



# HumanE AI:

## Toward AI Systems that Augment and Empower Humans by Understanding Us, our Society and the World Around Us

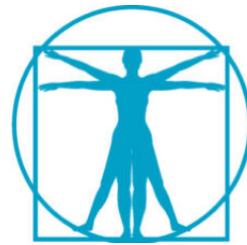
Grant Agreement Number: 761758  
Project Acronym: HumanE AI

Project Dates: 2019-03-01 to 2020-02-29  
Project Duration: 12 months

### ***D1.3 Humane AI Ethical Framework***

**Author(s):** Virginia Dignum  
**Contributing partners:** UMU  
**Date:** November 12 2019  
**Approved by:** Paul Lukowicz  
**Type:** Deliverable  
**Status:** First report  
**Contact:** virginia@cs.umu.se

Dissemination Level	
PU	Public
PP	Restricted to other programme participants (including the Commission Services)
RE	Restricted to a group specified by the consortium (including the Commission Services)
CO	Confidential, only for members of the consortium (including the Commission Services)



## DISCLAIMER

This living document contains material, which is the copyright of *HumanE AI* Consortium parties, and no copying or distributing, in any form or by any means, is allowed without the prior written agreement of the owner of the property rights. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the *HumanE AI* Consortium as a whole, nor a certain party of the *HumanE AI* Consortium warrant that the information contained in this document is suitable for use, nor that the use of the information is free from risk, and accepts no liability for loss or damage suffered by any person using this information.

This document reflects only the authors' view. The European Community is not liable for any use that may be made of the information contained herein.

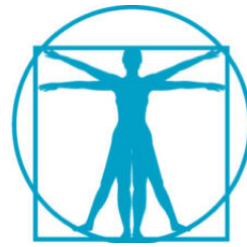
## DOCUMENT INFO

### 0.1 Authors

Authors	Institution	e-mail
Virginia Dignum	UMU	<a href="mailto:virginia@cs.umu.se">virginia@cs.umu.se</a>

### 0.2 Document History

Revision		
Date	Lead Author(s)	Comments
191112	Virginia Dignum	Initial draft



## TABLE OF CONTENTS

<b>Executive Summary</b> .....	<b>4</b>
<b>1 Introduction</b> .....	<b>4</b>
1.1 Accountability .....	8
1.2 Responsibility .....	9
1.3 Transparency .....	12
<b>2 Design for Values</b> .....	<b>15</b>
<b>3 Towards Responsible AI Development Life-Cycle</b> .....	<b>18</b>
<b>4 Legal Aspects of responsible AI</b> .....	<b>19</b>
<b>5 Concluding Remarks</b> .....	<b>21</b>
<b>6 References</b> .....	<b>21</b>



## EXECUTIVE SUMMARY

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. AI, both embedded in software systems and embodied in artefacts (e.g. robots), is everywhere. It affects everyone, and has the capability to transform public and private organisations, and the services and products they offer. 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 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 economic implications are taken into account.

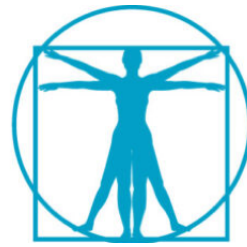
In this report, present the grounding principles of Responsible AI, namely Adaptability, Responsibility and Transparency. We then introduce the Design for Values methodology to guide the development of Responsible AI systems. In section 4, we discuss how these principles can be integrated into a system development life cycle framework. Finally, section 5 focuses on the legal issues, in particular legal protection by design (LPbD).

## 1 INTRODUCTION

AI will affect everybody. This demands that the development of AI systems ensures inclusion and diversity, that is, truly considers all humankind 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. The core of AI development should be the enhancement of human wellbeing and capacities, including the concepts of 'AI for Good' and 'AI for All'.

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.

Governance is necessary for the reduction of incidents, to ensure trust, and for society's long-term stability through the use of well-established tools and design practices. Well designed, regulations do not eliminate innovation, as some claim, instead can enhance innovation through the development and promotion of both socio-legal and technical means to enforce compliance (Brundage & Bryson, 2016).



Moreover, policy is needed to ensure the human responsibility for the development and deployment of intelligent systems, filling the gap that emerges from the increased automation of decision. The ultimate aim of regulation is to ensure our—and our societies'—well-being in a sustainable world. That is, research, development and use of AI should always be done in a responsible way, what is often referred to as responsible AI. When developing intelligent systems, besides the obvious necessity to meet legal obligations, societal values and moral considerations must also be taken into account weighing in the respective priorities of values held by different stakeholders in various multicultural contexts. Human responsibility to ensure flourishing and well-being of our societies should always be at the core of any technological development (Dignum, 2017).

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 AI is about the processes by which AI is developed (ethics in AI design and development), accountability for the results of AI system deliberation (ethics by design) and making sure that those developing AI systems are aware of their role and impact on the values and capabilities of those systems (ethics for designers) (Dignum, 2019), Design methods, verification techniques, and codes of conduct are all aspects that need to be developed alongside the computational design of algorithms.

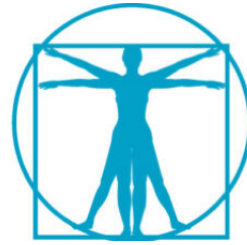
Responsible Artificial Intelligence is concerned with the fact that decisions and actions taken by intelligent autonomous systems have consequences that can be seen as being of an ethical nature. These consequences are real and important, independently of whether the AI system itself is able to reason 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.

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

# HUMANE AI



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.

Moreover, governance is not only about ethics despite the focus of most of these works. AI systems need to be legal and reliable as well. Abidance to law is the basis to any application, whereas ethics is the ‘sky’, the horizon we aim to attain (Bryson, 2018, Floridi et al., 2018, Winfield & Jirotko, 2018). However, as Hildebrandt (2020) points out: “whereas law and the Rule of Law introduce checks and balances and demand democratic participation (at least in constitutional democracies), ethics may be decided by tech developers or behind the closed doors of the board room of corporate business enterprise. It can thus obtain the force of technology.” That is, law is transparent and explicit and can be audited by external parties. If the law is flawed or does not represent its society’s best interests (or moral values at the time), it can be amended through the democratically elected legislative body that set the law. However, how and an AI system handles ethical dilemmas is the result of possibly implicit implementation decisions decided behind closed doors with little-to-none input by its users, let alone the society at large.

In this report 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 (Floridi & Sanders, 2004; Russell & Norvig, 2009)

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

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 should be complemented with design principles that ensure trust. Here, we follow the proposal of (Dignum, 2017) to complement autonomy with *responsibility*, interaction 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. 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 1. 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.

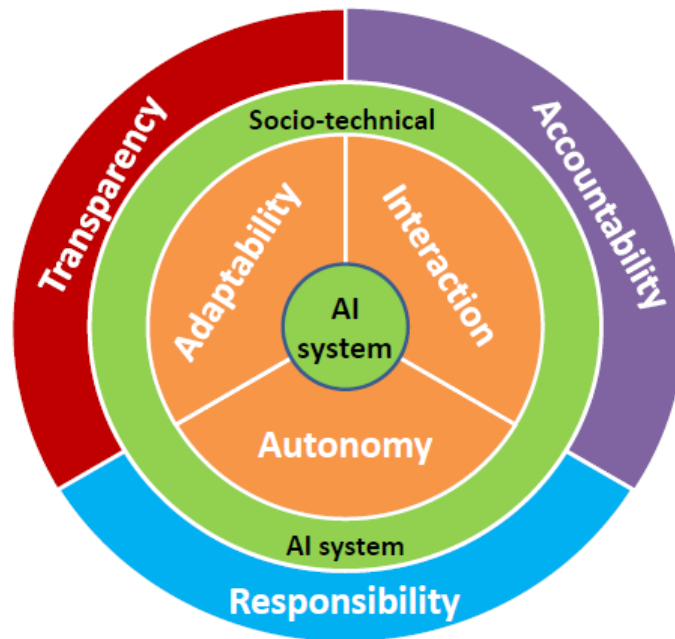


Figure 1: The ART principles refer to AI as a socio-technical system.

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), and 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 (Eisenhardt, 1989).

## 1.1 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<sup>1</sup> (Winikoff, 20170, Gunning, 2018). Another important aspect of accountability is to be able to rely on a safe and sound design process that accounts for and reports on options, choices and restrictions about the system's aims and assumptions (Friedman et al., 2006). 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 (Malle et al., 2015). 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.

---

<sup>1</sup> See also GDPR regulation: <http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf>



# HUMANE AI



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 (Dignum et al., 2018), 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 (Miller, 2019). According to Miller, explanations should be *contrastive*, i.e. answer questions of the form “why did you do \$\$\$ ... instead of \$\$\$?”; *selective*, i.e. select relevant factors and present those; and *social*, i.e. presented relative to what the explainer believes the listener 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 (Grice, 1975).

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 (Friedman et al., 2006; van den Hoven, 2007) 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 3.

## 1.2 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

# HUMANE AI



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 the apportionment of liability for the machine's decisions.

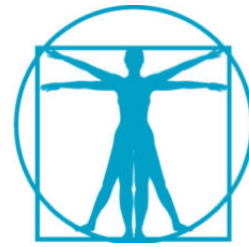
Autonomy of an AI system refers to its autonomy to develop its own plans and to decide between its possible actions. Therefore these actions can in principle be traced back to some user instruction (e.g. personalisation preferences), manufacturing setting or design choice. Even if the system has evolved, by learning from its interaction with the environment, what it learns is determined by the purpose for which it was built, 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 built 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 (Bryson, 2018); 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 (Shaw et al., 2018, Conitzer et al., 2017; Dennis et al., 2015).

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 latter 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”<sup>2</sup>. 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, 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*”<sup>3</sup>. 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

---

<sup>2</sup> See Motion for a European Parliament Resolution, with recommendations to the Commission on Civil Law Rules on Robotics. [http://www.europarl.europa.eu/doceo/document/A-8-2017-0005\\_EN.html?redirect](http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html?redirect), 2017

<sup>3</sup> See <https://g8fip1kplyr33r3krz5b97d1-wpengine.netdna-ssl.com/wp-content/uploads/2018/04/RoboticsOpenLetter.pdf>



from Hanson Robotics<sup>4</sup> 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, 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.

### 1.3 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 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,

---

<sup>4</sup> See <https://www.hansonrobotics.com/>



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<sup>5</sup>. 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.

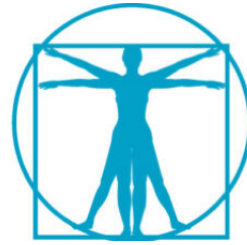
Transparency may be better served by openness and control over the whole learning and training process than by removing the algorithmic black box. Trust in the system will improve if we can ensure openness of affairs in all that is related to the system. This can be done by applying software- and requirement-engineering 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 2 describes possible questions to be considered to ensure transparency of design processes.

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<sup>6</sup> are collected and documented. 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.

---

<sup>5</sup> For more on this issue, see e.g. (O’neill, 2016)

<sup>6</sup> Requirements elicitation refers to both functional and non-functional requirements, which include the values that the system should enforce and/or follow.



### 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 2: Checklist for Transparency.

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.



## 2 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 (van den Hoven, 2005). 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 (Aldewereld et al., 2014).

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 3 depicts the high-level Design for Values approach for AI systems. 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 resource*, 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.

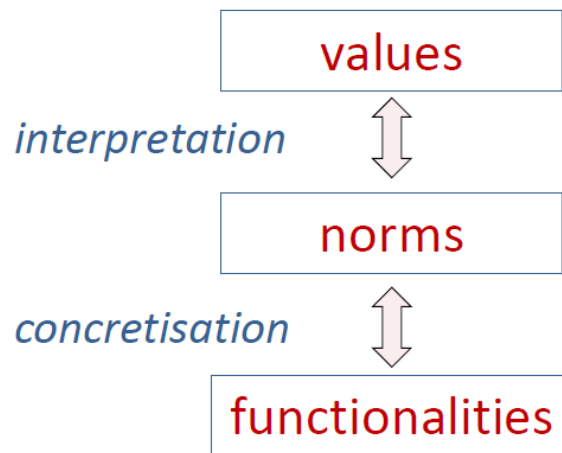


Figure 3: Design for Values: From values to norms to functions, and back

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<sup>7</sup> (Pedreschi et al., 2008) or equal odds<sup>8</sup> (Dwork et al., 2012) amongst others. These **functionalities** are quite different in their results. Without being explicit about which approach is taken to implement 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 (Aldewereld et al., 2010). 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 (vandePoel, 2013). This relation can also be turned around to indicate that the norm of equal opportunities *counts-as* fairness in a given context (Searle, 1995; Jones & Sergot, 1993).

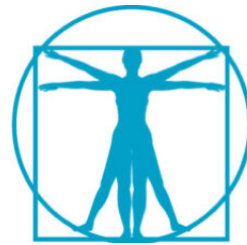
Precise interpretations, using formal verification mechanisms, are needed both to link values to norms, as well as to transform these norms into concrete system functionalities. Work on formal normative systems proposes a representation of such interpretations based on the formal concept of *counts-as*, where the relation *X counts-as Y* is interpreted as a subsumption that holds only in relation to a specific context (Grossi et al., 2006, Aldewereld et al., 2010).

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

<sup>7</sup> 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.

<sup>8</sup> Equal odds aims at balancing classification errors across protected attributes, towards achieving equal false positive rates, equal false negative rates, or both.





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' (Vazquez-Salceda et al, 2008).

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.

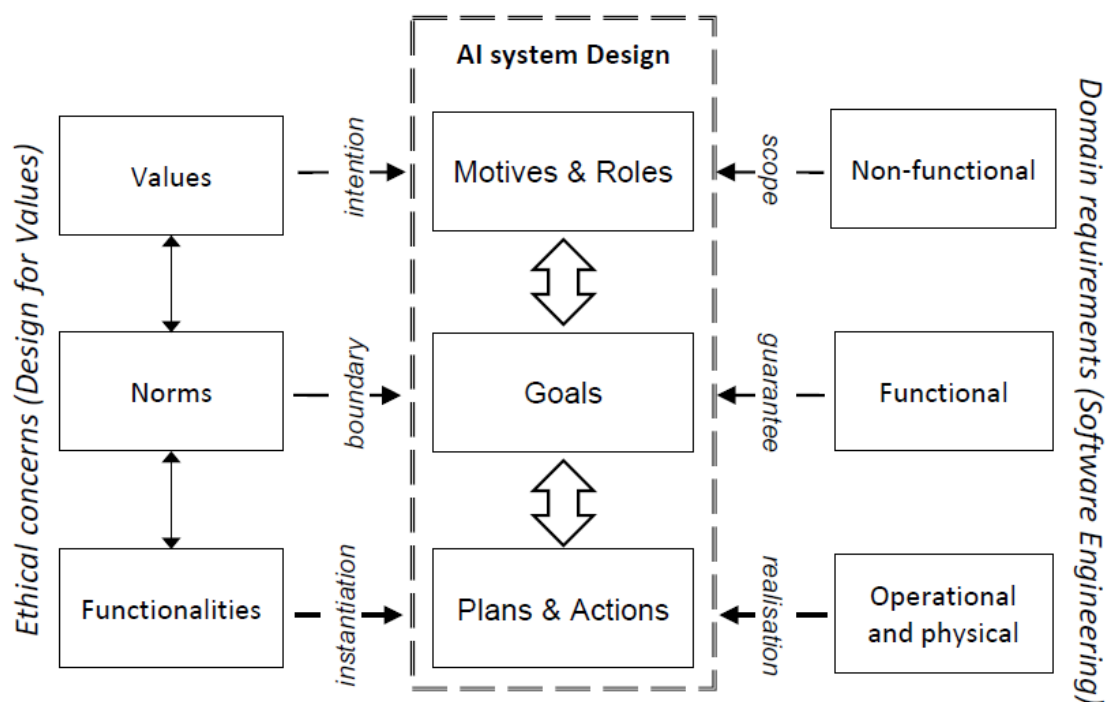


Figure 4: Responsible AI design: integrating ethical concerns and domain considerations.



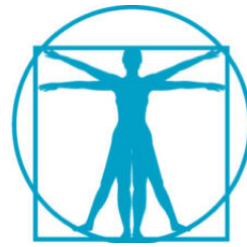
These issues point to the need for a multi-layered approach to software development where the links to the values are kept explicit. In the following, we present a possible design methodology for Responsible AI, based on the Value Sensitive Software Development (VSSD) framework proposed by (Aldewereld et al., 2014). 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, 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.

### 3 TOWARDS RESPONSIBLE AI DEVELOPMENT LIFE-CYCLE

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.

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 on evaluation and justification processes, as depicted in Figure 5.

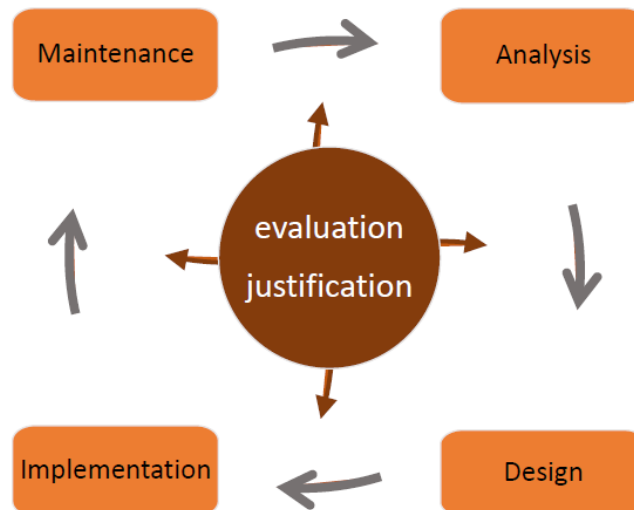


Figure 5: The Responsible Development Life Cycle for AI systems.

## 4 LEGAL ASPECTS OF RESPONSIBLE AI

Every AI system should operate within an ethical and social framework in understandable, verifiable and justifiable ways. Such systems must in any case operate within the bounds of the rule of law, incorporating fundamental rights protection into the AI infrastructure. That is, given AI systems are artefacts built for a given purpose, it is necessary to demand that these artefacts to stay within the realm of what is both legal and ethical, and do not learn other options by themselves. That is, AI systems should be seen as incorporating soft ethics, i.e, ethics as post-compliant to an existing regulatory system, and used to decide on what ought and ought not to be done over and above the existing regulation (Florida, 2018).

Theory and methods are thus needed for the Responsible Design of AI Systems as well as to evaluate and measure the ‘maturity’ of systems in terms of compliance with legal, ethical and societal principles. This is not merely a matter of articulating legal and ethical requirements, but involves robustness, and social and interactivity design. Concerning ethical and legal design of AI systems, we must clarify the difference between legal and ethical concerns, as well as their interaction (Hildebrandt, 2020), and ethical and legal scholars will work side by side to develop both legal protection by design and value-sensitive design approaches. The focus here is the prioritization of ethical, legal, and policy considerations in the development and management of AI systems to ensure responsible design, production and use of trustworthy AI.

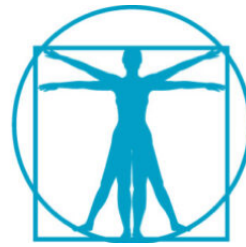


Legal aspects in responsible AI entail a focus on the preconditions for ethical conduct, for instance but not only: (1) accountability of those who take risks with other people's rights, freedoms and interests (by processing their data or targeting them based on data-driven inferences), (2) effective and meaningful transparency concerning the logic of automated decision systems that enables safe and meaningful interaction with such systems (both in the case of online search, social media and commerce and in the case of real-world navigation as in Internet of Things and robotics), (3) actionable purpose limitation to enable users (inhabitants) of AI environments to foresee the behavior of the myriad systems they may encounter (from connected cars to self-executing insurance contracts based on real-time data-driven input and care robots for the elderly), (4) reliable proportionality testing in the context of impact assessments, balancing the interests of providers against the rights, freedoms and interests of users or third parties that will suffer the consequences (whether algorithmic or data protection or safety and security impact assessment), in a way that enables them to contest such assessments, (5) built-in human-machine interaction that allows users or those targeted to exercise their fundamental rights and freedoms (enabling access to meaningful information, withdrawal from invasive targeting, detecting and contesting prohibited or unfair discrimination, and violations of the presumption of innocence).

Such Legal Protection by Design (LPbD) entails the incorporation of fundamental rights protection into the architecture of AI systems. This plays out at two levels:

1. This first concerns the checks and balances of the Rule of Law, notably a concrete and effective set of interventions at the level of the research design, the subsequent development of core code, choice of programming language, foreseeable system behaviors, design of the APIs, and various types of interfaces. This concerns the choice architecture instituted by law that confronts (1) developers, (2) manufacturers, (3) sellers, (4) users (e.g. service providers, governments), and (5) end-users, providing them with leeway, proper constraints, transparency, accountability and foreseeability.
2. The second level concerns requirements imposed by positive law that elaborates fundamental rights protection, such as the GDPR, non-discrimination legislation, labour law and the more. Here, the point is to follow up on concrete legal norms (e.g. the right to withdraw consent as easily as it has been given), translating them into technical requirements and specifications when developing applications. LPbD differs from ethics by design because it concerns legal obligations that are democratically legitimated and enforceable under the Rule of Law.

The choice of which norms must be built-in therefore does not depend on the ethical inclinations of e.g. developers or service providers, but on constitutional preconditions for ethical behavior (e.g. ensuring that those who act ethically will not be pushed out of the market), and on enforceable law. LPbD differs from mere techno-regulation, 'legal by design', or 'compliance by design' because it does not aim to nudge or technologically enforce e.g. administrative law, trying to turn legal obligations into technical measures, but instead aims to build transparency, accountability and contestability into these systems enabling a.o. protection against "compliance by design."



## 5 CONCLUDING REMARKS

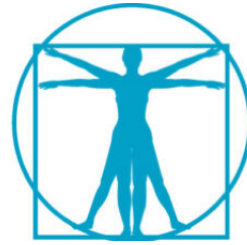
In this report, we presented the ART principles for Responsible AI: Accountability, Responsibility and Transparency, and described a potential design methodology to support development of AI systems that follow these principles. Achieving *ARTful* systems is a complex process, which requires at least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

## 6 REFERENCES

- Aldewereld, H., Álvarez-Napagao, S., Dignum, F., and VázquezSalceda, J. Making norms concrete. In 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010) (2010), International Foundation for Autonomous Agents and Multiagent Systems, pp. 807–814.
- 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.
- Brundage, M, Bryson, JJ. 2016. Smart Policies for Artificial Intelligence. (August 2016). Retrieved from <http://arxiv.org/abs/1608.08196>
- Bryson, J. J. Patency is not a virtue: the design of intelligent systems and systems of ethics. Ethics and Information Technology 20, 1 (Mar 2018), 15–26.

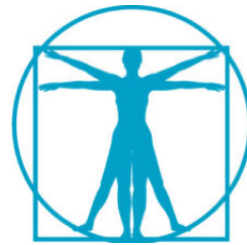


- Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., and Kramer, M. Moral decision making frameworks for artificial intelligence. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017) (2017), pp. 4831–4835.
- Dennis, L. A., Fisher, M., and Winfield, A. Towards verifiably ethical robot behaviour. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015).
- Dignum, V. Responsible Artificial Intelligence, Springer (2019).
- Dignum, V. Responsible autonomy. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017) (2017), pp. 4698–4704.
- Dignum, V. Ethics in artificial intelligence: introduction to the special issue. Ethics and Information Technology 20, 1 (Mar 2018), 1–3.
- Dignum, V., Baldoni, M., Baroglio, C., Caon, M., Chatila, R., Dennis, L., Genova, G., Kliess, M., Lopez-Sanchez, M., Micalizio, R., Pavon, J., Slavkovik, M., Smakman, M., van Steenbergen, M., Tedeschi, S., van der Torre, L., Villata, S., de Wildt, T., and Haim, G. Ethics by design: necessity or curse? In Proceedings of the 1<sup>st</sup> International Conference on AI Ethics and Society (2018), ACM, pp. 60–66.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. Fairness through awareness. In Innovations in Theoretical Computer Science 2012 (2012), ACM, pp. 214–226.
- Eisenhardt, K. M. Agency theory: An assessment and review. The Academy of Management Review 14, 1 (1989), 57–74.
- Floridi, L. Soft ethics and the governance of the digital. Philosophy & Technology 31, 1 (Mar 2018), 1-8.
- Floridi, L., and Sanders, J. On the morality of artificial agents. Minds and Machines 14, 3 (2004), 349–379.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. and Schafer, B., 2018. AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), pp.689-707.
- Friedman, B., Kahn, P. H., and Borning, A. Value sensitive design and information systems. Advances in Management Information Systems 6 (2006), 348–372.
- Grice, H. P. Logic and conversation. Academic Press, 1975.
- Grossi, D., Meyer, J.-J. Ch., and Dignum, F. Counts-As: Classification or constitution? An answer using modal logic. In Deontic Logic and Artificial Normative Systems: Proceedings of the Eighth International Workshop on Deontic Logic in Computer Science (DEON'06) (2006), L. Goble and J.-J. Ch. Meyer, Eds., vol. 4048 of Lecture Notes in Artificial Intelligence, Springer-Verlag.

# HUMANE AI

- Gunning, D. Explainable Artificial Intelligence (XAI). <https://www.darpa.mil/program/explainable-artificial-intelligence>, 2018.
- Hildebrandt, M. Law for Computer Scientists and Other Folk, Oxford University Press, 2020
- Jones, A., and Sergot, M. On the characterization of law and computer systems. In Deontic Logic in Computer Science: Normative System Specification (1993), J.-J. Meyer and R. Wieringa, Eds., Wiley, pp. 275–307.
- Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (2015), HRI '15, ACM, pp. 117–124.
- Miller, T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267 (2019), 1 – 38.
- O'Neill, C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown, 2016.
- Pedreschi, D., Ruggieri, S., and Turini, F. Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2008), pp. 560–568.
- Russell, S., and Norvig, P. Artificial Intelligence: A Modern Approach, 3rd ed. Pearson Education, 2009.
- Searle, J. The Construction of Social Reality. Simon and Schuster, 1995.
- Shaw, N. P., Stöckel, A., Orr, R. W., Lidbetter, T. F., and Cohen, R. Towards provably moral AI agents in bottom-up learning frameworks. In 2018 AAAI Spring Symposium Series (2018), pp. 69–75.
- 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.
- van den Hoven, J. Design for values and values for design. Information Age, Journal of the Australian Computer Society 7, 2 (2005), 4–7.
- 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.
- Vázquez-Salceda, J., Aldewereld, H., Grossi, D., and Dignum, F. From human regulations to regulated software agents' behaviour. Journal of Artificial Intelligence and Law 16 (2008), 73–87.
- Winfield, A., Jirotko, M.. 2018. Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos. Trans. R. Soc. A Math. Phys.

# HUMANE



# AI

Eng. Sci. 376, 2133 (November 2018), 20180085.  
DOI:<https://doi.org/10.1098/rsta.2018.0085>

Winikoff, M. Towards Trusting Autonomous Systems. In Fifth Workshop on Engineering Multi-Agent Systems (EMAS) (2017).