Artificial intelligence, Human responsibility

Triodos Bank position paper on ethical AI

Triodos Bank

Scope of the paper

The topic of artificial intelligence (AI) has been gaining increasing traction and sparking heated debates, particularly with the introduction of ChatGPT in November 2022. It has now become a business topic of paramount importance. Applications of AI systems for businesses and governments are flourishing, and while companies specialising in AI have become a hot discussion point in investor circles, there are substantial risks associated with AI-powered technology. Such risks need to be known and taken into account by society, companies and investors as well as by each of us individually.

This paper is intended to set out Triodos Bank's values-based position on AI. The primary focus is on identifying key issues and risks related to AI systems. In the paper, we also formulate precautionary principles for AI systems, identify highly controversial uses of AI technology, and discuss the role of financial institutions in nudging companies towards responsible practices for developing and using AI technology.
The paper does not classify existing AI systems and technologies, nor does it explore the positive impact – current or potential – of AI technology. It also avoids describing the macroeconomic implications of a large-scale adoption of AI technology, such as effects on the labour market.

Content

Triodos Bank

1. Introduction

"There's nothing artificial about AI. It's inspired by people, it's created by people, and – most importantly – it impacts people. It is a powerful tool we are only just beginning to understand, and that is a profound responsibility."

Fei-Fei Li, Co-Director, Stanford Institute for Human-Centered Artificial Intelligence

Technology has been shaping our societies and our lives since the beginning of human history. The second half of the 20th century was marked by the advent of computer technology, and now the accelerated development and adoption of artificial intelligence (AI) systems marks another key turn in the history of technology, and possibly of humanity.

The release of ChatGPT in 2022 has initiated a broad public conversation on AI. AI technology is now recognised as a transformative force changing our societies and economies, and it is already shaping our future. It is no longer a question of whether we are for or against AI, but rather how we can ensure that we develop and use it wisely, so that its power and potential are a force for good.

AI systems and their uses have evolved from demonstrating theorems[1] and playing checkers[2] to having significant and widespread commercial applications. Today, AI technology is having a profound effect on humanity by stimulating progress in many areas of science and the economy at an unprecedented speed. Increased capacity for data gathering has allowed us to leverage large amounts of data for the development of AI systems and to provide solutions to complex problems across businesses, institutions and civil society. Many industries have been using AI technology for years, including in life sciences for medical imaging and drug development, cybersecurity for anomaly and fraud detection, 3D design and visualisation, computer vision, and sales and marketing for predictive modelling.

However, progress has not come without concerns, and awareness is growing around the risks associated with such rapid technological development. There are many examples of algorithmic bias, and while not all such algorithms are powered by AI, they can all have very real consequences for individuals and society. For instance, the Dutch childcare benefits scandal (*toeslagenaffaire*) has put the spotlight on institutionalised discrimination, and accidents caused by self-driving vehicles have raised questions about accountability. Meanwhile, AI-generated content is making it difficult to distinguish between human and machine-generated content, and the introduction of AI systems in businesses is raising concerns about the displacement of large numbers of workers.

1.1 About this paper

This paper sets out the values-based position of Triodos Bank on AI technology, reflecting a precautionary approach. The focus is on risks and issues related to AI systems. We formulate precautionary principles on the development and use of AI systems and identify uses of AI technology that are highly controversial, and we believe should be no-go areas. We also highlight the role of financial institutions in setting requirements and expectations of the businesses they finance and invest in, as well as in their own operations and as corporate citizens. We close with a call to action for financial institutions to take a responsible stance on AI technology.

One important note: the focus on the criticalities of AI technology does not mean that Triodos Bank is not excited about its potential. As a values-based bank and responsible impact investor, we are very open to AI systems that display potential for positive impact in any of the transition themes that drive our impact strategy. However, we believe that positive impact does not result from the technology itself but from how it is used, whereas negative impact can arise from how the technology is designed and developed, as well as how it is used. As always, it is up to all of us to weigh the pros and cons by exercising good (human) judgement.

1.2 Talking about artificial intelligence

Discussions about the benefits and risks of technology often start without a good understanding of the technical background and terminology. It is important to identify the key concepts and definitions and to demystify the term artificial intelligence (AI).

There is no universally accepted definition of AI. The EU and the OECD[3] define an AI system as "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."

In short, **artificial intelligence** is the ability of a machine to perform tasks commonly associated with intelligent beings. Within AI technology, currently the most important subset of techniques is **machine learning**, which is the capability of a computer programme or a machine to learn and take actions without being explicitly encoded with commands. However, the term AI is used differently by different groups of people, and in the general discourse what is referred to as AI is often a set of techniques in the science of machine learning called **deep learning**. These techniques allow, loosely speaking, to artificially simulate a human learning process starting from unstructured data such as images or audio files. Deep learning techniques are responsible for the reputation of AI systems as 'black boxes', whereas other AI techniques allow for full explainability of the mechanisms leading to a model's output. **Generative AI (GenAI)** models are AI systems based on deep learning techniques that are capable of generating new written, visual or audio content. ChatGPT is currently the most famous example of this type of AI.

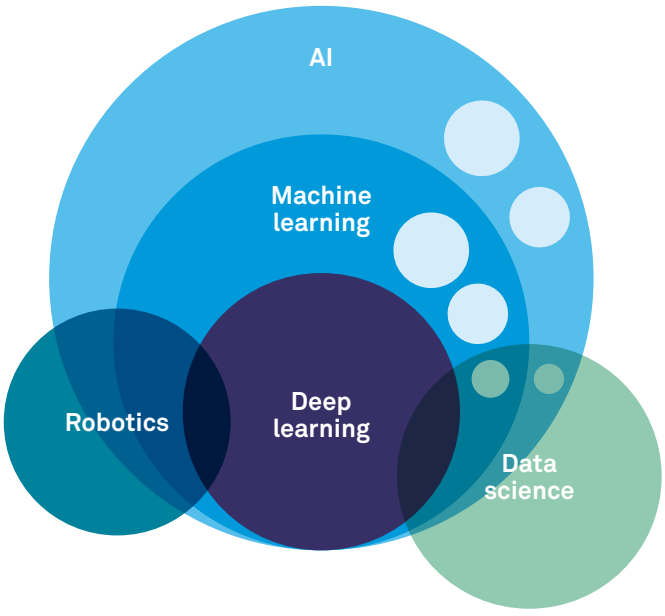

Figure 1: Illustration of interaction of different disciplines related to AI.
Source: Own elaboration from Andrew Ng, deeplearning.ai

It is important to state that not all AI is 'super intelligent'. On the contrary, current AI applications are known as **narrow AI**. This means they are AI systems designed to perform a specific set of tasks such as email spam filtering software, machine translation systems, shopping ads, chatbots (including ChatGPT) and self-driving cars. All these systems operate within the boundaries of a specific set of goals. The prospect of **artificial general intelligence (AGI)**, also called strong AI, or even more so, **artificial superintelligence (ASI)**, is currently raising questions and concerns. These types of artificial intelligence would be able to perform intellectual tasks in a way that can be compared to that of a human being, and in the case of ASI, beyond human intelligence, learning to adapt to new situations and not being limited to a specific set of tasks.

AI systems are not human-like **robots**. Robotics is a branch of mechanical engineering that can, but does not necessarily have to, use deep learning techniques (with impressive results). However, not all robots are equipped with AI systems, and not all AI systems move in the physical world. Furthermore, not all models that use data are powered with AI. **Data science** can make use of AI, but a simple set of linear regressions, largely used in data science, is not an AI system.

Demystifying AI: language choices

When discussing AI, it is essential to be aware of how language shapes our thinking and understanding of things. We have made a conscious choice to refer to 'AI systems' or 'AI models', rather than to simply use the term AI. This helps to demystify artificial intelligence, to avoid thinking of AI as a mysterious, possibly sentient technology and to highlight that humans are ultimately responsible for how artificial intelligence is designed, developed and used.[4]

1.3 Regulation is emerging

Until recently, the focus of regulators with respect to technological advancements was predominantly addressing risks related to the misuse of personal data. The rapid dissemination of AI systems has amplified concerns beyond data protection. Safety and ethical concerns regarding the development and use of AI technology have incentivised several national governments and international organisations to develop and adopt guidelines and frameworks for responsible AI. As of June 2024, some national governments and supranational institutions have also been developing regulatory initiatives addressing the development and use of AI.

Attempts to regulate AI are primarily aimed at preventing and mitigating risks related to algorithmic fairness, transparency and human oversight.[5] In May 2024, EU institutions approved the EU Artificial Intelligence Act (also called the AI Act), which at the time of this publication is expected to enter into force by July 2024. The regulation applies to providers placing AI systems on the EU market or putting them into service in the EU, therefore effectively having a global impact. It does not, however, apply to AI systems developed in the EU and exported outside its borders. Some sectors are left out of the regulatory document, including civil aviation and national security. On the other side of the Channel, the UK has taken a more flexible stance with a pro-innovation approach to AI regulation, publishing a white paper in 2023 that provides a risk management framework but focuses on supporting innovation.

In the United States, several initiatives are ongoing. The US presidency issued an Executive Order on Safe, Secure and Trustworthy AI in October 2023 indicating the government's intention to develop robust standards for AI safety and beyond. China, on the other hand, has opted for more targeted regulations of AI – starting with a regulation on recommendation algorithms[A] in 2021, rules on synthetically generated content in 2022 and a regulation on generative AI in 2023. Other notable efforts are taking place in Canada, Japan, South Korea, Australia, Singapore and India. Overall, countries in Africa and Latin America are less active on this front.

On a global scale, the UN launched an AI advisory body to make recommendations on the international governance of AI.[6] A global network of AI Safety Institutes, where cutting-edge AI models would be rigorously tested before public release, is also emerging with a commitment to close collaboration from various parties including the US and the UK.[7] Meanwhile, the G7 countries agreed on a voluntary code of conduct[8] for organisations developing advanced AI systems, complementing the various regulatory initiatives.

Multiple calls for regulation – but differing priorities

While the AI industry has a very positive stance on the potential benefits of AI for humanity, important industry actors have called for regulation and robust governance systems, particularly for artificial general intelligence (AGI) and superintelligence. AI companies are particularly vocal about the existential threats linked to the rapid AI development, showing concerns about our ability to control the pace and direction of AI development, especially given the substantial competitive and shareholder pressure these companies face.

In contrast, in calling for regulation, civil society organisations and academic researchers highlight the existing issues with AI development and use, particularly the algorithmic bias and violations of human rights and universal freedoms. In the EU, a coalition of civil society organisations led by the European Digital Rights Network (EDRi) has long been calling for the EU AI Act to protect and promote human rights, with particular attention to marginalised groups. The call has been echoed by renowned universities and by many independent organisations, which also warn of the risk of loopholes in the legislation allowing tech companies to self-regulate.

[A] These are algorithms that provide each user with personalised suggestions which are deemed most pertinent. They are the most deployed type of AI we find when browsing the internet.

2. Ethical and sustainability issues and risks

Many of the ethical and sustainability issues that surface in relation to AI technology are not new, and they are not specific to AI either. For instance, discussions about algorithmic bias and the energy needed to operate data centres have been ongoing since before AI systems were adopted at their current scale and pace.

Nevertheless, the pervasiveness of AI technology gives a new dimension to these issues. We believe this is due to an intrinsic characteristic of AI technology to amplify existing issues (amplification), as well as users' tendencies to rely too heavily on technology (overreliance) and the opportunity to shift responsibilities to machines (moral outsourcing).

We believe that these three aspects both of AI technology and of our attitudes towards it can provide a lens through which we can analyse the various risks and issues related to AI. These can be grouped into more classic categories: resource use or environmental risks, human rights and fundamental rights, and business ethics.

Roots of ethical and sustainability issues related to AI technology

- **Amplification:** AI systems have an unprecedented capacity to work at scale. This also means they have great potential to amplify unaddressed issues and to exacerbate biases and related threats to our social fabric. At the same time, the use of AI technology can trigger a shift in resource use, as technological innovation increases the need for natural resources over human labour.
- **Overreliance:** AI systems are often treated as highly rational and reliable sources of knowledge (and sometimes even wisdom). Their outputs are often left unscrutinised and uncontested, as if they were oracles. Users risk blindly trusting outputs obtained through AI systems and expecting AI systems to deliver solutions to potentially any problem (even those that can be solved within a reasonable amount of time and resource use with more primitive technologies).
- **Moral outsourcing:** Overreliance is partly explained by a tendency to refer to AI systems as sentient beings – speaking for example of 'racist AI' or 'xenophobic machines'. Careless use of language can result in the automatic transfer of responsibility of encoded biases and the functioning of AI-related products onto the products themselves. This can absolve AI system creators and users from moral obligations linked to technology dissemination.

2.1 Resource use

The carbon footprint of AI models has long been under public scrutiny. While widely recognised estimates of the environmental costs of AI are not yet available, here are three main areas of concern that translate into environmental costs of developing and deploying AI.

- **Energy consumption** of large datacentres linked to the training of AI systems as well as the energy needed to deploy these systems is hard to quantify. This is in part due to a lack of information from the companies that provide the technology. However, a recent study estimated that by 2027 energy consumption for AI systems is likely to be equivalent to that of a country like the Netherlands.[9] Other studies predict that by 2030 AI could account for 3-4% of global power demand.[10] With the rapid increase in use of AI systems, the carbon footprint of AI systems is going to play an ever-increasing role in contributing to environmental deterioration.
- **Water consumption and withdrawal** for AI technology is generally overlooked and is mainly linked to freshwater used in datacentres to generate electricity and to cool down servers. However, a recent study[11] described and estimated water consumption related to AI which replicates the scoping approach used to estimate carbon footprints. This study considered not only on-site server cooling (scope 1), but also off-site water for electricity generation (scope 2) and supply chain water for server manufacturing (scope 3). The results of the study paint a grim picture of water usage. Reliance on AI models also raises concerns in terms of water withdrawal, as fresh water is a finite resource and it is being depleted and polluted faster than it can replenish itself. This will lead to competition between different sectors of the economy for water use at a certain time in the area where the water is sourced.
- **Raw materials and e-waste:** The issue of raw material use and e-waste is not specific to AI, but applies to technology as a whole. In 2019, more than 50 million

metric tonnes of e-waste were produced annually.[12] AI applications can potentially be useful and effective in reducing the amount of raw material used in industrial processes, from production to waste management. However, the adoption of AI in the industry might take time. The rapid development of AI systems also means that AI hardware can quickly become outdated, contributing to the problem of e-waste. If not properly recycled and treated, e-waste releases hazardous chemicals such as lead, mercury and cadmium into the environment.

The environmental issues related to AI are primarily linked to the development phase of AI models, which use high amounts of resources and energy during model training. The deployment of AI systems is also a factor, but to a lesser extent. This is largely a drawback of the technology itself, rather than the result of an unethical approach to AI development. However, in addition to the technicalities of resource and energy use, we must recognise our strong reliance and sometimes overreliance on digital technology, including AI, as one of the drivers of consumption of digital technology. This therefore underlies the environmental issues described above.

2.2 Human rights and fundamental rights

Human rights are enshrined in the UN Declaration of Human Rights[13] as well as in the EU Charter of Fundamental Rights embedded in the Treaty of Lisbon[14] and in the International Covenant on Civil and Political Rights.[15] Widespread development and use of AI systems poses important questions about the violation of several human rights and fundamental rights of both individuals and collective entities.

- **The right to respect for a private life and privacy** is a fundamental human right, and it implies that one's personal information including official records, photographs, letters, diaries and medical records must be kept securely and only shared with the data subject's permission. AI algorithms are based and trained on data that may involve personal data. A user's data privacy is violated when their personal data is collected and used without their knowledge, for example for targeted advertising. Even more concerning is the presence of biometric recognition systems, such as facial recognition systems used in public spaces. This poses a very significant threat to privacy, as people's movements and behaviour can be monitored in real time without their explicit consent. Finally, while AI now plays a fundamental role in cybersecurity, AI systems can also be used to launch cyberattacks. Cybercriminals are increasingly targeting AI systems, which poses a growing threat to individual's and companies' privacy and data security.

- **Freedom from discrimination** is of key importance in the development of AI systems. AI systems can incorporate biases at any stage of the AI model lifecycle[16], from data collection and model development to deployment, monitoring and feedback integration. AI models are trained on large amounts of data which may itself be biased, perpetuating and amplifying existing stereotypes. The resulting AI systems may discriminate against people who were underrepresented or overrepresented in the training sample. Also, the teams of developers may have unconscious biases that are reflected in the data without being recognised. These types of issues have been identified in many fields such as healthcare[17] and banking (credit risk).[18] Algorithmic predictions have been demonstrated to be discriminating for minority groups and severely racially biased, for reasons that cannot be solely attributed to skewed datasets.

- **The right to life, liberty and personal security** is becoming more endangered as autonomous weapon systems (AWS) become more widely available.[19] Remote-controlled robotic vehicles are likely to be able to fight alongside conventional troops as soon as 2030, and fully autonomous remote combat vehicles, or (RCVs) would have their own self-contained AI systems.[20] These automated tanks, killer robots and kamikaze drones could be deployed to make life-and-death decisions without a human making the final call, in an extreme case of moral outsourcing. In addition to physical security, cybercrime and threats to individuals' property and digital identity are already a reality. Accessibility and public availability of sophisticated AI systems poses serious security risks, for individuals and society as a whole. Cybercriminals can use AI to generate malware rapidly, automate attacks and enhance the effectiveness of scams using deep fakes.[21]

- **The right to freedom of opinion and expression** (as well as the right to freedom of thought, conscience and religion, and peaceful assembly) can be easily violated when citizens are subjected to AI-powered surveillance which is then used by law enforcement agencies. Without robust regulatory safeguards, citizens risk being persecuted for exercising their right to criticise a ruling regime or participate in political demonstrations or cultural and religious gatherings. Moreover, AI tools can be used to censor media and independent journalists. The concentration of power in the hands of a few private online platforms increases the risk of information being unlawfully governed by these actors.[22]

- **The right to the protection of intellectual property** is also a recognised human right in the Declaration (article 27.2) and a fundamental right contained in the EU Charter (article 17), and it is essential to foster innovation. Infringements of intellectual property rights may occur if AI systems are trained or operate on available online data, such as texts and pictures,

without checking whether this is protected or private. AI systems may produce text results that partially replicate texts without providing sources or generate images that are very close to the source material.[23]

- **Labour rights** are a subset of human rights (article 23 of the Declaration), and large-scale societal risks arise from the advent of AI systems which could be used for tasks previously performed by humans.[24] AI can automate processes, generate texts and presentations and perform analyses at higher efficiency rates than most people. However, while these risks raise questions for future jobs, some labour rights are already being threatened or violated in the AI development process. The training of current AI systems requires human input from data labellers, which has reportedly led to an industry of millions of workers globally who perform repetitive tasks under precarious labour conditions and poverty wages. These workers are often exposed to violent or disturbing content.[25]

In summary, some AI technologies, such as live facial recognition and biometric identification and classification, raise specific concerns both in terms of their potential use for mass surveillance and their reliability from a non-discrimination perspective. This amplifies the risk of structural discrimination and limitation of personal freedoms. Autonomous weapon systems can shape warfare and defence and security practices by outsourcing responsibilities for life-and-death decisions to machines, which is a clear example of moral outsourcing. Finally, the issue of biased algorithms is a concern that can impact any use with direct or indirect implications for people's access to adequate standards of living.

2.3 Business ethics

Business ethics is the application of ethical values and moral principles to the way businesses and individuals engage in business activities. Business ethics covers a wide range of topics including corporate governance and is typically encoded in policies and procedures. Business ethics principles and the practices that derive from them contribute to building trust in businesses, and they often precede and supplement regulation. There are various stages in the lifecycle of AI systems, from conceptualisation, development and design, to deployment and use, where business ethics should be taken into account.

- **Lack of transparency and explainability** of AI systems is one of the most discussed shortcomings of AI technology, and it is the root cause of several of the issues presented in this chapter. While there are difficulties in establishing clear standards for transparency in AI systems[26], it is important to have a certain degree of explainability of AI systems' outcomes. This means underlying data sets should be available for review and algorithms should be replicable so that errors and flaws can be traced. For instance, if there is a suspicion of biased outcomes, it should be possible to check whether the underlying data sample is biased in any way or where in the system the problem is rooted. Ensuring transparency is key in business ethics as it is essential for building trust and accountability in both the business and its products.

- **Lack of clear accountabilities** is another area of concern. It is difficult to place responsibility on the various actors in the production chain of AI systems and AI-powered products because AI outcomes are hard to explain. For this reason, good product governance is of the utmost importance. Product governance is the active oversight of product design, development, compliance, risk management and protection of product trust. A clear example is self-driving cars. These vehicles use AI software and sensors to travel between destinations without human interference. While this technology can revolutionise travel and reduce human error that leads to accidents, the AI system may also fail to identify risks or pedestrians on the street. Such examples also raise questions about liability, product safety and user awareness.

- **Manipulative marketing practices** are hugely helped by AI models. AI is widely used in targeted marketing campaigns using browser history and cookies, as well as online shopping history. Companies are responsible for ensuring that their marketing practices are fair for their customers. However, there is a fine line between personalising the customer experience and manipulating it, and currently companies have to define this for themselves.

- **Addiction-inducing mechanisms** have been largely under scrutiny since the advent of social media and can be exacerbated by the latest advances in AI. Increasing dependence on machine-driven networks and tools can have a negative impact on people's cognitive and social skills. However, it is fairly common practice for many businesses in the gig economy to target user engagement, therefore explicitly developing systems that can induce serious addiction.

In summary, AI systems raise important concerns for transparent design and development and the capacity of businesses to ensure accountability and take full responsibility during the AI lifecycle. Moreover, AI systems can produce highly addictive features and can be used with manipulative intent.

3. Taking a stance on responsible AI

As outlined above, the unprincipled or irresponsible development or use of AI systems can bring about substantial risks to human and labour rights, business ethics and individual and collective security. These risks cannot be overlooked. AI technology is a tool that is not inherently good or bad. It can have positive impacts when used consciously, but negative impacts can result from how the technology is designed, developed and deployed.

For this reason, it is essential to define some high-level principles that lay the foundations for responsible development and use of AI technology at a broad societal level. It is also essential to set boundaries about what represents highly controversial practices and uses of AI systems.

Guiding principles

According to the EU Ethics guidelines for trustworthy AI, AI systems should be:

Lawful - respecting all applicable laws and regulations

Ethical - respecting ethical principles and values

Robust - from a technical perspective while taking into account the social environment

3.1 Principles for responsible use and development of AI technology

Several government bodies, international and civil society organisations, as well as companies, have outlined principles for trustworthy AI. The EU has set out Ethics guidelines for trustworthy AI, for which you can find the high-level guiding principles in the text box in this page. While the exact wording might change slightly, most entities have listed similar principles which we largely share. In our own words, these are the guiding principles that we think AI technology should abide by to be genuinely beneficial.

Humanity-centred AI: We believe that AI systems, and technology in general, must have human dignity at its core, and be not only human-centred but humanity-centred. This sets out the expectation that AI systems should be tailored to our needs and preferences as individuals as well as be designed and developed with broader wellbeing in mind. In other words, it is not enough for technology to be designed so that people can understand it and use it to satisfy their individual needs. It should also be designed, developed and manufactured in a way that upholds our collective needs and freedoms. This requires a conscious approach to the resources deployed. Technology must support humanity in achieving prosperous and healthy societies with individuals living on a healthy and hospitable planet.

Trustworthy AI: It is essential that AI systems are designed and developed with the highest standards of technical safety and robustness. They should be adequately tested over time before being circulated. The data used should be of high quality and accessed legitimately. Personal data and data in general should be collected, stored and used in a responsible manner in line with data privacy and data security standards. There should be transparency whenever we interact with machines rather than with other individuals and the functioning of the systems and their decisions should be adequately explained. Adherence to ethical principles and guidelines should be taken seriously, ensuring that fundamental rights-related risks are sufficiently mitigated before AI systems are distributed and deployed.

People in control: People should always be in control and maintain oversight of AI systems, both during development and use. We also believe that any decision on ethical issues that can fundamentally affect the rights and dignity of groups and individuals should never be fully outsourced to machines. In order for people to be fully in control, adequate and widespread awareness and digital literacy is of the utmost importance.

Adequate use: We should put in place mechanisms that help us avoid overusing technology and relying on it too much. For a truly healthy humanity and healthy lives, it is important that we actively preserve and nurture those human qualities and capabilities that are essential to maintain our creativity, mastery of manual skills and cognitive abilities. AI systems provide phenomenal support to human activities, but their development and use also comes at a cost. We should be mindful of the environmental and human resources being used and affected. AI systems should not be used to go beyond what is needed to achieve a legitimate aim. Overuse of the technology can bring about environmental costs as well as long-term societal costs, and it should not undermine our trust in human judgement.

Responsible development and use: AI systems should be developed and used in compliance with all existing laws and regulations, including those relating to human rights and intellectual property rights. The spirit of these laws should be taken into consideration when

regulatory loopholes are present that create legislative vacuums. As a financial institution that channels money towards local as well as remote and multinational businesses, Triodos Bank believes that companies are ultimately responsible for developing and using AI systems that meet the highest standards of safety and robustness and uphold strong ethical values, promoting awareness of our individual and collective digital rights and duties. Although development of the most advanced AI systems currently takes place mainly within private entities and corporations, the developments in this space are of much a broader, collective interest. Therefore, proper governance of AI systems should be established that involves all relevant stakeholders, to ensure that developments are carried out in the broader interest.

3.2 Controversial use of AI technology

Some AI-powered technologies and uses of AI technology are considered to be highly problematic and pose unacceptable risks. In some instances, such AI systems are expected to be explicitly banned by European institutions through the EU AI Act. At Triodos Bank, we consider that the following AI systems and uses are not in line with responsible AI principles and should be firmly condemned:

Lethal autonomous weapons: Autonomous weapons are weapons systems that, once activated, can detect and autonomously attack a target without human approval or intervention. Lethal autonomous weapons not only undermine the right to life, but they are currently being developed and used with no strict regulatory frameworks. They infringe the principle of human oversight in the use of AI systems and are a reprehensible example of moral outsourcing by delegating life-and-death decisions to machines.[27]

Biometric identification in public spaces: Biometric identification, specifically facial recognition, in publicly accessible spaces represents a substantial infringement of the right to privacy. Individuals are identified without their consent, and there is a high risk to data security leading to fraud and identity theft. Additionally, this technology can be used for mass surveillance, which poses a high risk for freedom of speech and association, and for democratic rights more broadly. Being observed can change the way we behave and affect our mental health and wellbeing.[28] It has also been demonstrated that accuracy varies by demographic, therefore creating very high risks of discrimination, and this can represent a threat to the rights of children and minors.[29]

Biometric categorisation and emotion recognition: Biometric categorisation is highly problematic both in its design and use and in the data collection process, which involves scraping biometric data from the internet, such as social media pages, without meaningful consent. Categorising people based on their physical features presents extremely high risks for discrimination and has often been shown to lack scientific basis.[30] Emotion detection also lacks a robust, trustworthy scientific basis. It can be extremely detrimental in sensitive contexts, such as the workplace and educational settings (according to EU regulation), as well as in law enforcement, criminal justice and border control.

Social scoring: AI systems for social scoring are used to classify or evaluate individuals based on social behaviour or known or predicted personal or personality characteristics.[31] These systems can lead to unjust and discriminatory treatment of individuals and groups of people. They can also compromise privacy and lead to profiling based on stereotypes and unrecognised biases, with important repercussions on people's democratic rights and access to adequate standards of living. For these reasons, we are particularly against predictive and profiling systems in law enforcement and criminal justice as well as automatic credit approvals without human oversight.

Cognitive and behavioural manipulation: AI-powered technology can be developed or used to influence human behaviour, such as persuasive marketing practices and systems designed to keep users engaged, with high risks of addictive behaviour. Another type of manipulative practice is the spread of misinformation, such as deepfakes (fake videos and audio material) and synthetic media. These can have dangerous repercussions, not least undermining democratic processes and trust in institutions. While regulating these practices is very difficult, we believe we should collectively refrain from encouraging the development and use of AI systems that can cause substantial harm through cognitive and behavioural manipulation.

We believe the AI applications mentioned above can have a serious negative impact on the wellbeing of individuals, the health of our social fabric and the future of humanity in general. Therefore, people everywhere in the world should be protected from them. In the next chapter, we will carve out the role of financial institutions in promoting responsible AI, based on the principles and considerations outlined above regarding controversial applications.

AI technology and Triodos Bank's transition themes

Triodos Bank's impact strategy revolves around five transition themes: food transition: energy transition; resource transition; wellbeing transition and societal transition. Technology, and therefore AI technology, can be play a vital role in each of the transitions. However, we want to emphasise how AI technology relates to people and society at large. This is why we look at it through the lens of the **wellbeing and societal transitions**. We consider how it affects our social fabric and social foundations, and this leads us to adopt a precautionary approach to technology. We advocate for putting human dignity at the centre of technological development and create the conditions to use technology in a healthy way that supports physical and mental wellbeing. Collectively, we are moving from human-centred to humanity-centred technology. This means that the design and use should focus on reducing inequalities, rather than exacerbating them, and support the building of just, cohesive and peaceful societies. Without holding these principles strong, we only have technology for its own sake.

4. The role of financial institutions

Financial institutions are key actors in our economy due to their double role as money intermediaries and corporate citizens. In both roles, they can (and do!) play a role in relation to various stakeholders. Triodos Bank takes both these roles seriously, and we want to highlight ways that financial institutions with societal wellbeing at heart can foster responsible practices in AI technology.

4.1 Financial institutions as money intermediaries: expectations on companies

In channelling money towards the private sector, financial institutions play a substantial role in setting expectations and incentives for companies. To date, AI technology is largely being developed by private companies, and in parallel, companies around the world are rapidly adopting AI systems for their own operations. As regulations for AI systems are only now being introduced and full implementation will take time, companies are largely in the driving seat in determining practices around AI systems development and use.

Financial institutions like banks and investors typically have two ways of encouraging good business practices. They can screen companies and only select those that display good enough practices for financing and investing. Alternatively, they can engage with companies in portfolio to raise them to higher standards. In the context of AI technology, we believe that responsible financial institutions should take a precautionary approach and demand that the companies they finance and invest in demonstrate accountability regarding the risks related to the technology they develop and use. In practice, this would mean moving in the following direction:

- **Commitment to responsible AI:** Companies that have substantial exposure to AI technology, either as users or developers, should have a public commitment on responsible AI in place. Importantly, these documents or statements should demonstrate substantial awareness by making explicit reference to the most relevant adverse human rights impacts and risks of the systems that the company develops or uses. They should also commit to ethics by design and technology that has humanity at its core.

- **Refrain from harmful activities:** Investors and financiers should expect and demand that companies commit not to use, develop or contribute to the development of AI systems for highly controversial uses and have proper due diligence systems in place to ensure that they fulfil such commitment. At a bare minimum, companies should be expected to adhere to regulations regarding controversial activities. However, financial institutions also have the power to set higher standards than those required by law. By doing so, they send a clear signal to businesses. A clear example of requirement beyond regulation would be demanding no involvement with the production and distribution of lethal autonomous weapons.

- **Strive for transparency and good governance:** Firstly, financial institutions can demand that companies are transparent about which AI systems they develop or use. Without a fundamental level of transparency, further scrutiny is very hard to implement. Companies that are involved in the research and development of AI systems and provide AI-powered products or related services to others, particularly AI system developers, should be able to explain how such systems were designed and trained, which data was used and how the models were tested. Moreover, they should have robust governance mechanisms in place to ensure responsible development and use of AI systems and define clear accountabilities. Companies that have a substantial involvement in AI technology, should be expected to have a dedicated ethics committee that oversees the responsible development and use of AI systems across the company. The committee should have clear accountability and decision-making power and conduct recommended ethical impact assessments (EIA) and fundamental rights impact assessments.

- **Targeted policies in place:** Depending on the nature of the business, financial institutions can demand that companies that use AI systems have targeted policies and governance mechanisms covering the use of AI technology. For example, companies that use AI systems for marketing purposes should do so in the framework of a policy on responsible marketing practices that explicitly addresses risks associated with the use of AI systems.

Engaging with businesses to improve their practices and advocate for best practices in the industry is a viable next step for both investors and banks providing loans to businesses. Investors and banks can engage in dialogue with businesses regarding the need for adequate training for digital literacy and specifically on the use of AI technology and related risks to employees at different levels of the company. They could also discuss the need to provide clear user guidelines for AI-powered applications. If AI is used to cut costs and increase efficiency at the expense of jobs, companies

should transparently explain how they intend to apply relevant safeguards for the people made redundant. Furthermore, investors could encourage their investees to take a clear stance on the issue of the rapid development of AI technology, and support initiatives that call for regulations and a halt to the further development of the most advanced AI technologies until comprehensive regulations and proper governance are in place. As always, requirements for companies involved in AI technology development or use need to be reasonable and proportionate to the size of the business.

What are Triodos Bank's expectations of companies

As a values-based bank and responsible impact investor, Triodos Bank applies high ethical standards and expectations to the companies it finances and invests in. We believe it is time to do the same for companies developing or using AI technology.

At a very minimum, Triodos Bank expects the companies it finances and invests in to comply with the EU AI Act requirements once the regulation is implemented.

Triodos Bank applies its minimum standards to the companies it finances and invests in. This includes exclusionary criteria that implicitly cover most of the controversial uses of AI technology. This means for example excluding companies from investments that facilitate the development and use of AI systems in autonomous weapons (minimum standards on weapons, arms and munitions), and of AI systems that are highly controversial for their negative impacts on human rights (minimum standards on human rights).

We screen companies to ensure that they meet our minimum standards, to the best of our knowledge. If any company in our portfolio is found to be involved in controversies, this triggers a process to clarify the situation with the company and could lead to exclusion from the portfolio. AI controversies can include the provision of AI systems for highly controversial uses, export of critical technology to sanctioned states or entities, algorithmic bias and data privacy and security, as well as infringements of intellectual property rights and regulations and AI-powered manipulative practices.

While some of our standards go beyond what is required by regulation, particularly for autonomous weapons (which are not covered by the EU AI Act), at the moment of publishing this paper we do not formally expect companies to meet requirements beyond those established by the EU AI Act and other relevant regulations.

However, we intend to scrutinise companies' policies and programmes related to AI more closely as our own knowledge and internal practices on the matter improve. Our aim is to stimulate companies to adopt the practices described above. This is why Triodos Investment Management, the investment management arm of Triodos Bank, recently joined the World Benchmarking Alliance's Collective Impact Coalition for Ethical AI. We aimto leverage the power of collaborative engagement from investors to advance responsible AI practices, based on the principles and recommendations outlined in this paper.

4.2 Financial institutions as corporate citizens: own operations and public voice

Financial institutions, like all companies, are corporate citizens. They have a duty to serve their customers, but they also have a broader responsibility to society. And they can also use their voice in the public arena and towards institutions, provided that they do so in an open and transparent manner.

When it comes to a financial institution's own operations, banks and investors mustn't underestimate their role in ensuring that AI systems are developed and used according to the highest ethical and technical standards. The financial sector has been increasingly adopting AI for various applications. When AI systems are applied to credit approval processes or credit scoring, they are currently at high risk of biased outcomes, perpetrating existing biases and leading to credit access and price discrimination[32, 33]. Data use by financial institutions is a source of widespread concern. According to the Dutch Central Bank, when financial institutions use customers' data as a commercial asset it can undermine trust in the institutions and the financial system[34], even if it is done within existing legal frameworks. Cybersecurity is another area of concern. Generative AI can generate more sophisticated phishing messages and facilitate malicious actors impersonating individuals. Moreover, algorithmic trading already represents a significant source of possible instability for the

financial system, which could be further exacerbated by more powerful models. The same is true for AI systems applied to risk management.[35]

Banks and investors should therefore live up to the same standards they require of their investees and financed businesses, and work to ensure proper governance around AI technology, whether they only use it or develop it internally.

As corporate citizens with an undeniably powerful influence, financial institutions can be vocal about the financial and broader societal risks related to the development and dissemination of AI technology. They can take advantage of many opportunities to influence regulators and decision-makers, provided they are transparent and consistent in how they present their views in both private and public settings. In the text box

AI systems at Triodos Bank

Triodos Bank uses AI systems in different departments for transaction monitoring, anti-money laundering and fraud detection, for identity verification using facial recognition as part of clients' digital onboarding, for gathering information for company research and for initial translations of marketing documents.

Our investees and business clients may use AI systems in a wide range of areas to accelerate positive impact, including smart energy management, environmental monitoring, resource efficient design, supply chain tracking and optimisations, diagnostics and drug research.

Triodos Bank has formulated its own internal principles for ethical AI along with guidelines for employees to improve digital literacy and raise awareness of the opportunities and risks related to the use of AI systems.

Triodos Bank's view on the EU AI Act

Triodos Bank directly operates in several EU countries and the UK, and it is active globally through its investment activities. As such, the bank is interested in regulatory developments related to AI technology.
The final draft of the EU AI Act has come a long way in regulating AI-powered technologies, introducing rules to mitigate risks related to AI systems and preparing to create dedicated supervisory authorities. While Triodos Bank is currently not planning to take an active advocacy role on topics related to AI, we are aware of the risks and issues related to AI systems that are not fully addressed by the regulation. In particular, a few points stand out:

- Triodos Bank does not agree with the regulatory exemption on AI systems for national security, as this leaves room for the development and use of AI-powered autonomous weapons.
- Triodos Bank believes that the same requirements that apply to AI systems used in the EU should also apply to AI systems developed in the EU and exported outside of its borders. This would ensure that EU companies and governments do not benefit from exporting potentially abusive technology.
- Triodos Bank does not approve of the use of remote biometric identification in publicly accessible spaces or emotion recognition in sensitive settings such as law enforcement, criminal trials and border control.
- Triodos Bank hopes that further steps in the implementation of the regulation will lead to a clear definition of what constitutes a high-risk AI system, addressing the regulatory loopholes that currently allow companies to decide for themselves whether the regulation applies to the AI systems they develop.

5. Call to action

AI systems are widely used and have many valuable applications. AI technology is not inherently good or bad, but the widespread use of AI systems is a transformative force that is already affecting our economies and needs to be consciously regulated. Although regulations are being introduced, companies still largely have the responsibility of managing risks and impacts on stakeholders related to human rights and business ethics, while developing and deploying AI systems in a humanity-centred way.

Triodos Bank has developed its own views and high-level expectations for the businesses it finances and companies it invests in. These expectations aim to ensure commitment to responsible AI and good practices in AI development and use. Triodos Bank is also working on improving its internal mechanisms and knowledge of AI technology to ensure responsible use.

We recognise the role that investors and banks can play in establishing the responsible development and use of AI technology. Therefore, Triodos Bank calls on other financial institutions to refrain from blindly spurring and profiting from the irresponsible rapid development of AI technology in an unregulated and highly competitive industry. Instead, they should take a precautionary approach by setting expectations and engaging on responsible AI technology. We all have a profound responsibility to understand the power of AI technology and the implications for humanity.

References

1 Gugerty, L. (2006). Newell and Simon's logic theorist: historical background and impact on cognitive modeling. Proceedings of the human factors and ergonomics society annual meeting, Vol. 50, No. 9, pp. 880-884). Sage CA: Los Angeles, CA: SAGE Publications.
2 Medium (December 4, 2020). The first of its kind AI Model- Samuel's Checkers Playing Program.
3 OECD (2024), Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.
4 Washington Post (March 26, 2023). There's no such thing as artificial intelligence.
5 Harvard Business Review (2021). AI regulation is coming.
6 United Nations (October 26, 2023). New UN Advisory Body aims to harness AI for the common good.
7 Politico (October 31, 2023). UK, US slated to announce AI safety partnership.
8 European Commission (October 30, 2023). Hiroshima Process International Code of Conduct for Organisations Developing Advanced AI Systems.
9 BBC News (October 10, 2023). Warning AI industry could use as much energy as the Netherlands.
10 S&P Global Commodity Insights (October 16, 2023). Power of AI: Wild predictions of power demand from AI put industry on edge.
11 Li, P. et al. (2023). Making AI less" thirsty": Uncovering and addressing the secret water footprint of AI models. arXiv preprint arXiv:2304.03271.
12 Forti, V. et al. (2020). The global e-waste monitor 2020. United Nations University (UNU), International Telecommunication Union (ITU) & International Solid Waste Association (ISWA), Bonn/Geneva/Rotterdam, 120.
13 United Nations, General Assembly (1948). Universal Declaration of Human Rights.
14 European Union, European Convention (2000). Charter of Fundamental Rights of the European Union.
15 United Nations, General Assembly (1966). International Covenant on Civil and Political Rights.
16 Varona, D., Suárez, J.L. (2022). Discrimination, bias, fairness, and trustworthy AI. Applied Sciences, 12(12):5826.
17 Harvard T.H. Chan School of Public Health (March 12, 2021). Algorithmic bias in health care exacerbates social inequities — How to prevent it.
18 MIT Technology Review (June 17, 2021). Bias isn't the only problem with credit scores—and no, AI can't help.
19 Dresp-Langley, B. (2023). The weaponization of artificial intelligence: What the public needs to be aware of. Frontiers in Artificial Intelligence, 6:1154184.
20 Forbes (January 6, 2021). The U.S. army's robot tanks will make great bait.
21 Treleaven, P. et al. (2023). The future of cybercrime: AI and emerging technologies are creating a cybercrime tsunami. Social Science Research Network (SSRN).
22 OSCE (2020). Global Conference for Media Freedom: Freedom of the media and artificial intelligence.
23 Harvard Business Review (April 7, 2023). Generative AI has an intellectual property problem.
24 Financial Times (November 6, 2023). What AI means for ESG.
25 Noema (October 13, 2022). The exploited labor behind artificial intelligence.
26 Council on Foreign Relations (October 25, 2023). Governing artificial intelligence: A conversation with Rumman Chowdhury.
27 Pax (October 17, 2023). Increasing complexity - Legal and moral implications of trends in autonomy in weapons systems.
28 Smith, M. J. et al. (1992). Employee stress and health complaints in jobs with and without electronic performance monitoring. Applied Ergonomics, 23(1), 17-27.
29 UNICEF (2019), Faces, fingerprints and feet.
30 Barrett, L. F. et al. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. Psychological Science in the Public Interest, 20(1), 1-68.
31 European Commission (2021), Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts.
32 Garcia, A. C. B. et al. (2023). Algorithmic discrimination in the credit domain: What do we know about it?. AI & Society, 1-40.
33 MIT Technology Review (June 17, 2021). Bias isn't the only problem with credit scores—and no, AI can't help.
34 De Nederlandsche Bank (2019). General principles for the use of artificial intelligence in the financial sector.
35 Danielsson, J. et al. (2022). Artificial intelligence and systemic risk. Journal of Banking & Finance, 140, 106290.

Address
Hoofdstraat 10, Driebergen-Rijsenburg
PO Box 55
3700 AB Zeist, The Netherlands
Telephone +31 (0)30 693 65 00
www.triodos.com
www.triodos-im.com
www.triodos.com/en/regenerative-money-centre

Published
July 2024

Text
Federica Masut and Johanna Schmidt, Triodos Bank

Design and layout
PI&Q, Zeist

Acknowledgments
Triodos Bank would like to thank Iris Muis (Data School, Utrecht University) for the insightful feedback on this paper.

Triodos Bank