Chapter 4
Taking Responsibility

> "How will machines know what we value if we don't know it ourselves?"
>
> John C. Havens

Where we discuss how to ensure that AI systems are developed in a responsible way.

4.1 Introduction

Now that we have discussed what Artificial Intelligence (AI) is and how ethical theories can be relevant to understanding the impact of AI, we turn our attention to the practical development of AI systems in ways that are aligned with human values and can ensure trust.

AI has huge potential to bring accuracy, efficiency, cost savings and speed to a whole range of human activities and to provide entirely new insights into behaviour and cognition. However, the way AI is developed and deployed for a great part determines how AI will impact our lives and societies. For instance, automated classification systems can deliver prejudiced results and therefore raise questions about privacy and bias, and the autonomy of self-driving vehicles raises concerns about safety and responsibility. AI's impact concerns not only the the research and development directions of AI, but also how these systems are introduced into society. There is debate concerning how the use of AI will influence labour, well-being, social interactions, healthcare, income distribution and other areas of social relevance. Dealing with these issues requires that ethical, legal, societal and economical implications are taken into account.

AI will affect everybody. This demands that the development of AI systems ensures inclusion and diversity, that is, truly considers all humankind

V. Dignum, *Responsible Artificial Intelligence*, Artificial Intelligence: Foundations, Theory, and Algorithms, https://doi.org/10.1007/978-3-030-30371-6_4

when determining the purpose of the systems. Therefore, Responsible AI also requires informed participation of all stakeholders, which means that education plays an important role, both to ensure that knowledge of the potential impact of AI is widespread, as well as to make people aware that they can participate in shaping societal development. At the core of AI development should lie the idea of ‘AI for Good’ and ‘AI for All’. We will discuss further this issue in Chapter 7.

Researchers, policymakers, industry and society at large, all are increasingly recognising the need for design and engineering approaches that ensure the safe, beneficial and fair use of AI technologies, that consider the implications of ethically and legally relevant decision-making by machines, and that evaluate the ethical and legal status of AI. These approaches include the methods and tools for system design and implementation, governance and regulatory processes, and consultation and training activities that ensure all are heard and able to participate in the discussion.

In this endeavour, it is important to realise that AI does not stand by itself, but must be understood as part of socio-technical relations. A **responsible approach to AI is needed**. One that not only ensures that systems are developed in a good way, but also that they are developed for a good cause. The focus of this chapter is on understanding what such an approach should look like, who are the responsible parties and how to decide on which systems can and should be developed.

Responsible Artificial Intelligence is concerned with the fact that decisions and actions taken by intelligent autonomous systems have consequences that can be seen as being of an ethical nature. These consequences are real and important, independently of whether the AI system itself is able to reason about ethics or not. As such, Responsible AI provides directions for action and can maybe best be seen as a code of behaviour — for AI systems, but, most importantly, for us.

> **Responsible AI** is more than the ticking of some ethical ‘boxes’ or the development of some add-on features in AI systems.

In all cases, the processes by which systems are developed entail a long list of decisions by designers, developers and other stakeholders, many of them of an ethical nature. Typically, many different options and decisions are taken during the design process, and in many cases there is not one clear ‘right’ choice. These decisions cannot just be left to be made by those who engineer the systems, nor to those who manufacture or use them, but require societal awareness and informed discussion. Determining which decisions an AI system can take, and deciding how to develop such systems, are both ethically based decisions that require a responsible approach. Most of all, this means that these choices and decisions must be explicitly reported and open for inspection. This is fundamentally different from but at least as important as the discussion of whether or not AI systems are capable of ethical reasoning, which will be discussed further in Chapter 5.

At all levels and in all domains, businesses and governments are, or will soon be, applying AI solutions to a myriad of products and services. It is fundamental that the general public moves from passively adopting or rejecting technology to being in the forefront of the innovation process, demanding and reflecting on the potential results and reach of AI. The success of AI is therefore no longer a matter of financial profit alone but how it connects directly to human well-being. Putting human well-being at the core of development provides not only a sure recipe for innovation but also both a realistic goal as well as concrete means to measure the impact of AI. We will discuss the issue of education further in Chapter 6.

In this chapter, we first examine how Responsible Research and Innovation (RRI) approaches support the development of technologies and services, and how Responsible AI can learn from such approaches. We then present the grounding principles of Responsible AI, namely Adaptability, Responsibility and Transparency. We then introduce the Design for Values methodology to guide the development of Responsible AI systems. Finally, we discuss how these principles can be integrated into a system development life cycle framework.

4.2 Responsible Research and Innovation

Given the fundamental and profound impact of AI systems in human society, the development of AI technology cannot be done in isolation from its sociotechnical context. A full understanding of societal, ethical and policy impacts requires us to analyse the larger context of its implementation. In this section, we describe how a Responsible Research and Innovation (RRI) vision can be applied to the development of AI systems.

RRI describes a research and innovation process that takes into account effects and potential impacts on the environment and society.

There are many approaches to and views on RRI, some focused on environmental effects and others on societal impact, but fundamentally RRI is grounded on participation. That is, RRI requires that all societal actors (researchers, citizens, policymakers, business, non-governmental organisations, etc.) work together during the whole research and innovation process in order to better align both the process and its outcomes with the values, needs and expectations of society. RRI should be understood as a continuous process that does not stop at the drawing table but continues through the whole process until the introduction of resulting products and services into the market.

4.2.1 Understanding the RRI Process

RRI has been defined as a "transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products" [131].

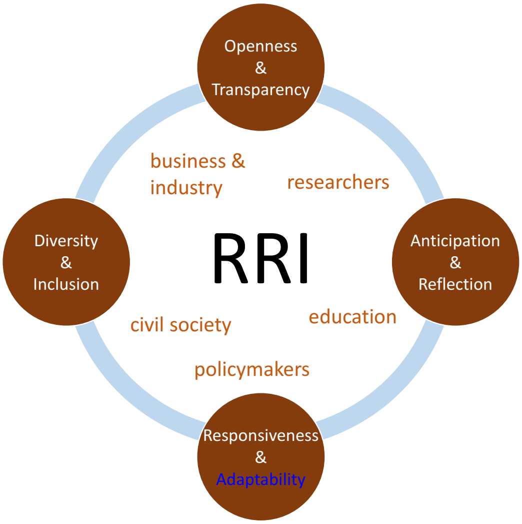

Figure 4.1: The Responsible Research and Innovation process

The RRI process[1] is depicted in Figure 4.1. The process should ensure that all parties participate during the process of defining research and innovation directions. The issue of *Diversity and Inclusion* refers to the need to involve a wide range of stakeholders in the early innovation process, to ensuring diversity and inclusion within system development teams and stakeholders, broadening and diversifying the sources of knowledge, expertise, disciplines and perspectives. *Openness and Transparency* require open, clear communication about the nature of the project, including funding/resources, decision-making processes and governance. Making data and results openly available ensures accountability and enables critical scrutiny, which contribute to build public trust in research and innovation. *Anticipation and Reflexivity* are needed to understand the current context for research and innovation from a diverse range of perspectives. They imply the need to consider the environmental, economic and social impact in the short and long term. They also refer to the need to identify and reflect on individual and institutional values, assump-

[1] Adapted from https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation

tions, practices and responsibilities. Finally, *Responsiveness and Adaptiveness* are needed to deal with a dynamic context and with possibly emerging knowledge, data, perspectives, views and norms. They require an ongoing interaction with stakeholders and an ability to change patterns of thought and behaviour as well as roles and responsibilities in response to emerging perspectives and insights in the context.

4.2.2 RRI in the Development of AI Systems

Advances in computational autonomy and machine learning are rapidly enabling AI systems to decide and act without direct human control. This means that special attention should be given to the analysis of the evolution of the system, and how to ensure that it does not lead to undesirable effects.

A responsible approach to Artificial Intelligence is needed to ensure the safe, beneficial and fair use of AI technologies, to consider the ethical implications of decision-making by machines, and to define the legal status of AI. The focus of this process should be on ensuring wide societal support for the AI applications being developed, which is achieved by focusing on human values and well-being. Moreover, for the whole of society to truly be able to benefit from all AI developments, education and an honest and accessible AI narrative are needed. Only then, will everybody be able to understand AI's impact and truly benefit from its results. RRI in AI should therefore include steps to ensure proper and wide education of all stakeholders present and future, alongside governance models for responsibility in AI. We will discuss these issues further in Chapter 6.

Ensuring that systems are designed responsibly contributes to our trust in their behaviour, and requires accountability, i.e. being able to explain and justify decisions, and transparency, i.e. being able to understand the ways systems make decisions and the data being used. To this effect, we have proposed the principles of Accountability, Responsibility and Transparency (ART). ART follows a Design for Values approach as outlined in Section 4.4 to ensure that human values and ethical principles, and their priorities and choices are explicitly included in the design processes in a transparent and systematic manner.

True responsibility in AI is not just about how we design these technologies but how we define their success. Even if a system is built to be safe and robust, comply with legal regulations, and be economically viable, it can still have dramatic negative consequences on human and societal well-being. Issues such as mental health, emotions, identity, autonomy, or dignity, which are key components of what makes us human, are not those that are measured by the usual Key Performance Indicators. Multiple metrics are already in use that measure well-being through Indicators such as the United Nations'

Human Development Index[2] and the Genuine Progress Indicator.[3] Business leaders and governments alike have been working for years to implement a Triple Bottom Line[4] mindset honouring societal and environmental issues along with financial concerns. Many are also aligning business with the UN's Sustainable Development Goals.[5] Responsible AI development must thus include the consideration of measuring performance in terms of human and societal well-being.

4.3 The ART of AI: Accountability, Responsibility, Transparency

Following the characterisation of AI given in Chapter 2, in this chapter we assume an intelligent system (or agent) to be a system that is capable of perceiving its environment and deliberating how to act in order to achieve its own goals, assuming that other agents possibly share the same environment. As such, AI systems are characterised by their *autonomy* to decide on how to act, their ability to *adapt*, by learning from the changes effected in the environment, and how they *interact* with other agents in order to coordinate their activities in that environment [57, 104].

These properties enable agents to deal effectively with the kinds of environments in which we live and work: environments that are unpredictable, dynamic in space and time, and where one is often faced by situations one has never encountered before. If AI systems are capable and expected to act in such environments, we need to be able to trust that they will not exhibit undesirable behaviour. Or, at least, we need to limit the effects of unexpected behaviour. Therefore, design methodologies that take these issues into account are essential for trust and the acceptance of AI systems as part of a complex socio-technical environment.

To reflect societal concerns about the impact of AI, and to ensure that AI systems are developed responsibly, and incorporating social and ethical values, these characteristics of autonomy, adaptability and interaction, as discussed in Chapter 2, should be complemented with design principles that ensure trust. In [40] we have proposed to complement autonomy with *responsibility*, interactiveness with *accountability*, and adaptation with *transparency*. These characteristics relate most directly to the technical system. However, the impact and consequences of an AI system reach further than the technical system itself, and as such the system should be seen as a socio-technical system, encompassing the stakeholders and organisations involved.

[2] See http://hdr.undp.org/en/content/human-development-index-hdi

[3] See https://en.wikipedia.org/wiki/Genuine_progress_indicator

[4] See https://en.wikipedia.org/wiki/Triple_bottom_line

[5] See https://sustainabledevelopment.un.org/

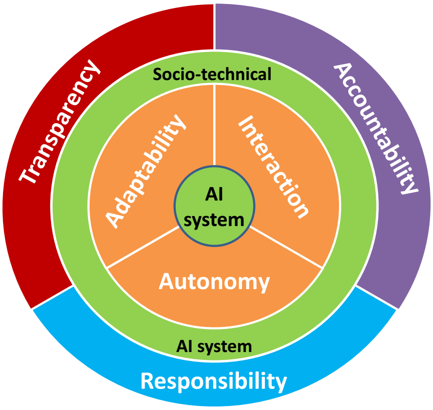

Figure 4.2: The ART principles: Accountability, Responsibility, Autonomy

The ART principles for responsible and trustworthy AI apply then to the AI socio-technical system. That is, addressing ART will require a socio-technical approach to design, deployment and use of systems, interweaving software solutions with governance and regulation. Moreover, even though each of the ART principles can apply to all aspects of AI systems, each is imperative for a specific characteristic, as is depicted in Figure 4.2. That is, truly responsible AI cannot have autonomy without some form of responsibility, interaction without accountability, nor adaptability without transparency. From the perspective of system development, ART requires new methods that support the integration of the ethical and societal impact of AI systems into the engineering process. Above all, ART requires training and awareness of all stakeholders, including researchers, designers, programmers, managers, providers, users, and all of society to enable each of them to understand and assume their role in the overall process.

The ART principles for Responsible AI can be summarised as follows:

- **Accountability** refers to the requirement for the system to be able to explain and justify its decisions to users and other relevant actors. To ensure accountability, decisions should be derivable from, and explained by, the decision-making mechanisms used. It also requires that the moral values and societal norms that inform the purpose of the system as well as their operational interpretations have been elicited in an open way involving all stakeholders.
- **Responsibility** refers to the role of people themselves in their relation to AI systems. As the chain of responsibility grows, means are needed to link the AI systems' decisions to their input data and to the actions of stakeholders involved in the system's decision. Responsibility is not just about making rules to govern intelligent machines; it is about the

whole socio-technical system in which the system operates, and which encompasses people, machines and institutions.

- **Transparency** indicates the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and the provenance and dynamics of the data that is used and created by the system. Moreover, trust in the system will improve if we can ensure openness of affairs in all that is related to the system. As such, transparency is also about being explicit and open about choices and decisions concerning data sources and development processes and stakeholders. Stakeholders should also be involved in decisions about all models that use human data or affect human beings or can have other morally significant impact.

Given this characterisation, we further define the ART principles in the following sections of this chapter. As a whole, these principles inform the design of AI systems. That is, ART imposes requirements on AI systems' design and architecture that will condition the development process and the systems' architecture.

Note that there is a fundamental difference between accountability and responsibility, even if these terms are often used interchangeably, as synonyms. Putting it simply, accountability refers to the ability to explain, or report on, one's role in events or actions, whereas responsibility is the duty to answer for one's actions. Responsibility entails liability and exists before the task or action is done. Accountability is only evident after the action is done, or not done. When a person delegates some task to an agent, be it artificial or human, the result of that task is still the responsibility of the delegating person (principal), who is the one who will be liable if things don't go as expected. The agent however, must be able to report on how the task was executed, and to explain eventual problems with this execution. This is the basis of the principal-agent theory that is often used to explain the relationship between people and autonomous systems [48].

4.3.1 Accountability

Accountability is the first condition for Responsible AI. Accountability is the capability to give account, i.e. to be able to report and explain one's actions and decisions. A key factor for people to be willing to trust autonomous systems is that the system is able to explain why it took a certain course of action [134, 70].[6] Another important aspect of accountability is to be able to rely on a safe and sound design process that accounts for and reports on op-

[6] See also GDPR regulation: http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf.

tions, choices and restrictions about the system's aims and assumptions [60]. In the following we further discuss these two aspects of accountability.

Explanation is relevant for trusting AI systems for a number of reasons. Firstly, explanation can reduce the opaqueness of a system, and support understanding of its behaviour and its limitations. Secondly, when things do go wrong, *post-mortem* explanation, using some sort of logging systems (such as the black boxes used in aviation) can help investigators understand what went wrong.

Even though explanation is of particular importance when the AI system has made an error, it is also crucial when the system does something good, but unexpected, e.g. it takes a course of action that would not occur to a human, but is appropriate, either because the human is not aware of information, or because they don't think that way. And, even if to err is human, decision-making by an AI system seems to be held to a higher standard than human decision-making [86]. One reason for this could be that some of the justifications for a mistake, such as feeling distracted or confused, are only valid arguments or 'excuses' for people and do not apply to machines. Another reason for the need for explanations is that, machines are assumed to be incapable of moral reasoning, unlike humans who are assumed, by default, to be moral agents. Given this lack of moral agency, and also of empathy, of machines, society will require a proof or certification of the (ethical) reasoning abilities of a machine, or at least a guarantee about the scope of the decisions that the system can make. Currently we do not have any clear description let alone consensus on the nature of these proofs [42], which will require much more research.

In developing explanation mechanisms, it is important to be mindful that the explanations should be comprehensible and useful to a human, and therefore we should consider the relevant social sciences literature [92]. According to Miller [92] explanations should be *contrastive*, i.e. answer questions of the form "why did you do X ... instead of Y?"; *selective*, i.e. select relevant factors and present those; and *social*, i.e. presented relative to what the explainer believes the listener (i.e. explainee) knows. Given that the explanation processes can be seen as a conversation between the system and its user, it should therefore also follow Grice's conversation maxims of quality, quantity, manner and relevance [67].

Accountability also means that we understand the rationale beyond the design of the system. As such, the system's design should follow a process that is sensitive to the societal, ethical and legal impact, and to the characteristics of the context in which it will operate. Decisions made during the design process have ethical implications. That is, design is not only enabling of function, but also constitutive: it shapes practices and society in important ways. In order to take normative considerations into the design process, the first step is to identify and articulate the overall (ethical) objectives of the system, the human values at stake in a particular design context, and the stakeholders that are affected by the AI system being designed. Design for Values methodology approaches [60, 125] have been successfully applied to the design of many different technologies, and have the potential to guarantee the accountable development of AI systems. We will further discuss the Design for Values methodology for development of AI systems in Section 4.4.

4.3.2 Responsibility

Currently, never a day goes by without news and opinion articles concerning the capabilities of AI systems and raising questions about their role in society. This raises many questions about responsibility for the system and by the system. What does it mean for an AI system to make a decision? What are the moral, societal and legal consequences of their actions and decisions? Can an AI system be held responsible for its actions? How can these systems be controlled once their learning capabilities bring them into states that are possibly only remotely similar to their initial design?

In order to answer these questions, it must first and foremost be clear that whatever the system's level of autonomy, social awareness and ability to learn, AI systems are tools, i.e. artefacts, constructed by people for a given purpose. That is, even if the system is designed for accountability and transparency, human responsibility cannot be replaced. This implies that, even if the system will be able to modify itself by learning from its context of use, it does so based on that purpose. Ultimately, we, people, are the ones determining that purpose.

Theories, methods, and algorithms are needed to integrate societal, legal and moral values into technological developments in AI, at all stages of development (i.e. analysis, design, construction, deployment and evaluation). These frameworks must deal with the autonomic reasoning of the machine about issues that we consider to have ethical impact, but most importantly, must identify who are the 'we' that are the focus and the guides of design decisions, and ensure the apportionment of liability for the machine's decisions.

In Chapter 2, we have discussed the issue of autonomy and how that is understood and dealt with in AI. In particular, in Section 2.3.2 we reflected on the fact that in most cases, the autonomy of an AI system refers to its autonomy to develop its own plans and to decide between its possible actions. Therefore these actions can in principle be traced back to some user instruction (e.g. personalisation preferences), manufacturing setting or design choice. Even if the system has evolved, by learning from its interaction with the environment, what it learns is determined by the purpose for which it was build, and the functionalities it is endowed with. A robot vacuum cleaner will never by itself learn how to do the laundry or clean the windows. Nor will a self-driving car learn how to fly, even if that may be the most suitable answer to a user's request. Not only are these systems limited by their physical characteristics, they are also limited in their cognitive abilities: the way a system learns to use its input is determined by the purpose the system was build for.

Although currently much discussion goes on concerning the responsibility of the AI system itself, where it concerns current state-of-the-art systems, basically two things can happen either: (i) the machine acts as intended and therefore the responsibility lies with the user, as is the case with any other tool [26]; or (ii) the machine acts in an unexpected way due to error or malfunction, in which case the developers and manufacturers are liable. The fact that the action of the machine is a result of learning cannot be seen as removing liability from its developers, as this is in fact a consequence of the algorithms they've designed. This is, however, a consequence that can be hard to anticipate and assure, which is why methods to continuously assess the behaviour of a system against given ethical and societal principles are needed. These include methods to prove, either by verification or by observation, that AI systems behave ethically [113, 30, 37].

Note that the capability to learn, and thus to adapt its behaviour, is an expected characteristic of most AI systems. By adapting, the system is then functioning as expected. This makes the clear specification of objectives and purpose even more salient, as well as the availability of tools and methods to guarantee that learning doesn't go awry. Current research on this issue includes the definition of fall-back procedures (e.g. the system switches off, or requests the intervention of a human operator), and testing the system for vulnerability to adversarial attacks (e.g. by exposing the system to various malignant situations).

Responsibility refers thus to the role of people as they develop, manufacture, sell and use AI systems.

Responsibility in AI is also an issue of regulation and legislation, in particular where it respects liability. Governments decide on how product liability laws should be regulated and implemented, and courts of law on how to interpret specific situations. For example, who will be liable if a medicine pump modifies the amount of medicine being administered? Or when a predictive policing system wrongly identifies a crime perpetrator? The builder of the software? The ones that have trained the system to its current context of use? The authorities that authorised the use of the system? The user that personalised the system's decision-making settings to meet her preferences? These are complex questions, but responsibility always relates to the humans involved, and liability can often, for a large part, be handled by existing regulations on product and service liability. Existing laws describe how and when manufacturers, distributors, suppliers, retailers and others who make products available to the public are held responsible for the injuries and problems that those products cause, and can, to some extent, ensure liability in the case of AI applications. However, there are also many arguments for developing new regulation specifically for AI, ranging from mere modifications of existing liability laws to more extreme approaches such as granting AI legal personhood, so that one can identify the responsible party. The later has been suggested, amongst others, by the European Parliament, in a motion from February 2017. This motion, focusing on smart robots, proposed the creation of a specific legal status for robots "so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently" [50]. It must be noted that the European Parliament was not aiming at recognising robots as conscious entities, or living systems like rivers and forests, but as legal persons, with responsibilities, rights and obligations, for the aim of facilitating business and legal processes. Nevertheless, this proposal was strongly contested by many researchers and practitioners, based on technical, as well as ethical and legal arguments. In an open letter [1], experts in AI and Robotics indicated that *"from a technical perspective, this statement offers many bias based on an overvaluation of the actual capabilities of even the most advanced robots, a superficial understanding of unpredictability and self-learning capacities [and] a robot perception distorted by Science-Fiction and a few recent sensational press announcements"*. Moreover, these experts express their condemnation of this proposal based on existing legal or ethical precedents, given that in such a case, *"the robot would then hold human rights, such as the right to dignity, the right to its integrity, the right to remuneration or the right to citizenship, thus directly confronting the Human rights. This would be in contradiction with the Charter of Fundamental Rights of the European Union and the Convention for the Protection of Human Rights and Fundamental Freedoms"*. Moreover, legal personhood models imply the existence of human persons behind the legal person to represent and direct it, which is not the case for AI or robots. All in all, the area of AI regulation is one where much activity can be expected in the coming years.

Finally, responsibility also relates to design decisions about embodiment and human-likeness of AI systems. When and why should an AI system exhibit anthropomorphic characteristics? Just recently, widespread public outcry followed the release of the Google Duplex demo that showed a chatbot that would behave in a way that led its user to believe it was a person. There is also much discussion around the robot Sophia from Hanson Robotics[7] and the meaning of its interventions at the United Nations assembly or the European Parliament, just to name a few. These events lend the system a level of expectation in terms of its intelligence and autonomy that it just does not possess. In fact, it is not the robot Sophia that speaks to the United Nations or other audiences, but the PR department of Hanson Robotics. The use of a,

[7] See https://www.hansonrobotics.com/

seemingly autonomous, puppet should not obfuscate this fact. Even though it is well known that people will tend to anthropomorphise all types of objects (toys, cars, computers, ...), the deliberate use of human-like characteristics in the design of AI systems requires much attention and deep understanding of the consequences of these choices. In particular when dealing with vulnerable users, such as young children or dementia patients, huge responsibility lies with the designers for their choices of which human-like characteristics they implement in the system.

The more realistic these human-like characteristics are, the higher the expectations of the capabilities of the system are. On the other hand, deliberately attempting to impersonate another's identity can and will be a source of liability for designers and manufacturers of which they should be well aware.

4.3.3 Transparency

The third ART principle is transparency. Currently, much effort is put into Algorithmic Transparency, the principle that the factors influencing the decisions made by algorithms should be visible, or transparent, to the people who use, regulate and are impacted by those algorithms. In a strict sense, this is a red herring, solvable by making code and data open for inspection. However, this 'solution' does not suffice: not only may it violate intellectual property and business models of those that develop the algorithms, but mostly, the code would not make much sense to most users.

Opacity in Machine Learning, the so-called 'black-box' algorithms, is often mentioned as one of the main impediments to transparency in Artificial Intelligence. Machine Learning algorithms, as we discussed in Section 2.4.2, are developed with the main goal of improving functional performance. Even though each component function is usually not very complex (often implementing some statistical regression method, the sheer number of components renders the overall system intractable to analyse and verify. These complex algorithms are optimised to provide the best possible answer to the question at hand (e.g. recognise pictures, analyse x-ray images or classify text) but they do it by fine-tuning outputs to the specific inputs, approximating a function's results without giving any insights into the structure of the function that is being approximated.

On the other hand, Machine Learning algorithms are trained with and reason about data that is generated by people, with all its shortcomings, biases, and mistakes. To promote transparency, an increasing number of researchers, practitioners and policymakers are realising the need to deal with bias in data and algorithms. However, this is easier said than done. All people use heuristics to form judgements and make decisions. Heuristics are simple rules that enable efficient processing of inputs, guaranteeing a usually appropriate reaction. Heuristics are culturally influenced and reinforced by practice, which means that heuristics can induce bias and stereotypes when they reinforce a misstep in thinking, or a basic misconception of reality. Moreover, sometimes bias is not a misstep, but reflects aspects of reality, e.g. the relation between socio-economical level and crime rates, or access to credit. In fact, even if particular attributes are not part of a dataset, these can still be learned and used as a proxy by AI systems, based on this type of correlation, reinforcing racial differences.[8] Therefore, bias is inherent in human thinking and an unavoidable characteristic of data collected from human processes.

Because the aim of current Machine Learning algorithms is to identify patterns or regularities in data, it is only natural that these algorithms will follow the bias existing in the data. In fact, data reflects aspects of reality, such as, e.g. correlations between race and address. So, even if it may be illegal to use certain attributes in decision-making, such as race, these correlations are discovered by the Machine Learning algorithm, and the system can discover them and use them as a proxy, thus reinforcing bias. The aim of so-called algorithmic transparency is to ensure that the machine will not be prejudiced, i.e. act on these biases in the data. Removing the algorithmic black box, that is, providing means to inspect and evaluate the algorithms used, will not eliminate the bias. You may be able to get a better idea of what the algorithm is doing but it will still enforce the biased patterns it 'sees' in the data. Another complexity in the attempt to remove bias from data is that there are different measures of bias, and they are in tension. Nevertheless, Machine Learning can help identify overt and covert bias that we may not be aware was reflected in data. Besides bias, other problems with data include incompleteness (not enough information about all of the target group), bad governance models (resulting in tampering and loss of data), and outdatedness (data no longer representative of the target group or context). Transparency is also needed to deal with these.

Transparency may be better served by openness and control over the whole learning and training process[9] than by removing the algorithmic black box. Trust in the system will improve if we can ensure openness of affairs in all that is related to the system. This can be done by applying software- and requirement-engineering principles to the development of AI systems. By ensuring the continuous and explicit reporting of the development process, decisions and options can be analysed, evaluated and, if necessary, adapted. The following analysis guidelines, exemplify the type of information that must be maintained and made available for inspection by stakeholders in order to support openness and transparency. The checklist in Figure 4.3 describes possible questions to be considered to ensure transparency of design processes.

[8] For more on this issue, see e.g. [97]

[9] cf. Figure 2.4 for an overview of the Machine Learning process.

Checklist for Transparency

1. Openness about data
   - What type of data was used to train the algorithm?
   - What type of data does the algorithm use to make decisions?
   - Does training data resemble the context of use?
   - How is this data governed (collection, storage, access)
   - What are the characteristics of the data? How old is the data, where was it collected, by whom, how is it updated?
   - Is the data available for replication studies?
2. Openness about design processes
   - What are the assumptions?
   - What are the choices? And the reasons for choosing and the reasons not to choose?
   - Who is making the design choices? And why are these groups involved and not others?
   - How are the choices being determined? By majority, consensus, is veto possible?
   - What are the evaluation and validation methods used?
   - How is noise, incompleteness and inconsistency being dealt with?
3. Openness about algorithms
   - What are the decision criteria we are optimising for?
   - How are these criteria justified? What values are being considered?
   - Are these justifications acceptable in the context we are designing for?
   - What forms of bias might arise? What steps are taken to assess, identify and prevent bias?
4. Openness about actors and stakeholders
   - Who is involved in the process, what are their interests?
   - Who will be affected?
   - Who are the users, and how are they involved?
   - Is participation voluntary, paid or forced?
   - Who is paying and who is controlling?

Figure 4.3: Checklist for Transparency

Many of these issues can be addressed by applying proper Software Engineering procedures to the development of AI systems. According to the IEEE, software engineering is *"the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software."* This ensures that stakeholder requirements[10] are collected and documented.

[10] Requirements elicitation refers to both functional and non-functional requirements, which include the values that the system should enforce and/or follow. See more in Section 4.4 on Design for Values.

Moreover, the use of systematic, methodical, quantifiable methods supports comparative analysis, support code maintenance and testing strategies and allows the precise specification of data governance and provenance.

In any case, there is a need to rethink the optimisation criteria for Machine Learning. As long as the main goal of algorithm design is to improve functional performance, algorithms will remain black boxes. Demanding a focus on ensuring ethical principles and putting human values at the core of system design calls for a mind-shift of researchers and developers towards the goal of improving transparency rather than performance, which will lead to a new generation of algorithms. This can be enforced by regulation, but also supported by education. We will discuss this issue further in Chapter 6.

4.4 Design for Values

In this section, we discuss practical ways through which the ART principles described in the previous section can direct the development of AI systems. Design for Values is a methodological design approach that aims at making moral values part of technological design, research and development [124]. Values are typically high-level abstract concepts that are difficult to incorporate in software design. In order to design systems that are able to deal with moral values, values need to be interpreted in concrete operational rules. However, given their abstract nature, values can be interpreted in different ways. The Design for Values process ensures that the link between values and their concrete interpretations in the design and engineering of systems can be traced and evaluated.

During the development of AI systems, taking a Design for Values approach means that the process needs to include activities for (i) the identification of societal values, (ii) deciding on a moral deliberation approach (e.g. through algorithms, user control or regulation), and (iii) linking values to formal system requirements and concrete functionalities [5].

AI systems are computer programs, and therefore developed following software engineering methodologies. But, at the same time, fundamental human rights, including respect for human dignity, human freedom and autonomy, democracy and equality, must be at the core of AI design. Traditionally, limited attention is given to the role of human values and ethics in the development of software. The link between values and the application being developed is left implicit in the decisions made and the choices taken. Even though ethical principles and human values are at the basis of the system's requirements, the requirements elicitation process only describes the resulting requirements and not the underlying values. The problem with this process is that, due to their abstract nature, values can be translated into design requirements in more than one way. If the values and their translation to requirements are left implicit in the development process, one cannot analyse the decisions that led to the specific definition chosen, and, moreover, one loses the flexibility of using alternative translations of those values.

Figure 4.4 depicts the high-level Design for Values approach for AI systems.

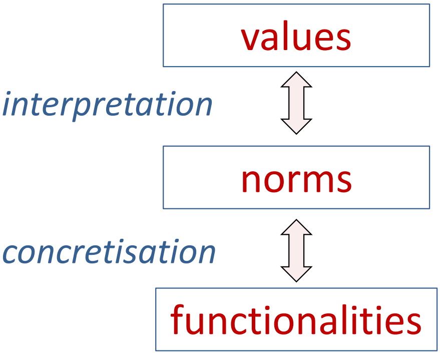

Figure 4.4: From values to norms to functions, and back

To understand the Design for Values approach, consider for example the development of a system to decide on mortgage applications. A **value** that can be assumed for this system is *fairness*. However, fairness can have different normative interpretations, i.e. it can be interpreted by different **norms**, or rules, e.g. *equal access to resources*, or *equal opportunities*, which can lead to very different actions. For instance, in a very simplistic case, giving everybody equal access to a given property would mean that the decision would be done based on the value of the property, whereas equal opportunities would mean that decisions would be taken based on e.g. income and age independently of the value of the property. It is therefore necessary to make explicit which interpretation(s) the design is using. This decision may be informed by domain requirements and regulations, e.g. a choice for *equal opportunities* meets the legal requirement established in national law, but it may also be due to some preference of the person or team creating the system.

We also need to be explicit about how norms are implemented in the system. This will depend on the context but is also influenced by the personal views and cultural background of those deciding on the design. For instance, the Machine Learning literature identifies different implementations of the equal opportunities view of fairness, e.g. demographic parity[11] [98] or equal odds[12] [47] amongst others. These **functionalities** are quite different in their results. Without being explicit about which approach is taken to implement

[11] Demographic parity means that a decision, e.g. accepting or denying a mortgage application, is independent of a given attribute, e.g. gender, which is called the protected attribute.

[12] Equal odds aims at balancing classification errors across protected attributes, towards achieving equal false positive rates, equal false negative rates, or both.

the concept of fairness, it is impossible to compare different algorithms or to understand the implications of their decisions towards different population groups.

The Design for Values approach enables us to formalise these choices and their links to support verification and adaptation in case motivating views change [3]. In the description above, we followed a top-down view of the Design for Values process, which indicates how norms, and functionalities, come about based on given values, or norms. So, e.g. the norm of equal opportunities is there *for-the-sake-of* fairness [122]. This relation can also be turned around to indicate that the norm of equal opportunities *counts-as* fairness in a given context [109, 76].

Precise interpretations, using formal verification mechanisms, are needed both to link values to norms, as well as to transform these norms into concrete system functionalities. Work on formal normative systems proposes a representation of such interpretations based on the formal concept of *counts-as*, where the relation X *counts-as* Y is interpreted as a subsumption that holds only in relation to a specific context [68, 3].

Formalising and making these links explicit allows for improvements in the traceability of (the effects of) the values throughout the development process. Traceability increases the maintainability of the application. That is, if a value needs to be implemented differently, the explicit links between values and the application make it much easier to determine which parts of the application should be updated. In the same, way, if one needs to change or update a system functionality, it is necessary to be able to identify what are the norms and values that this functionality is associated with, and ensure that changes maintain these relations intact. Moreover, the relation between values and norms is more complex than a mere 'translation', but requires also that there is knowledge about the concepts and meanings that hold in the domain, i.e. the ontology of the domain. For instance, whether something counts as personal data and should be treated as such depends on how the application domain interprets the term 'personal data' [127].

A Design for Values approach provides guidelines to how AI applications should be designed, managed and deployed, so that values can be identified and incorporated explicitly into the design and implementation processes. Design for Values methodologies therefore provide means to support the following processes:

- Identify the relevant stakeholders;
- Elicit values and requirements of all stakeholders;
- Provide means to aggregate the values and value interpretations from all stakeholders;
- Maintain explicit formal links between values, norms and system functionalities that enable adaptation of the system to evolving perceptions and justification of implementation decisions in terms of their underlying values;

- Provide support to choose system components based on their underlying societal and ethical conceptions, in particular when these components are built or maintained by different organisations, holding potentially different values.

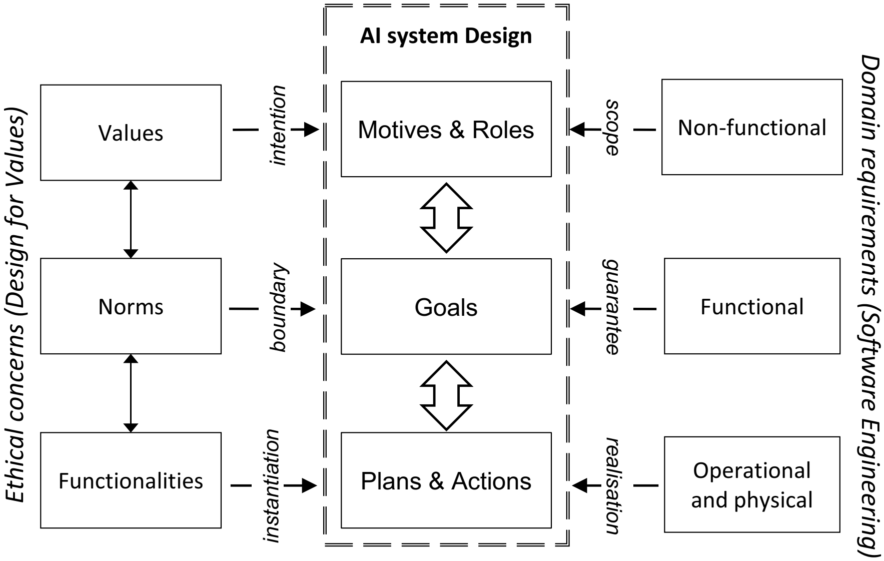

Figure 4.5: Responsible AI design: integrating ethical concerns and domain considerations in the design of AI applications

These issues point to the need for a multi-layered approach to software development where the links to the values are kept explicit. In the following, we present a possible design methodology for Responsible AI, based on the Value Sensitive Software Development (VSSD) framework proposed by [5]. In software design, architectural decisions capture key design issues and the rationale behind chosen solutions. These are conscious and purposeful development decisions concerning the application as a whole, which impact (non-functional) characteristics such as software quality attributes. A fundamental result of software engineering methods is ensuring that these architectural decisions are made explicit.

This framework, depicted in Figure 4.5, connects traditional software engineering concerns with a Design for Values approach, to inform the design of AI systems. On the one hand, and as described above, Design for Values (left-hand side of the figure) describes the links between values, norms and system functionalities. On the other hand, domain requirements (right-hand side of the figure) shape the design of software systems in terms of the functional, non-functional and physical/operational demands of the domain. An

AI system must obey both orientations, i.e. meet domain demands and at the same time ensure alignment with social and ethical principles.

By structuring the design of an AI system in terms of high level motives and roles, specific goals, and concrete plans and actions, it becomes possible to align with both the Design for Values and Software Engineering approaches. As such, at the top level, values and non-functional requirements will inform the specification of the motives and roles of the system by making clear what is the intention of the system and its scope. Norms will provide the (ethical-societal) boundaries for the goals of the system, which at the same time need to guarantee that functional requirements are met. Finally, the implementation of plans and actions follows a concrete platform/language instantiation of the functionalities identified by the Design for Values process while ensuring operational and physical domain requirements. These decisions are grounded on both domain characteristics and the values of the designers and other stakeholders involved in the development process. Taking a Design for Values perspective, it becomes possible to make explicit the link to the values behind architectural decisions. In parallel, the system architecture must also reflect the domain requirements, which describe specific contextual concerns. Making these links explicit allows for improvements in the traceability of values throughout the development process, which increases the maintainability of the application. That is, if a value is to be interpreted differently, having explicit links between the values and the functionalities that contribute to the realisation of that value makes it much easier to determine which parts of the application should be updated.

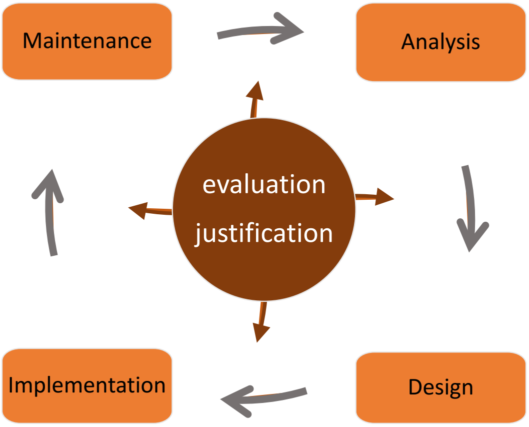

Figure 4.6: The Responsible Development Life Cycle for AI systems

A responsible use of AI reduces risks and burdens, and ensures that societal and ethical values are central to development. It is, however, not always obvious to organisations and developers how best to approach Responsible

AI in their development processes. Most software development methodologies follow a development life cycle that includes the steps of analysis, design, implementation, evaluation, and maintenance. However, a responsible approach to the design of AI systems requires the evaluation process to be continuous during the whole development process and not just a step in the development sequence. Moreover, the dynamic and adaptable nature of AI systems also requires evaluation to be continuous because the system is continuously evolving.

The responsible development life cycle for AI systems must therefore ensure that the whole process is centred around evaluation and justification processes, as depicted in Figure 4.6.

4.5 Concluding Remarks

In this chapter, we presented the ART principles for Responsible AI: Accountability, Responsibility and Transparency, and described a potential design methodology to support development of AI systems that follow these principles. Achieving ARTful systems is a complex process, which requires at least the following steps:

- Align system goals with human values. This requires that core values, as well as the processes used for value elicitation, must be made explicit and that all stakeholders are involved in this process. Furthermore, the methods used for the elicitation processes and the decisions of who is involved in the value identification process are clearly identified.
- Use explicit interpretation mechanisms. Values are per definition and per necessity of an abstract nature and therefore open to be understood in different ways by different actors and under different conditions.
- Specify reasoning methods that handle ethical deliberation, describing both the decisions or actions made by the system, and those that would have been considered of an ethical nature when performed by a person, and indicate the priorities given to which values in the context of the application.
- Specify governance mechanisms to ensure that responsibility can be properly apportioned by the relevant stakeholders, together with the processes that support redressing, mitigation and evaluation of potential harm, and means to monitor and intervene in the system's operation.
- Ensure openness. All design decisions and options must be explicitly reported, linking system functionalities to the social norms and values that motivate them in ways that provide inspection capabilities for code and data sources, and ensure that data provenance is open and fair.

Finally, Responsible AI requires informed participation of all stakeholders, which means that education plays an important role, both to ensure that knowledge of the potential impact of AI is widespread, as well as to make people aware that they can participate in shaping its societal development. We will discuss these issues further in Chapter 6.

4.6 Further Reading

The special issue on "Ethics and Artificial Intelligence" [41], published in 2018 in the Springer journal *Ethics and Information Technology* contains several papers closed related to the topics discussed in this chapter:

- Rahwan, I. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* **20**, 1 (Mar 2018), 5–14
- Bryson, J. J. Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology* **20**, 1 (Mar 2018), 15–26
- Vamplew, P., Dazeley, R., Foale, C., Firmin, S., and Mummery, J. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology* **20**, 1 (Mar 2018), 27–40
- Bonnemains, V., Saurel, C., and Tessier, C. Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology* **20**, 1 (Mar 2018), 41–58
- Arnold, T., and Scheutz, M. The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology* **20**, 1 (Mar 2018), 59–69

On the issue of bias and the impact of automated decision-making, see O'Neill, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown, 2016.

For further information on Design for Values methodologies, see:

- Friedman, B., Kahn, P. H., and Borning, A. Value sensitive design and information systems. *Advances in Management Information Systems* **6** (2006), 348–372
- van den Hoven, J. ICT and value sensitive design. In *The Information Society: Innovation, Legitimacy, Ethics and Democracy. In honor of Professor Jacques Berleur S.J.*, P. Goujon, S. Lavelle, P. Duquenoy, K. Kimppa, and V. Laurent, Eds., vol. 233 of *IFIP International Federation for Information Processing.* Springer, 2007, pp. 67–72
- van de Poel, I. Translating values into design requirements. In *Philosophy and Engineering: Reflections on Practice, Principles and Process*, D. Michelfelder, N. McCarthy, and D. Goldberg, Eds. Springer Netherlands, 2013, pp. 253–266
- Aldewereld, H., Dignum, V., and Tan, Y. H. Design for values in software development. In *Handbook of Ethics, Values, and Technological*

Design: Sources, Theory, Values and Application Domains, J. van den Hoven, P. E. Vermaas, and I. van de Poel, Eds. Springer Netherlands, 2014, pp. 831–845