardson and Bartz 2023) that the US National Telecommunications and Information Administration (NTIA) was curious to know if regulatory measures could be applied that 'provide assurance that AI systems are legal, effective, ethical, safe, and otherwise trustworthy'. The Biden Administration, the news source noted, was concerned about whether these tools were 'dangerous', with the NTIA expected to draft a report that would inform the administration as to AI's present and potential future. This cautious approach is not a surprise, especially from a US agency that advises the President on policy direction.

The key point to remember is that the United States is not acting in a vacuum and there will be similar initiatives in progress in other countries.

To date, the number of actual regulatory processes that are ready to be signed is non-existent. Current AI regulations in the US and Europe, for example, fall into two camps: laws that could apply to the use of AI systems and proposed regulatory frameworks or blueprints for the future.

The US does not have any specific AI regulations, but the White House Office of Science and Technology Policy (OSTP) published a 'Blueprint for an AI Bill of Rights' (White House 2022) in October 2022 that establishes five principles for the design, use and deployment of what the agency refers to as 'automated systems'. According to the OSTP, this framework is a response to the concerns of public citizens and will be a significant step towards protecting people from the (unnamed) threats of artificial intelligence.

The principles and the associated practices are exactly what you would expect from an exploratory guideline – safe and effective systems, algorithmic discrimination protections, data privacy, notice and explanation, human alternatives, consideration, and fall-back. What is more interesting is that the federal government is not only pushing for a defined policy for AI – with the very public backing of President Biden – but also drawing on laws or regulations that could apply to AI systems.

For instance, the Federal Trade Commission Act (FTC Act) prohibits unfair or deceptive trade practices. This could potentially include using AI systems in a way

that is deceptive or harmful to consumers. In addition, the Cybersecurity and Infrastructure Security Agency (CISA) has published several guidance documents on how to secure AI systems. Both approaches are still nascent in terms of regulating AI in its current incarnation, but they do provide some evidence that algorithmic systems are likely to become regulated sooner rather than later. The US Chamber of Commerce recently released a report that called for discussions around the use of AI to be 'ramped up', noting that the technology had the capacity to hurt economic growth or become a national security risk – a stance that goes against its usual anti-regulation rhetoric (US CoC 2023).

Europe, on the other hand, has adopted a more strict approach to AI regulations. The EU's General Data Protection Regulation (GDPR) became enforceable in 2018, governing the way in which public companies process personal data that relate to an individual. This legislation adheres to a set of defined principles that include fair and lawful processing, purpose limitation, and the concept of data minimization and retention. GDPR was designed to protect our personal data, with the law stating that individuals must be informed about how their data are used and their data protection rights.

In the five years since it came into effect, GDPR has changed the way that data are viewed, even more so when the data become the subject of a breach or unauthorized use. And while the law doesn't specifically mention AI or algorithmic systems, there is a tacit understanding that they could fall under GDPR's Article 22. Automated decision making has become more of a concern in recent years, and it is expected that the new wave of generative AI will increase conversations over whether the provisions of GDPR can be brought into the regulatory mix.

From a pure AI standpoint, however, the European Union has been proactive. In 2021, the EU published a proposal for an Artificial Intelligence Act (AIA). If adopted, the AIA would create a comprehensive regulatory framework for AI systems and classify them into four risk categories:

- *Unacceptable risk*: AI systems that pose a serious threat to safety, fundamental rights, or public security would be prohibited.
- *High risk*: AI systems that pose a significant risk to safety, fundamental rights, or public security would be subject to strict requirements, such as mandatory conformity assessment and market surveillance.
- *Limited risk*: AI systems that pose a limited risk would be subject to certain requirements, such as transparency and documentation.
- *Minimal risk*: AI systems that pose a minimal risk would not be subject to any specific requirements.

This proposal has been making its way through the requisite regulatory processes since it was introduced and, while it is seen as a positive step forward, the attempts by the EU to address the various elements of AI within a standardized framework have (at time of writing) not produced a concrete piece of legislation. Granted, the EU is focused on risk management, but it can be argued that the AIA is the first law that

deals specifically with the impact that AI will have on public health, safety, and well-being.

High risk systems, for example, would fall under the most stringent of regulations. This would include any AI that is deemed to be a safety component of a regulated product – medical devices, machinery, hazard detection in motor vehicles, to name just three. Biometric identification and categorization of natural persons by law enforcement would also be in this category, with developers of AI systems required to establish the intended purpose of the algorithms.

In addition, any AI regulations would have to consider the end user and the outcomes that the AI produces – conversational AI is, for instance, a part of child development with millions of children interacting with voice–user interfaces such as Amazon's Alexa or Google Assistant. When you factor in a familiarity with recommendation systems via child-friendly apps, then the risks associated with AI take on a new level.

According to the lawmakers behind the legislation, the EU AI Act (European Union 2021) could become a global standard. There is a need to determine the extent to which AI can be a positive as opposed to a negative influence on society, and it is already reportedly providing a framework for other countries to consider their own regulatory stance on AI. The caveat is that there are some identified flaws with the AIA, including the fact that it is (in its current form) relatively inflexible and reflects the point in time in which it was first drafted.

In my opinion, there are three significant problems with the proposed EU laws. First, we should not be looking to regulate the use of the technology but focus on the problems and sectors that the technology impacts. There is an element of fear around the potential for human harm, but we should be approaching AI in the same way that we enact oversight on food or healthcare. In other words, we have to draft regulations that work for all possible technologies and not just the versions that exist in the here and now.

Second, we must understand that risk itself is not static. Rather, it exists as a continuous variable. Companies that choose to integrate AI into their business practices are unlikely to want to allocate it to the four risk categories proposed by the AIA, and there will be the potential for a conflict of interest in terms of self-evaluation. A better way of looking at the potential risks is to compare AI regulation to improvement mechanisms such as the US Food and Drug Agency – an accepted path that takes a phased approach to the approval of a medicine or device through study and human validation.

The third and final problem that the proposed regulations don't address is the issue of systems that do not use AI. Not all algorithmic systems are human-trained, rather they rely on interpreting the data that they are presented with and decide based on the information itself. The use of automated decision-making tools is extremely prevalent in sectors such as recruitment and finance, and there is an argument to be made that the EU AIA Act should apply to all algorithmic systems, not only AI-based systems. Otherwise, there is an easy loophole to skip regulation.

This becomes even more of a concern when we come back to the influence of generative AI. Large language models are the face of AI in 2023, and there is already a backlash against tools such as ChatGPT and its presumed successors. Large language models are being promoted as productivity tools, with companies already using them to automate customer service, generate reports, create content, streamline sales or marketing, and more (Leighton 2023).

This increase in AI usage and awareness has now forced various governmental agencies to fast-track discussions about what regulatory actions will be needed. Notwithstanding the aforementioned EU Laws, the *Financial Times* reported that the EU is set to introduce a 'sweeping set of regulations on the use of AI', with developers required to be more transparent on how these models are trained, and whether they infringe copyright (Espinoza and Johnston 2023).

Italy actually banned ChatGPT, citing a breach of strict privacy regulations. That ban lasted a month, and the tool was only restored after OpenAI agreed to make the changes required by Italian regulators. Germany and France have taken a less gung-ho approach, albeit that both countries are waiting to see how other EU members are dealing with the increase in AI usage; Spain, and Ireland, for instance, have decided to launch investigations into its usage and potential harm, but these are not expected to mirror the Italian approach.

No longer beholden to the EU's decision-making process, the United Kingdom government recently released a white paper (UK Government 2023) that sets out proposals for regulatory activity, although its authors were careful to note that the principles under discussion were more akin to practical guidance as opposed to concrete rules around AI use. The UK is also planning to both review the AI market and assess the potential for 'business domination' by a single tech entity or vendor. According to BBC News (Thomas 2023), the Competition and Markets Authority (CMA) is concerned that two of the companies aligned with the current AI revolution – Microsoft and Google – could have a significant localized impact, with the software behind large language models (LLMs) having the power to 'transform the way businesses compete as well as drive significant economic growth'.

Away from Europe, we must not overlook how China and the US are dealing with this wave of innovation and adoption. China does not have access to ChatGPT, and its draconian internet censorship laws could be a blessing in disguise from a regulatory standpoint. The Chinese Communist Party is considering measures (Digichina 2023) for the 'Management of Generative Artificial Intelligence Services' to fall in line with the country's existing Cybersecurity Law, so it is safe to assume that regulations to govern the use and commercial development of AI are not far away. Companies such as Alibaba, Huawei, Baidu and Tencent are all working on generative AI, and it will be interesting to see how and when they decide to integrate these experiments into their products.

In the US, the Biden Administration is keeping a close eye on developments but has taken no actual action (to date) towards investigating how the tools are being used or introducing regulations. However, the White House did announce (White House 2023) a series of actions that will 'promote responsible AI innovation', which

included a $140 million investment into seven National AI Research Institutes and a public assessment of existing generative AI systems. The latter included a meeting on 4 May 2023 between President Biden and the CEOs of Anthropic, Google, Hugging Face, Microsoft, NVIDIA, OpenAI, and Stability AI, all of whom reportedly agreed to an evaluation of their systems as part of the Blueprint for an AI Bill of Rights and AI Risk Management Framework.

The Brookings Institute reported (Engler 2023) that the 'U.S. federal government's approach to AI risk management can broadly be characterized as risk-based, sector-specific, and highly distributed'. According to the institute, any advantages to this approach can be negated by the fact that it contributes to a perception of 'uneven development of AI policies', with the authors of the report noting that the publication of several guiding documents has not led to the required AI regulations.

At the beginning of this section, we noted that AI regulation is a work-in-progress. And while the discussions around what it should look like are gathering pace, the current fascination with and deployment of algorithmic systems make regulation less of a talking point and more a required action. The question that we need to ask, however, is what timeline are we hoping for?

The EU Parliament just approved the new version of the EU AI Act in early May 2023 (including an article on generative AI) and expects to sign AIA into law in 2024, but there are no guarantees that global regulation will be enacted anytime soon. The challenge, thereby, is not whether we need regulation (which we do), but what that regulation will look like and how it will affect both the need for responsible AI and the principles that underlie its implementation. Until we know the answers to this simple question, then the analogy of the Wild West is likely to be in play for the foreseeable future.

Conclusions

Mark Weiser, the acknowledged father of ubiquitous computing, once said that a good tool is an invisible tool. His analogy was that the blind man tapping the cane feels the street, not the cane itself. In other words, it must be something that does not intrude on your consciousness and, by association, is not there to cause you harm. AI can be that tool, but the human element is what gives it responsible and ethical guidelines.

Science fiction has arguably become our current reality. The global village predicted by Marshall McLuhan back in the 1960s is no longer a vision of the future, it has become part of our connected present, a digital ecosystem that touches every part of our lives. Devices that we use daily have augmented our humanity, an evolution that provides us with access to a second brain whenever we need it. These connected devices don't just define who we are, they present a digital self. In addition, automated systems that have been designed and integrated by human beings to not only alleviate mundane tasks, but also allow companies to be more streamlined and efficient are ubiquitous in every industry sector.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

When these two factors are combined, they create a virtual footprint that is intended to both enhance the lives that we lead and the work that we do. The caveat is that the tools that we use and the solutions they provide are, essentially, unthinking systems. And while they may have been developed with good intentions, the unintended ethical consequences of machines programmed by humans could be a real and present danger.

Ultimately, responsible AI should not be an afterthought, but until we have the regulatory guidelines and guard rails in place, AI systems will need to be integrated in such a way that all stakeholders can feel comfortable with how they are being used and, importantly, their impact on society. AI can be a game-changer on so many levels, but all games need to have rules. And the players must follow them. If they don't, then this bright and shiny new tech tool has the capacity to impact our lives in ways that we will find difficult to predict or even plan for.

Acknowledgements

Thanks to David Bolton for his English revision of this article.

References

ACM (2022) *Principles for Responsible Algorithmic Systems*. Lead authors: Baeza-Yates R and Matthews J. https://www.acm.org/articles/bulletins/2022/november/tpc-statement-responsible-algorithmic-systems

Canca C (2019) *The PiE Model: Our Ethics Model for Innovation*. AI Ethics Lab. https://aiethicslab.com/pie-model/

Digichina (2023) *Translation: Measures for the Management of Generative Artificial Intelligence Services*. Stanford University. https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/

Engler A (2023) The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings Institute. https://www.brookings.edu/research/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

European Union (2021) *The AI Act*. https://artificialintelligenceact.eu/

Espinoza J and Johnston I (2023) European Parliament prepares tough measures over use of AI. *Financial Times*. https://www.ft.com/content/addb5a77-9ad0-4fea-8ffb-8e2ae250a95a

IDC (2023) Worldwide spending on AI-centric systems forecast to reach $154 billion in 2023. https://www.idc.com/getdoc.jsp?containerId=prUS50454123

Leighton N (2023) 6 ways business leaders should integrate ChatGPT. Forbes. https://www.forbes.com/sites/forbescoachescouncil/2023/02/22/6-ways-business-leaders-should-integrate-chatgpt/?sh=535371326c61

Northeastern Global News (2022) Northeastern launches AI Ethics Advisory Board to help chart a responsible future in artificial intelligence. https://news.northeastern.edu/2022/07/28/ai-ethics-board/

Open AI (2023) Our approach to AI safety. https://openai.com/blog/our-approach-to-ai-safety

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press

Shepardson D and Bartz D (2023) US begins study of possible rules to regulate AI like ChatGPT. Reuters. https://www.reuters.com/technology/us-begins-study-possible-rules-regulate-ai-like-chatgpt-2023-04-11/

Thomas D (2023) AI investigated in UK over fears of domination. BBC News. https://www.bbc.com/news/business-65478156

UK Government (2023) AI regulation: a pro-innovation approach. Department for Science, Innovation and Technology and Office for Artificial Intelligence. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

US Chamber of Commerce (2023) Report of the Commission on Artificial Intelligence, Competitiveness, Inclusion, and Innovation. https://www.uschamber.com/assets/documents/CTEC_AICommission2023_Exec-Summary.pdf

White House (2022) Blueprint for an AI Bill of Rights. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

White House (2023) Biden–Harris administration announces new actions to promote responsible AI innovation that protects Americans' rights and safety. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/

Zicari R (2022) ODBMS Industry Watch: on responsible AI. Interview with Ricardo Baeza-Yates. https://www.odbms.org/blog/2022/02/on-responsible-ai-interview-with-ricardo-baeza-yates/

About the Author

Ricardo Baeza-Yates is Director of Research at the Institute for Experiential AI of Northeastern University. He is also a part-time Professor at Universitat Pompeu Fabra in Barcelona and Universidad de Chile in Santiago. Before, he was the CTO of NTENT, a semantic search technology company based in California, and prior to these roles, he was VP of Research at Yahoo Labs, based in Barcelona, Spain, and later in Sunnyvale, California, from 2006 to 2016. He is co-author of the best-selling textbook *Modern Information Retrieval* published by Addison-Wesley in 1999 and 2011 (2nd edn), which won the ASIST 2012 Book of the Year award. From 2002 to 2004 he was elected to the Board of Governors of the IEEE Computer Society and between 2012 and 2016 was elected to the ACM Council. Since 2010 he has been a founding member of the Chilean Academy of Engineering. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow, among other awards and distinctions. He obtained a PhD in computer science from the University of Waterloo, Canada, in 1989, and his areas of expertise are web search and data mining, information retrieval, bias and ethics on AI, data science and algorithms in general. Regarding responsible AI, he is actively involved as an expert in many initiatives, committees, and advisory boards all around the world: Global Partnership on AI, Global AI Ethics Consortium, ACM's US Technology Policy Committee, and IEEE's Ethics Committee. He is also a co-founder of OptIA in Chile, an NGO devoted to algorithmic transparency and inclusion, and a member of the editorial committee of the new Springer journal *AI and Ethics*, in which he co-authored an article highlighting the importance of research freedom on AI ethics.

https://doi.org/10.1017/S1062798723000145 Published online by Cambridge University Press