Responsible Innovation: Ethics, Safety and Technology 2nd edition

Joost Groot Kormelink (Editor)

TU Delft open

Responsible Innovation:

Ethics and risks of new technologies

How to deal with risks and ethical questions raised by the development of new technologies.

This book is based on the Massive Open Online Course Responsible Innovation which was first offered by the TU Delft in November 2014-January 2015 on the edX-platform. This book contains all the content covered by the web lectures and some additional content.

This is the link to the re-run in 2018/19: https://www.edx.org/course/responsible-innovation-ethics-safety-delftx-ri101x**.**

A large number of teachers provided input for the MOOC on RI. In Annex 3 you will find an overview of the teachers (with a link to further information) including a link to the weblectures on YouTube and reference to the related paragraph in this book.

Editors

First edition
- Naveen Srivatsa (course moderator)
- Sofia Kaliarnta (course moderator)
- Joost Groot Kormelink (course manager)

Second edition
- Joost Groot Kormelink, TU Delft, Faculty of Technology, Policy and Management

Delft, September 2019

Introduction

0.1 When is innovation good for society?

Innovation may bring a lot of good to society, but innovation is not a good in itself. History provides many examples of innovations and new technologies that had serious negative consequences, or simply failed to address significant problems and make meaningful contributions to society. Well known examples are carcinogenic asbestos or the ecological devastation caused by DDT.

Figure 0-1: Asbestos

Figure 0-2: New technologies

New technologies come with many ethical questions, controversies: and unknown risk. Think about nanotechnology, biotechnology, artificial intelligence, autonomous weapons, genomics, big data analytics, and so on.

At the same time, we do need new technologies to find solutions for great societal challenges, such as the scarcity of energy sources, ageing demographics, water management and food security.

It is therefore of the utmost importance - our duty even- , to define an adequate and shared conception of responsibility for our innovations and technologies. Just think about questions like:

- Can our innovations save lives?
- Will they produce more jobs?
- Can they save the planet, or do they only contribute more waste and pollution?
- Are they safe for users and secure from abusers?
- Do they respect the values and basic human rights we hold dear, like privacy, freedom, autonomy and equality? If not, how can we make them so? If not us, who? If not now, when?

The term "Responsible Innovation" was first introduced in 2006 in the context of the Dutch Research Council (NWO) Program entitled Socially Responsible Innovations. It has now been incorporated into the larger Research and Development agenda of the European Union (EU).

In November 2014, the policy was endorsed and extended in the **Rome Declaration on Responsible Research and Innovation**.

In Annex 1 you will find an overview of some main EU projects.

Our goal is to provide in-depth knowledge of what responsible innovation entails: an ethical perspective to help shape socio-technical solutions and innovations for global and regional problems. However, this reader is a comprehensive but by no means exhaustive primer to responsible innovation.

0.2 Content of this book and learning objectives

In this reader, we will start with a general introduction to RI (Part 1). How can we define RI? After that we will look at RI from different angles:

- Applied ethics and societal values as a starting point for innovation (Part 2)
- Safety and risk (Part3)
- Different types of innovations and processes including frugal innovations (Part 4)
- Design for Values (Part 5).

This reader will also highlight examples and case studies throughout the different chapters.

The learning objectives for the MOOC RI and this book are:

- Understand the concept of responsible innovation and its key ethical dimensions
- Become familiar with various ways and instruments to analyse the risks of new technologies, both forward-looking as backward-looking (e.g. the causes of accidents)
- Learn how to deal with known and unknown risks (deep uncertainty) when it comes to new technologies
- Become familiar with various types of innovation (e.g. radical, niche, incremental, frugal) and the conditions for success
- Apply the concept of Value Sensitive Design (VSD)
- Learn to critically reflect on new technologies from an ethical and risk perspective
- Be able to demonstrate how to think about - and translate - our moral values (e.g. privacy, safety, sustainability, inclusiveness) as technical requirements for new technologies.

0.3 Acknowledgements

This book could have not have been accomplished without the support of Professor Jeroen van den Hoven, the course director Saskia Roselaar for her thorough and insightful proofreading and the TU Delft Library for giving me the opportunity to publish this book as an open textbook and managing the production process.

Part I:
General Introduction to RI

"Making new technologies work for society….. without causing more problems than they solve"

(Hilary Sutcliffe, Director, SocietyInside)

Figure 1.1 : Lead in Petrol was only phased out in the seventies although it was known much longer that it causes neurological diseases.

1. Introduction to responsible innovation

1.1 The real-world context of responsible innovation: dilemmas

Before getting into the definition of **responsible innovation** (abbreviated to **RI** in this book), we will put the discussion into context.

Try to reflect on the following four dilemmas and underlying questions which have been designed to get you thinking about RI and get a feeling for the issues which will be discussed in this book.

- *Dilemma 1: Dealing with hazards.*
  New technologies can bring dangers and it is very well possible that we are unable to control or contain the outcomes. We expect a certain level of risk with every innovation. Some risk is unavoidable, but how much harm to human health, the environment and society is acceptable? Furthermore, it is essential to consider whether the danger is controllable. For instance, if we find out something is hazardous, would we be able to restrict its effects by removing the specific technology from society, stopping its effects, or even reverse the effects? And should we restrict these effects, even if it limits the usefulness of the innovation?

  To what extent do you think hazards should be controllable? Should they be fully controllable or do you think that allowing for some risk or hazard is part and parcel of life, and comes with each innovation?

- *Dilemma 2: Knowledge of outcomes*
  There is a certain level of knowledge required to make a comprehensive and reliable assessment of new technology. How can we get that knowledge? What level of certainty do we have that hazards may or may not occur?

  The level of knowledge can range from no knowledge (ignorance) to uncertainty about the likelihood, to knowing the probability of failure or having certain knowledge. If we are not certain of the outcomes, who is responsible for finding out, monitoring and taking precautions against hazards?

  When assessing a new technology, how much knowledge about the hazards and risks is enough, before deciding to introduce the technology in society? Should we assume that important risks and hazards will occur every now and then, and that it is not possible to anticipate and assess them beforehand? Or should we be certain of all possible hazards and risks beforehand, and thus have the capability to prevent or contain negative outcomes to some extent?

And what about the use of potentially hazardous technologies? Should we monitor every aspect of such technologies? Or is constant monitoring not necessary, since critical issues will become apparent anyway, so we only need to find a way to report and respond to any issues?

- *Dilemma 3: Distribution of risks and benefits*
  How should risks and benefits be distributed? Should they be distributed equally across all social groups and generations? Or, as it is often the case in real life, is it impossible to distribute benefits and dangers equally? What constitutes fair distribution?

  Essentially, this line of questioning explores the expected social benefits and hazards of a technology, and how these are distributed among stakeholders, including the environment and future generations.

Figure 1-2 Protest in Congo against children having to work in Cobalt mines under terrible circumstances

- *Dilemma 4: Feedback and democratic influence*
  Should ordinary citizens have some level of influence on the design and availability of new technologies; or not? To what extent can societal actors, NGOs, citizens and other public groups influence technological development? Should they have the power to block the development of potentially harmful technologies, if need be? Or do only producers and experts have enough knowledge and capability to make critical decisions?

Figure 1-3: Protest in London against Killer Robots

1.2 Why discuss responsible innovation?

Innovation often brings wonderful and unimagined new functional abilities that are in high demand and may lead to new business, new jobs and thus economic prosperity. And innovation does not only bring monetary profits: it also brought us penicillin, clean water and sanitation. As a result of these kinds of innovations, our life expectancy has gone up dramatically and hundreds of millions of people have been lifted from poverty and disease. Clearly, many types of innovation are desirable.

But surely innovation is not a good in itself. If we agree that something is really innovative and brings interesting new functionalities, it still makes perfect sense to ask: "but is it good?" There are plenty of examples of innovations which initially seemed a blessing, but later gave rise to serious moral concerns, like pesticides with DDT and building materials with asbestos. These innovations were once sold as wonderful new technological inventions, but are now associated with a greatly increased risk of illness and even death.

The **UN Sustainable Development goals** and the **EU's Grand Challenges** provide a list of urgent moral goals for innovation and applied science on a global scale; the EU has allocated a large part of its budget to fast-track work along these lines.

Figure 1-4: the UN 17 sustainable development goals

So, innovation in our time is no longer about building bigger SUVs, but instead about saving the planet and handing it down to future generations in good shape. We worry - as we should - about climate change, renewable energy, autonomous vehicles, big data and privacy, nuclear power and proliferation of nuclear weapons. We know by now that many of our innovations have a vast impact: they affect people in remote corners of the earth, the planet as a whole and generations in distant futures.

Our innovations have even started to alter what it means to be human: cochlear implants give the deaf back their hearing, advanced prosthetic devices and artificial organs bring functionality to the ill and disabled, cognitive neuro-enhancement may make some of us smarter.

Whether these are acceptable innovations will depend on their precise features and on how we shape this technology. This means we have to take responsibility for our innovations and realize that technology is never neutral, but always value-laden.

Many scholars in the past have realized that technology inherits the values of its maker. A couple of low-tech examples may serve to illustrate this point: the entrance to Bethlehem's Church of the Nativity is referred to as the "Door of Humility", because visitors must bend down to enter. Over the centuries, the entrance has been made smaller in order to keep thieves from entering the basilica on horseback; the sturdy but low door has nothing to do with humility, but is actually a security feature.

Langdon Winner, in his famous essay **"Do artefacts have politics?"**, argued that the low-hanging overpasses in New York in the beginning of the 20th century were low by design (see image), so as to prevent busses going from poor black neighbourhoods to the white middle-class beaches.

Figure 1-5: Low-hanging overpasses

Subsequently, this basic idea of values expressed and embodied in technology and design was elaborated in the field of Science and Technology Studies. Recently, studies in software engineering have drawn attention to the fact that information and communications technology is an important new carrier of values. It has been demonstrated how search engines, financial software and geographical information systems (GIS) may contain controversial algorithms and models that shape our behaviour and our thinking when we work with them.

If we do not critically and systematically assess our technologies in terms of the values they support and embody, people with perhaps less noble intentions may insert their views on sustainability, safety and security, health and well-being, privacy and accountability. In our case studies we will show you some examples.

Therefore, not only will our innovations have to be geared towards solving the world's great challenges, they will themselves have to be expressions of our shared moral values. Technology is too central, and the science underlying it too fundamental, to be ignored. We should not wait for outcomes and only reflect after the fact. This is why we need to think and act to promote responsible innovation, either by making the values embedded in our existing technologies explicit and clear, or by finding ways to develop the values we desire into practical, deployable design parameters.

1.3 Defining RI

Given the fact that we pursue many different values at the same time, we find it hard - and sometimes impossible - to choose between them or to compromise. We highly value privacy, health, sustainability, efficiency, equity, security, accountability and so much more, and all of them at the same time. .

Figure 1-6: Safety of privacy: What is more important?

But we cannot meet them all simultaneously, there is a trade-off*.*
In other words: We often find we have more moral obligations than the situation allows us to satisfy, and this can lead to situations of moral overload. We will discuss this in greater detail later in the book (Chapter 2).

Usually, this is seen as a problem. However, it may actually trigger creativity and the commitment to try and accommodate conflicting values by smart design and innovation.

Some examples are:

- Fairphone is a start-up that makes smartphones from conflict-free metals, so that human rights, sustainability, fairness and security are accommodated in one design.
- In the Netherlands, large storm surge barriers have been built to protect the country against flooding, but they are also ways to manage the ecosystem and generate tidal energy at the same time.
- Privacy-enhancing technology gives us access to the wonderful benefits of computers without the privacy drawbacks.
- Clean tech gives us the opportunity of industrial production and economic prosperity without environmental damage.
- The zero-tolerance policy against fatal road accidents in Sweden has triggered a great deal of innovation in the automotive industry. Volvo is now a leader in the production of safe cars.

1.4 Substantive and process aspect of RI

Substantive aspect

Innovation can thus also be construed as a moral concept in the sense that it helps to change the world, so that the set of moral obligations we can satisfy is amplified.

There is no guarantee, of course, that there will always be perfect solutions to our pressing moral problems, and in some cases we may need to apply more drastic and fundamental approaches. However, we do have an obligation to see whether there are possibilities to use innovation to meet conflicting values. ***This, one could say, is the outcome or substantive aspect of RI.***

Process aspect.

However, there is also a *process aspect to RI*.

In order to appreciate how <u>responsibility</u> is assigned in a complex (multi-actor) system, we have to look at the criteria that can determine who can be held responsible (e.g. knowledge, intention, non-coercion, contributory fault and capacity). This list corresponds nicely with excuses people tend to give when they want to deny responsibility: "I didn't know", "I didn't mean it", "I was forced", "It wasn't me", "I didn't understand".

Everything we do, we can do in such a way so as to extend our responsibility - or we may undercut or weaken our own responsibility, in order to make it more difficult for others to hold us responsible or accountable. There are many strategies to remain ignorant or pretend one is ignorant, in order to orchestrate plausible deniability. Think about the risks associated with new materials and chemical substances: "We could not have foreseen this. Our competitors also used asbestos. It was not us, but actually our subcontractors who were at fault. Our company did not have the resources at the time to critically consider this."

1.5 EU-definition of RI

This brings us to the defining clauses of RI, as given in the EU report **Options for Strengthening Research and Innovation**. The report considers that, for an innovative organization or process to be praised as being "responsible", this would imply - among other things - that those who initiated it and were involved in it must be acknowledged as moral and responsible agents. In other words, they have to:

- Obtain - as much as possible - relevant knowledge on (i) the consequences of the outcomes of their actions and (ii) the range of options open to them;
- Evaluate all outcomes and options effectively in terms of relevant moral values (including, but not limited to, well-being, justice, equality, privacy, autonomy, safety, security, sustainability, accountability, democracy and efficiency). In light of the "design for values" concept (see chapter 9 of this book) and the possibility of resolving problems by design, another aspect of RI is the capability of relevant moral agents;
- Use these two considerations as requirements for the design and development of new technology, products and services, leading to moral improvement.

In essence, responsible innovation refers to a transition to a new situation - and an amplification of possibilities - to meet more obligations and honour more duties to fellow human beings, the environment, the planet and future generations than before

Box: The EU and Responsible Research and Innovation (RRI)

The EU always speaks about Responsible *Research* and Innovation (RRI), and defines this concept as follows: an interactive process where societal actors, researchers and innovators actively cooperate to co-define, co-design and co-construct solutions, services and products that are socially acceptable and sustainable and resolve important societal issues.

RRI is a cross-cutting/overall priority in Europe-funded programs that encourages societal actors to work together during the whole research and innovation process in order to better align its results with the values, needs and expectations of society. This means that researchers, scientists and policymakers should interact with each and with other societal actors to create mutual awareness, co-define and co-design new initiatives and identify solutions to societal challenges. RRI focuses on how research and innovation can become more beneficial to society and simultaneously protect the environment.

Research and innovation (R&I) may contribute to finding solutions to society's main challenges, e.g. a circular economy, prevent climate change, mitigate demographic changes, improve well-being, energy security, food safety and secure societies. The European Union recognizes these challenges and considers RRI one of the main approaches to address them. Moreover, RRI can be an excellent vehicle to connect science with policy; this will enable policy-makers to be better informed and equipped to formulate improved policies and to achieve ecologic and economic goals.

Figure 1-7: EU and RRI

# Part II: Applied Ethics for Responsible Innovation
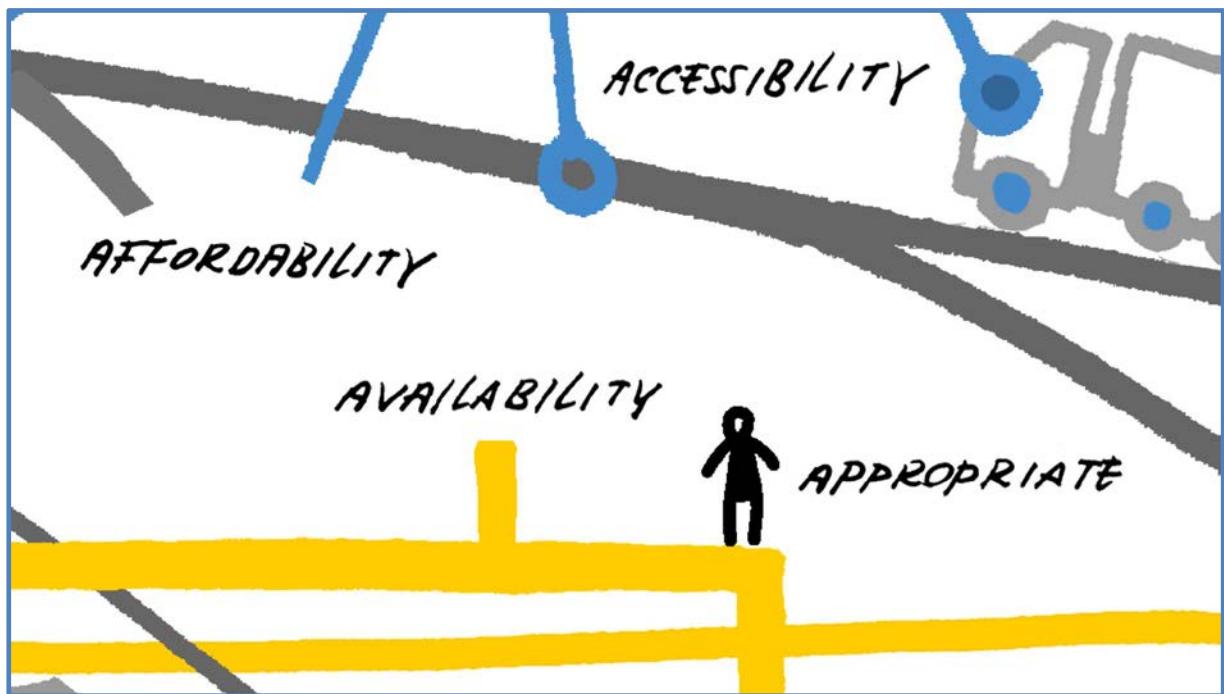
Figure 2.1: Potentially conflicting requirements: for, for example, energy transition.

2. Applied ethics for responsible innovation

2.1 Applied ethic: thought experiments

To freely explore moral and ethical nuances in an abstract manner, philosophers have traditionally come up with thought experiments. Thought experiments typically set up a carefully orchestrated dilemma, asking readers to pick their preferred course of action and justify why their choice would be the lesser evil. In this way, there is an opportunity to explore the philosophical implications of different responses to a dilemma. When we speak of responsible innovation, it becomes important to truly understand what we mean by the word "responsible" - that is to say, who is responsible, how, when and why. The "Trolley Problem" is one such thought experiment that could serve this purpose.

The Trolley Problem

The "Trolley Dilemma" (or the "Trolley Problem") consists of a series of hypothetical scenarios developed by **British philosopher Philippa Foot** in 1967: each scenario presents an extreme environment that tests the subject's ethical prowess. In 1985, American philosopher Judith Jarvis Thomson scrutinized and expanded on Foot's ideas in *The Yale Law Journal*.

The Trolley Problem is a thought experiment in ethics whose general form is as follows: there is a runaway trolley barrelling down the railway tracks (see image). Further ahead on the track, five people are tied up and unable to move. The trolley is headed straight for them! You are standing further away in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, on the side-track one person is tied up.

So you have two options:

- Do nothing, and the trolley kills the five people on the main track.
- Pull the lever, diverting the trolley onto the side-track, where it will kill one person.

What would you do?:
- Flip the switch to maximise the number of lives saved (one person dies, so five can live).
- Flip the switch because you are a compassionate person and it is the right thing to do.
- Do not flip the switch as it would lead to killing, and killing is inherently wrong.
- Do not flip the switch because you feel aiding in a person's death is culturally inappropriate, not to mention illegal.
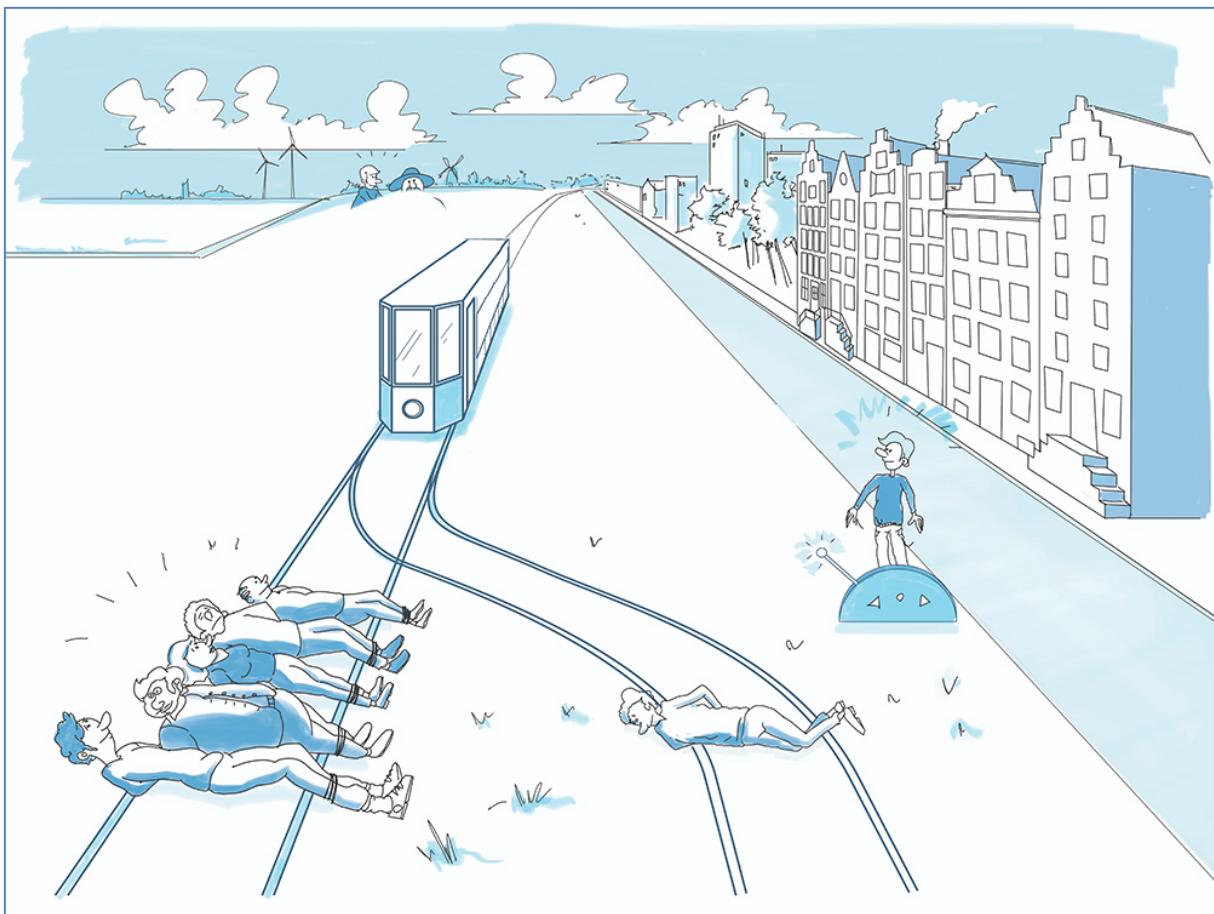
Figure 2.2: The famous Trolley Problem. What would you do? Save 5, although one person loses his life as a result?

Given the Trolley Problem as explained above, what would you do? Is it morally permissible to pull the lever, or do you even have a moral obligation to do so? Almost all philosophers in the last three decades have been raised on such so-called "trolley cases". If you would like to do a PhD in trolley problem analysis, it would be a respectable topic in philosophy departments around the world, assuming you would be able to add something new to the vast literature.

The reason why we discuss this artificial thought experiment is not to introduce you to the very extensive body of literature surrounding it, but rather *to illustrate how thinking about RI requires a point of view on making moral choices and responsibility* that is different from the philosophical ones used to analyse trolley scenarios.

Perhaps it adds a valuable dimension to our thinking about responsibility in a high-tech world.

A simple calculation in the Trolley Case shows that one can save four lives by throwing the switch. The majority of people think, after some reflection and calculation, that it is morally permissible - and most of them even think one has a moral obligation - to save five, although one person loses his life as a result.

The "Fat Man" case

Now suppose we change the story in the Trolley Problem a bit and take the switch out of the story. There are still five people tied up on the track and the trolley is barrelling towards them, but there is a fat man standing on a bridge over the track (see image). By pushing the fat man onto the track one can stop the trolley before it hits the five people. One would expect that people would react in the same way to this case as to the original version, since it implies the same numbers and basically the same calculation: saving five by causing the death of one.
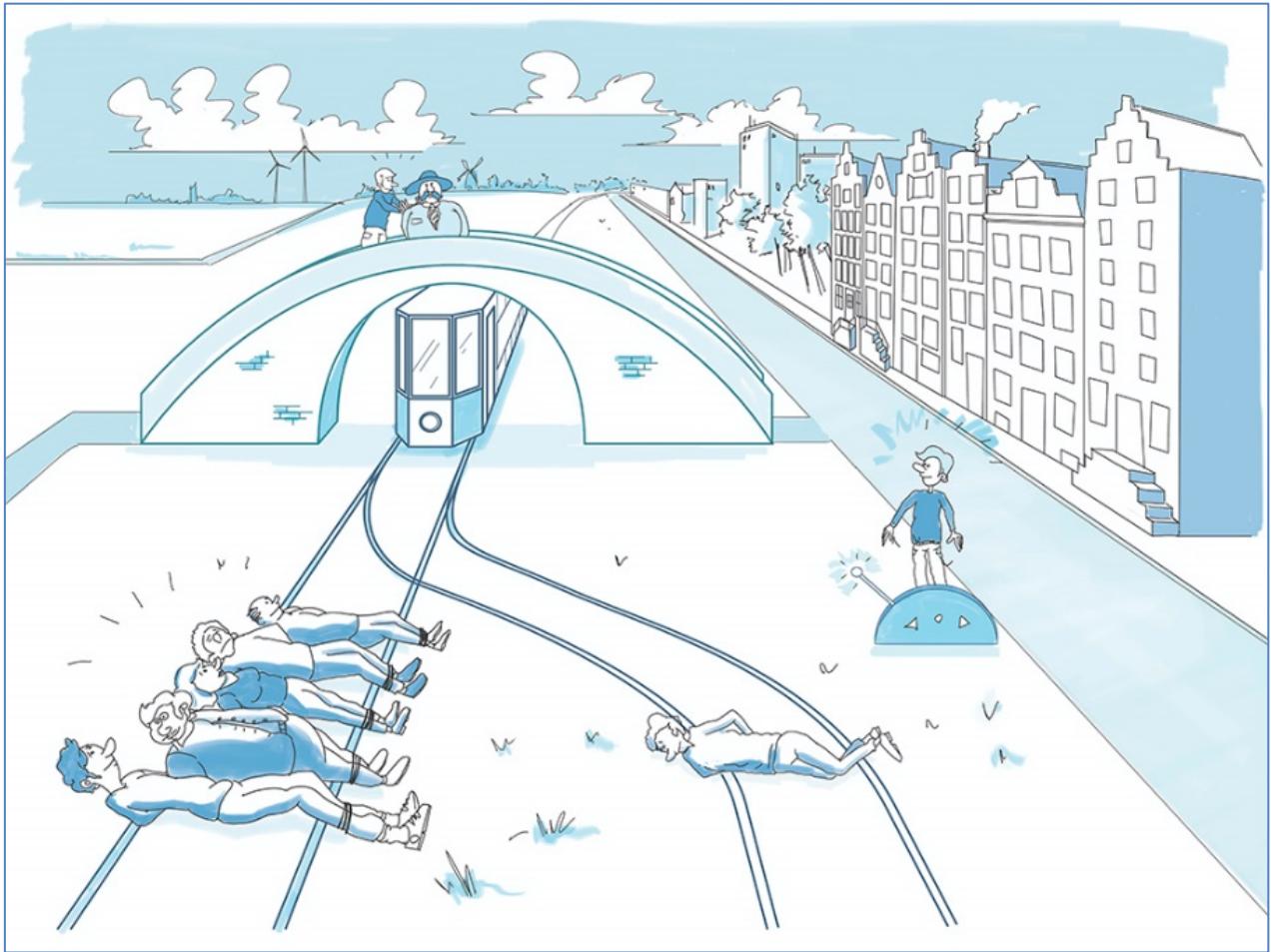
Figure 2.3 : The fat man case

Empirical research shows, however – some argue even that brain imaging studies point in this direction - that we react in a different way to this case, although the numbers and the calculations are the same. In the lever case, we primarily rely on cold reasoning and calculation in terms of lives lost. Given the option of pushing the "fat man", however, we tend to react with disgust or laughter. It seems preposterous to use a person as an obstacle and by doing so, killing him.

Nevertheless, the Trolley Problem is more than a thought experiment. For self-driving vehicles, we have to make similar decisions. What happens if an accident cannot be avoided? Does it protect the driver or the pedestrians? See the picture below.
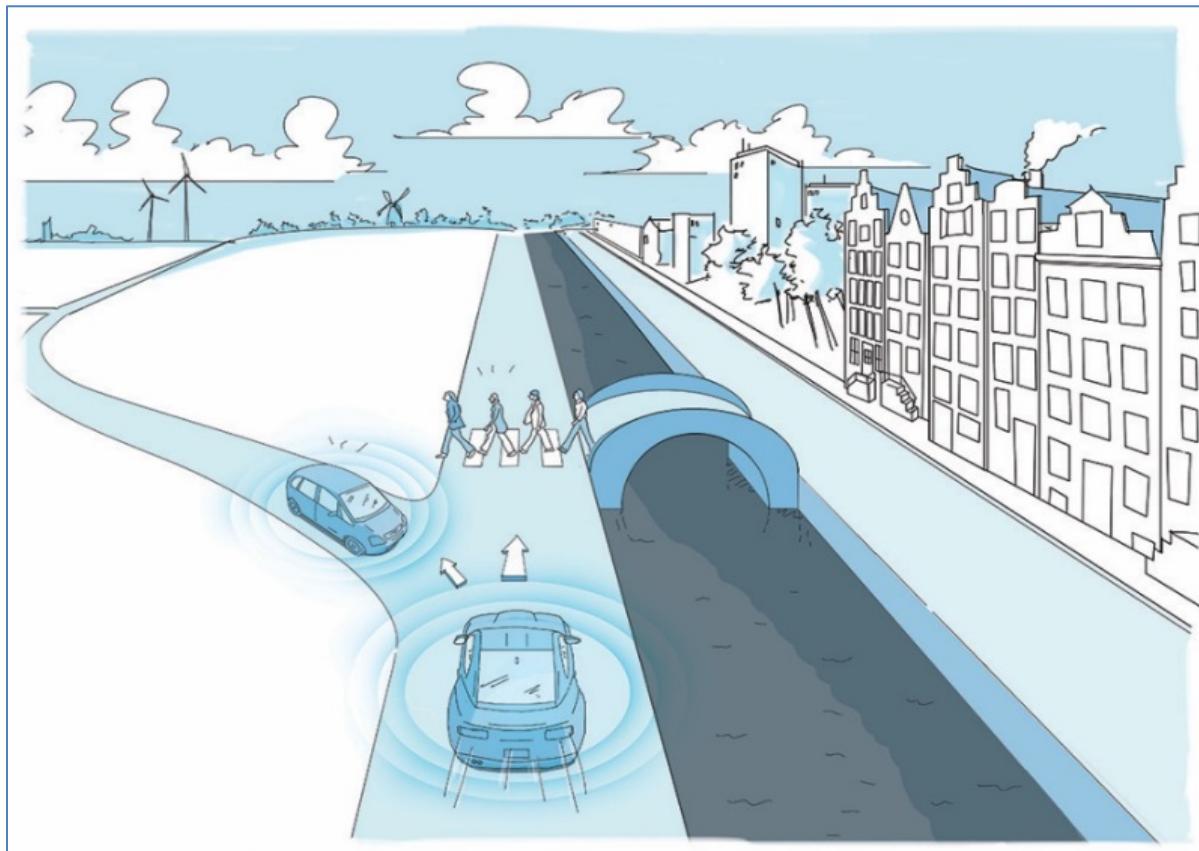
Figure 2-4: Trolley Problem for self-driving vehicles.
Check out this website for much more examples **http://moralmachine.mit.edu/**

2.2 How engineers answer the Trolley Problem

We've seen the philosophical questions that such dilemmas bring to the surface. What you will *not* find in the trolley literature, however, is the following: engineers and designers of technical products often reply to trolley cases by saying that the railway infrastructure is badly designed - and they would be right! Engineers especially would immediately start to think of better system designs and innovations, which would prevent this tragic situation from arising in the first place.

The infrastructure should have included, as they would suggest, early warning systems, automated breaking systems and kill-switches in order to prevent the need for the operator to make such a tragic choice near the switch. This may not be a legitimate move in a philosophy seminar, because solving the dilemma is not the goal. However, this line of reasoning is a very interesting move in another context, namely the one that pertains to preventing deaths in rail transport and maintaining the safety of rail infrastructure.

What this response clearly brings to the surface is that in such "trolley cases" the situation is a given and therefore unchangeable, as you would expect in a thought experiment. Engineers,

however, with their characteristic unwillingness to take the status quo for granted, would have difficulty accepting such stipulations in the thought experiment. Their goal is to change the world for the better by better designs which avoid tragic choices.

This dominant mode of moral thinking about trolleys, where conditions are given and immutable, draws attention away from the fact that problematic situations in reality typically do not come about as a result of the hard work of imaginative philosophers preparing for an academic paper.

They are the result of numerous prior design decisions by many others, and not necessarily of the final agent who faces the choice. Moral dilemmas in daily and professional life - and certainly ones that involve technology - are almost always the result of hundreds, if not thousands, of decisions and choices by different agents in complex processes. Design histories do matter in the real world and therefore it is just as important to learn how to prevent dilemmas from arising, as it is to learn how to think about them once they have come into existence.

Whether we are thinking about designing or developing intelligent or autonomous cars, IT infrastructures, new materials, foods, drugs or energy options, we are inevitably shaping the choice architectures (that is, the design of different ways in which choices can be presented) of future users. Engineers know that the best way to deal with moral problems for these situations in real life is often by anticipating failure scenarios and addressing these concerns, not just waiting for dilemmas to present themselves.

The types of moral considerations such as the trolley problem trigger discussions that can be fruitfully used in the design of high-tech innovations, systems or infrastructures. Moreover, these discussions reveal different sets of values that people have. It is therefore important to involve all stakeholders and address their values.

Responsible innovation is about anticipating moral choices and taking responsibility for others, whether those others are our fellow citizens or our grandchildren. It concerns designing and shaping technology in the understanding that future users, consumers, patients, citizens and future generations will be stuck with the choices that engineers and applied scientists come up with today, and have thought about - or forgot to think about - long before. Their ability to take responsibility will be a function of a long and detailed design history. And this applies equally to energy options, internet protocols, smart cities, new materials as to any other innovation deployed in society in some way or form.

2.3 Individual moral responsibility

Now that we have been introduced to a one class of thought experiments, let us look at some other scenarios which introduce more complexity, so that they more closely resemble real-world scenarios, and thus they include more grey areas to consider. We will use these examples to specifically examine various notions of individual moral responsibility. Understanding these distinctions is important within the context of responsible innovation, because one of the most important goals is to design and innovate in a way that promotes responsibility. In order to know how to promote responsibility, however, one needs to have a clear understanding of what responsibility is.

For now, we will focus on backward-looking responsibility - i.e. judgement of past actions - rather than considering their future obligations. Here are examples:

- ***The "Wrong Switch" case***

Let's consider a chemical accident case. We'll call this case the "Wrong Switch". Imagine that an operator at a chemical plant notes that leakage is coming from a tank, and in an attempt to contain the spill, accidentally turns the wrong switch. Imagine furthermore that an immediate consequence of this is that an explosion occurs, killing another worker. Given this information, it seems reasonable to conclude that the operator is causally responsible for the worker's death. After all, it was the flipping of the switch that caused the explosion. So, one way to test whether an agent is causally responsible for an outcome is to ask whether the same outcome would have occurred if the person did not act as (s)he did.

This way of understanding causal responsibility seems uncontroversial and seems to apply in the "Wrong Switch" case. But notice here that it's a different question to ask whether the operator was morally responsible for the worker's death. Being merely causally responsible for an outcome doesn't seem enough to conclude that one is also morally responsible for it. The operator's moral responsibility seems to depend on the explanation for why (s)he turned the wrong switch - in this case, it was an accident.

- ***The "Mixed Wires" case***

Let's now consider a version of the case that includes some additional information which explains in more detail what went wrong. Suppose that the wiring of the switches was mixed up and that the operator couldn't have known this. Because of the wiring, the operator flips what (s)he believes to be the right switch, but instead of stopping the leak, there is an explosion killing another worker. Importantly, in this "Mixed Wires" case, the operator tries to stop the explosion, but it is too late. Again, it seems uncontroversial to claim that the operator is causally responsible for the death of the worker. If (s)he didn't turn the switch, then the explosion would not have happened. But again, causal responsibility doesn't entail moral responsibility, and so we still have to ask: is the operator in "Mixed Wires" case morally responsible for the worker's death?

One way of reaching an answer is to consider a related question, which is whether the operator in this case was to blame for the death of the worker. Given that the operator couldn't have known that the wires were switched, and that (s)he couldn't prevent the explosion from occurring, despite trying to do so, it seems to be a mistake to assign the blame for the worker's death to the operator. That is to say, it would be inappropriate to blame the operator for the worker's death.

- ***The "Hateful Operator" case***

In order to understand this nuance clearly, it will help to compare this case with a version of the case where the operator is obviously blameworthy. So let's take another version of this case, and call it "Hateful Operator". Here, the situation is rather different: the operator intentionally and knowingly turns the wrong switch in order to kill the worker.

This case differs from the "Mixed Wires" case in two important respects. Firstly, the operator holds ill will towards the worker who dies in the explosion, whereas in the "Mixed Wires" case the operator held no such ill will and was actually motivated to try to stop the explosion. The second difference is that in the "Hateful Operator" case the deadly explosion is avoidable. The operator knew that (s)he was going to turn the wrong switch, and did so intentionally in order to bring about the worker's death. So, in the "Hateful Operator" case, it is intuitive to think that the operator is both causally responsible and morally culpable for killing the worker. Both the fact that the operator did something that causally brought about the worker's death

and the fact that the operator held ill will toward the worker entail that the operator is morally at fault. Importantly, this kind of moral culpability is just one way in which we can say that a person is morally responsible for some event or outcome.

- *The "Extra Effort" case*

This next case shows that it is possible to be morally responsible for something, without being culpable for it. Let's call this case "Extra Effort". This case is similar to the "Mixed Wires" case, in that the operator doesn't know and couldn't have known (s)he was turning the wrong switch. Imagine, however, that when the operator realizes that (s)he has turned the wrong switch, there are just a few seconds to turn another switch that will prevent the explosion. Imagine that turning this other switch is not the normal procedure, and that it takes some effort. Finally, imagine that the operator succeeds and that the worker is saved.

In this case, the operator is clearly causally responsible for saving the worker, (S)he had to think very quickly and had to carry out a very difficult action in order to save the worker's life. (S)he was motivated to go the extra mile in order to save the victim, and that seems to be good reason for thinking that the operator is morally praiseworthy. In this case, it is important to note that this operator is morally responsible for saving the operator's life. Being morally praiseworthy is yet another way in which someone can be morally responsible.

- *The "Routine Procedure" case*

The final version of the chemical spill scenario highlights yet another important aspect of moral responsibility. Let's call this case "Routine Procedure". In this case, there is a chemical spill and the operator turns the right switch. There is no mixed wiring and turning the switch required no extraordinary effort. In this case, you might not be inclined to think that the operator is praiseworthy for turning the switch, given that his/her actions were perfectly ordinary and didn't require a tremendous amount of effort or achievement.

It also seems obvious the operator is not culpable, given that (s)he did nothing wrong. The operator did the right thing freely and intentionally, and (s)he knew what (s)he was doing. For these reasons, it makes sense to conclude that the operator is morally responsible for avoiding the death of the worker.

What does individual moral responsibility entail?

So, what is the lesson we can learn from all of these cases? The lesson is that there seem to be several different notions of responsibility:.

- The minimal level of responsibility exists in case of causal responsibility. Recall that in the "Mixed Wires" case, the operator was causally responsible, but not morally responsible for the worker's death.
- The second notion of responsibility is moral responsibility. In the "Routine Procedure" case, the operator was morally responsible for preventing the worker's death, even though (s)he doesn't merit either praise or blame.
- The *third notion* of responsibility involves cases where the agent's actions merit praise or blame. We saw that the agent was commendable when (s)he went above and beyond the call of duty to do the right thing, and we saw the operator was culpable when (s)he knowingly and intentionally killed the worker. These agents are indeed morally responsible, but we are inclined to add that they are also commendable or culpable.

It's especially important to understand the connections between these different notions of responsibility. The first connection is that moral responsibility presupposes causal responsibility.

The operator has to cause the worker's death, in order for him/her to be morally responsible for it. Without causal responsibility, we cannot have moral responsibility. The second connection is that both commendability and culpability presuppose moral responsibility. For example, if the operator could not have avoided causing the worker's death, then (s)he is not morally responsible, and therefore not culpable either. Thus, judgments of commendability and culpability both assume that the person in question was both causally and morally responsible for the outcome.

Having distinguished between these different notions of responsibility, let's apply them to responsible innovation. As an innovator or designer, reflection on the various factors that affect the attribution of responsibility should help design processes to adapt them in a way that reduces the likelihood that something goes wrong without someone being morally responsible for it. Agents should have clear and timely information about the process and their role in it, and the system itself should be designed with multiple fail-safes that are easy to access.

2. 4 Collective moral responsibility

Introduction

In some cases, individual moral responsibility alone is not enough to address key concerns, especially when other parties who have equal influence to affect the outcome are also involved. We will be discussing a problem of collective action, which is sometimes called the "*tragedy of the commons*".

This problem can arise in the context of shared resources, such as rivers, the atmosphere and national parks. We shall focus on a typical example of a tragedy of the commons-scenario, namely overfishing.

Imagine that small fishermen from a seaside village rely on fishing for their economic livelihood. Each fishing boat in the village must compete with the others to bring in a catch. Because of this competition and the constant demand for fish, overfishing occurs.

This eventually leads to the fisheries to become depleted. The "commons" here refers of course to the natural stock of fish in the sea. But what's the tragedy? In this case, the tragedy has to do with the way that overfishing seems to be inevitable, namely due to the fact that the individual fishermen act in their own rational self-interest.

It is important to notice that it is in each individual fisherman's rational self-interest to catch as many fish as he can. If fisherman A catches less fish than his maximum capacity, he will make less money, and meanwhile his competitors, fishermen B, C, and D, will catch the fish he didn't catch. This shows that there is simply nothing to be gained and indeed there is only something to lose, namely profit, by catching fewer fish than the maximum amount. Thus, it is in fisherman A's rational self-interest to maximise his catch. Importantly and unfortunately, the same logic holds for the other fishermen. As each fisherman only acts in his individual rational self-interest, the common stock of fish is soon depleted.

Figure 2.5 Overfishing

Although the individual fishermen apparently take rational action, this behaviour does not contribute to the best interests of everyone collectively in the long term. The community's interests are damaged, because they risk losing an important source of income, the basis of their diet and economy. In addition, the individual fishermen's interests are also set back because they are losing their livelihood. Given these effects of depleting the fish stock, it is clear that when considered as a collective, the individual fishermen's actions were irrational. So, even individually rational actions can turn out to be collectively irrational.

The solution in order to avoid this tragedy is to co-operate. Rather than trying to catch as many fish as they can, individual fishermen should practice sustainable fishing. Sustainable fishing means taking an amount of fish from the ocean that is consistent with the continued health of the fish stock. This would mean that sometimes, individual fishermen would have to leave some fish in the ocean, even when they are fully capable of catching them. Sustainable fishing can be realised in a co-operative scheme, such as a fishing quota scheme, which limits the size of the catch for each boat. However, in order for this to work, the whole community, and especially the fishermen, must agree to it. That is, they must come together to establish the quota of fish that is consistent with sustainable fishing, and they must stick to it.

The problem of freeriding
But you might be wondering why the fishermen would stick to this scheme. Think back to individual rational self-interest and consider only fisherman A. If all the other boats comply with the quota scheme, then it is in fisherman A's rational self-interest to fish more than the quota. This is called freeriding. The same reasoning would once again apply to all the other fishermen as well. So, although the point of the collectively rational co- operative scheme was to avoid depleting the common stock of fish, it would actually be undermined by individual rational self-interested free-riding.

So what options are there for getting individuals to stick to a collective quota scheme? What would actually motivate cooperation in this case? One thing that might motivate individual

fishermen is morality. But what moral considerations might there be in this context? In fact, there seem to be several. First, fishermen might see as a moral reason for sticking to the quota the fact that sustaining the stock of fish is a shared and desirable goal and the quota is the means to this shared, desirable end. They may thus be motivated to take the necessary means to achieve the shared, desirable aim of sustaining the commons. Secondly, the fishermen may be motivated by the fairness of the cooperative scheme, if it were designed in a way that sustains the stock, while not giving any one fisherman an unfair share or advantage. Even if individual rationality encourages free-riding, fishermen who are motivated by the morality of the quota system might stick to it.

Note, however, that even though moral motivation may be necessary, it's not sufficient for actually realizing sustainable fishing. This is because we simply cannot count on everyone to be motivated by moral considerations. Many will only do what they morally should do, if they are forced in some way to do it. In order to make up for the lack of sufficient moral motivation, we may rely on enforcement. For the quota system to work, some significant degree of compliance must be achieved. There are several options for enforcing compliance. For example, if the community authorizes a maritime police to enforce the quota system, even those who aren't morally motivated may avoid free-riding. Through fines or penalties, such as revoking the license to fish, this enforcement shifts the individual rational self-interest to align with collective rationality.

The limits of enforcement

Unfortunately, even enforcement measures are not sufficient in themselves. Given the sheer number of fishing boats and the large area in which they fish, it is practically impossible for the maritime police to ensure compliance. Moreover, the maritime police itself, if it is acting in its own individual rational self-interest, may be lax on enforcement, either by taking bribes or simply by being lazy.

What is the solution to the tragedy of the commons, then? So far, we have seen that a co-operative fishing quota might be the best way to sustain the fisheries. However, the moral motivation to achieve a collective good is challenged by the individual self-interest to take advantage of the situation. This means that some kind of enforcement becomes necessary, although this in itself is not sufficient either. What if both the fishermen and the maritime police were morally motivated to sustain the fishing quota? Making such moral considerations salient to all parties, particularly when they might be tempted to disobey the rules, is an interesting design problem that responsible innovators should try to tackle.

2.5 Responsibility in complex systems

Introduction

So far, we have seen cases where it is easy to assign responsibility - and therefore blame too - when something goes wrong, by finding out who is causally or morally responsible. Unfortunately, the real world is very complex, with multiple stakeholders working together, influencing each other's outcomes. It becomes much harder to pinpoint who is causally or morally responsible, and who is to blame.

What we see is that the actions of all stakeholders *together* lead to a dramatic outcome, but none of the individuals involved can be held responsible. *This phenomenon is called "the problem of many hands*". Because there are many people involved, it is impossible to identify one single person that is responsible. This problem is very urgent in engineering, because there are often

Figure 2-6: Complexity

many people involved in the development of technology, even in risky technology - if anything were to go wrong, there could be serious consequences. How can we deal with the distribution of responsibility in complex socio-technical systems?

The conditions for moral responsibility

Let us start with the responsibility of engineers. Engineering often takes places in teams or networks of many people. Before we can discuss the responsibility of these groups, we first have to question what we mean when we say that an individual person is responsible. Usually we say that a person is responsible if the following four conditions are met:

1) **The freedom condition:**
   The person should be free to act and not be under external pressure. If I put a gun to someone's head and ask this person to do something illegal or immoral, this person cannot be held responsible. S/he was not free to do otherwise.

2) **The knowledge condition:**
   A person should have the knowledge that his/her action would lead to a negative outcome. If the person does not know this, s/he will generally not be held responsible. If, for example, someone painted the door of his house without putting on a notification that the door was wet, and you happen to touch the door and thereby destroy the paint job, it is not fair to hold you responsible or to blame you. You did not and could not know that the door had just been painted and that therefore you should not have touched it.

3) **The causal connection:**
   there should be a causal connection between the act of the person and the negative outcome. I cannot be held responsible for things I did not causally contribute to. However, note that sometimes doing nothing is the wrong act: if one has the possibility to save another from harm, not doing anything is the wrong act.

4) **The transgression of a norm:**
   if what you did was somehow faulty, then we can say you transgressed a norm. This can be a legal norm, but also an ethical or social norm. This is a difficult condition,

The problem of many hands

Now let us look at an example in which several people are involved: the development and use of a new fire-resistant material. There are four people involved:

- Person A is working in the laboratory and is doing fundamental research into the atomic properties of this new material
- Person B is hired by the fire brigade to design a new outfit for the firemen, using this promising new material;
- Person C is the director of the fire brigade who hired the designer,
- and Person D works at the fire brigade and is responsible for cleaning the firemen's outfits; s/he brings them to a dry-cleaning store for cleaning.

As it turns out, this promising new material becomes carcinogenic when brought into contact with washing powder. One of the employees of the dry-cleaning store develops a lethal type of cancer and eventually dies. Can we say that one of the persons A, B, C or D is morally responsible for the death of the cleaner?

Looking at the four people, we find that all of them made some causal contribution. But the other conditions listed above are probably not fulfilled; at least, we can say that none of the individuals fulfils all conditions. The person working in the laboratory may have known that this material could have a chemical reaction with other materials, but he could not foresee how others would use the material. The other persons probably did not know about the carcinogenic properties of the material. One may even say that the person responsible for cleaning was not really free to act differently, as there were no other options for cleaning.

So, the actions of the four people together led to an unfortunate dramatic outcome, but none of the individual persons can be held responsible. This case shows how the "problem of many hands" works in practice. Because there are many people involved, it is impossible to identify one single person who is responsible. This problem is very urgent in the engineering of complex or dangerous technologies, because there are often a great many people involved in the development of the technology, not to mention that there is a potentially high impact when things go wrong.

An example is the oil spill of the BP platform in the Mexican gulf, an industrial disaster that began on 20 April 2010. The impact of this disaster was huge and it immediately prompted the question of who was responsible for this disaster.

Figure 2.7: BP-disaster

The problem of many hands is often discussed in a backward-looking sense, that is, after a negative event has happened. However, we can also frame it in a forward-looking sense. We can then check against the conditions of moral responsibility to see if a person has the ability to fulfil his/her responsibility: does this person have the freedom to act? Does (s)he have the necessary information? Are the right norms in place?

Building responsibility into technology

This brings us to an interesting topic: the relationship between responsibility and technology. The autopilot in an airplane is a clear example of technology taking over responsibility from a person. But equally, can technologies be developed in such a way that they enable people to assume responsibility? We think that technology can indeed take up this role, but in order to ensure this, we should pay attention to specific aspects of responsibility when technology is being developed. Here are two examples:

Example 1: V-chip

The first example we consider is the V-chip. The V-chip is a technological device designed to prevent children from watching mature television content. TV stations broadcast a rating as part of a program. Parents can program the V-chip by setting a threshold level rating, so that all programs above that rating are automatically blocked by the V-chip when it is turned on. Thus, children watching TV cannot view the blocked programs. **Some people** argue that by using the Vchip, parents transfer responsibility to the TV stations, because the TV stations decide the exact rating of each program and thus determine whether this program will be shown on television or not. From this viewpoint, the V-chip limits the freedom of parents. Others say, however, that the V-chip provides parents with more information on mature content; as such, it gives them more freedom to control what their children are watching. Whether the Vchip limits or enhances parents' responsibility is open for discussion, but the example clearly shows that technology can and does affect a person's responsibility.

Example 2: Control Room

Another example would be a control room. A control room is a central space from which a large facility or service can be monitored and controlled. These rooms are often equipped with multiple monitors and screens (see image). The people working in the control room have to make decisions on the basis of huge amounts of information.

That means that the layout of these rooms, and the way the information is presented, determines the extent to which people are able to make the correct decisions.

We could argue that a badly designed control room may hinder people from assuming their responsibility. Vice versa, a well-designed control room may enhance a person's ability to carry out his/her responsibility.

Thus, technology can empower, but also hinder people in carrying out their responsibilities. One aspect of responsible innovation is therefore to develop technology in such a way that it may facilitate or strengthen people in their ability to carry out their responsibilities

Figure 2.8: example control room

2.6 Emotions and values

Introduction

The risks arising from technologies raise important ethical issues for people living in the 21st century. Consider the possibility and potentially disastrous consequences of accidents, pollution, occupational hazard or even environmental damage. Due to the subjective perception of such risks, controversial technologies can trigger strong (negative) emotions, including fear and indignation, which often leads to conflicts between experts and laypeople.

Emotions are generally seen as an annoyance in debates about risky technologies, because they seem irrational and immune to factual information. However, we will argue here that emotions can be a source of practical rationality. Natural emotions, like fear, sympathy and compassion, can help us to grasp the morally salient features of risky technologies, such as fairness, justice, equity and autonomy, that might be otherwise overlooked in conventional technocratic approaches to risk.

The difference between risk and risk perception

Responsible innovation is especially challenging in the context of risky technologies, such as nanotechnology, synthetic biology and information technologies. These technologies often give rise to heated and emotional public debates. While experts emphasize scientific studies that point out the supposedly low risks, the public is often concerned about the impact of such technologies on society. Experts like to point out that the worries of the public are due to a lack of understanding, but this makes them no less real.

Policy makers usually respond to this in one of two ways: they either ignore the emotions of the public or they take them as a reason to prohibit or restrict a technology. Let us call these two extremes the technocratic pitfall and the populist pitfall respectively. In both pitfalls, there is no genuine debate about the emotions, public concerns and moral values. However, this should be rectified.

Social scientists, psychologists and philosophers have argued against the technocratic approach for decades. They have pointed out that risk is *more* than a quantitative, scientific notion. Risk is

more than the probability of an unwanted effect that we can assess through cost-benefit analysis, as conventional, technocratic approaches assume. In other words, the experience of risk is something quite different than an calculation of risk.

Risk concerns the wellbeing of humans and it involves ethical considerations such as fairness, equity and autonomy. There is a strong consensus amongst risk scholars that ethical considerations should be included in any risk assessment. Interestingly, as we know from the influential work of **psychologist Paul Slovic**, these considerations do come up in the risk perceptions of laypeople. Apparently, the pre-theoretical connotations that people have about risk include ethical considerations that are normally excluded from the quantitative-oriented approach to risk that experts are using. As such, several risk scholars have argued that laypeople have a different, but equally legitimate rationality as experts.

It has become more and more clear that laypeople's risk perceptions are largely influenced by their emotions. Social scientists struggle to deal with this, as they understand emotions to be irrational, which seems to undermine the idea that laypeople might employ an alternative, legitimate rationality concerning risks.

Emotions as a guide to acceptable risk

However, emotions are not necessarily a threat to rationality. The neuropsychologist Antonio Damasio has **famously shown** that without emotions, we cannot be practically rational. Indeed, the dominant approach in emotion research in current philosophy and psychology is the so-called cognitive theory of emotions, according to which emotions are a form or source of cognition and knowledge. These ideas can shed a completely new light on the role of emotions in debates about risky technologies. Rather than being opposed to rationality and hence inherently misleading, emotions can be seen as an invaluable source of wisdom when it comes to assessing the moral acceptability of risk.

The emotions of the public can provide insight into reasonable moral considerations that should be taken into account in moral decisions about risky technologies and responsible innovation.

Experts might feel responsible and even worried about the technologies they develop. This worry and fear can point out concerns about the unforeseen negative consequences of a technology. Fear can indicate that a technology is a threat to our wellbeing. We often feel disgust when confronted with clones and human-animal hybrids, for example; this in fact indicates that creating such beings is ambiguous from a moral point of view. Meanwhile, indignation may be an indication of a violation of autonomy, in case of risks to which we are exposed against our will.

It is often thought that emotions are by definition opposed to technology and therefore one-sided, but this is not necessarily the case. Enthusiasm for a technology, for example, may suggest that is has benefits for our well-being. Sympathy and empathy can contribute to our understanding of a fair distribution of risks and benefits.

As such, emotions can draw our attention to important moral considerations that may otherwise be insufficiently addressed. These insights allow for a different way of dealing with emotions about risk in public debates, by avoiding both the technocratic pitfall and the populist pitfall.

Figure 2.9: Protest against nuclear energy

This alternative approach, which we call an "emotional deliberation approach to risk", gives the public a genuine voice, in which their emotions and concerns actually get heard and discussed. It can provide us with ideas on how to communicate about risks in a morally responsible way. Moral emotions in turn can provide important insights into moral constraints and the desirable parameters of responsible innovation. For example, in debates, experts should not only focus on the small probabilities of possible risks, but they should also provide a balanced outlook on both positive and negative consequences, allowing individuals to make an informed assessment.

Involving emotions in deliberation and communication about risks can also contribute to necessary changes in behaviour. For example, appealing to emotions in campaigns about climate change can increase the currently lacking "sense of urgency" , and at the same time provide the motivation to contribute to environmentally-friendly behaviour. After all, emotions are an essential source of motivation and should therefore be harnessed to stimulate change.

When developing risky technologies, we argue that emotions and moral concerns have to be taken seriously in order to come to a well-grounded ethical assessment. At the same time, this approach can help overcome the gap between experts and laypeople that occurs over and over in debates about risky technologies. Thus, the public will feel that their concerns are taken seriously, which will contribute to participative and responsible innovation.

2.7 Moral dilemmas and moral overload

Scientists, engineers and designers often feel the obligation to make the world a better place: to make the world safer and more sustainable, to create new jobs while simultaneously protecting privacy and fighting terrorism, to give autonomy and freedom to future users, to improve the quality of life for future generations. They want to achieve all of these things, but, like many people who want to do the right thing, they encounter the problem of moral overload.

The problem of moral overload is that there is just too much good to be done; we have too many obligations that we cannot fulfil, at least not all at once. We want economic prosperity and jobs for all, but also sustainability. We value our security, but also our privacy. We demand safety, but are not willing to sacrifice our freedoms. We require accountability, but insist on the right to confidentiality.

Such moral problems in science and technology often take the form of a moral dilemma. The most basic definition of a moral dilemma is the following syllogism:

a) *The agent ought to do A.*
b) *The agent ought to do B.*
c) *The agent cannot do both A and* B.

It is important for a better understanding of responsible innovation to become acquainted with some of the peculiarities of moral dilemmas. Specifically, we would like to demonstrate how innovation and design may be a way of dealing with moral dilemmas.

Dealing with moral dilemmas

To the extent that technologies embody some of our values, they too can simultaneously call into question which value we desire more, presenting at best an uneasy compromise.

Consider the following examples:
- CCTV cameras: do we value our privacy or our security more?
- Nuclear power plants: do we want energy security and lower CO_2 emissions, or less exposure to risk?
- Drones: do we want our soldiers to be safe or accountable?

Dilemmas such as these are typical moral dilemmas. Anyone who is confronted with a dilemma has a number of obligations but cannot fulfil all of them.

So what do you do?   There are various strategies for dealing with moral overload or moral dilemmas.
One way to deal with a moral dilemma is to look for the option that is best all things considered. Although this can be done in different ways, it will usually imply a trade-off among the various relevant value commitments. In other words, it will create a "moral residue".

 Moral residue here refers to the fact that even if we may have made a justified choice in the case of moral overload or a moral dilemma, there remains a duty unfulfilled, a value commitment not met.

However, the moral residue (or guilt is) , not just an unfortunate by-product we have to live with.

 The philosopher Ruth Barcan Marcus has **proposed a view** of this moral residue, or residual regret, that we believe is very relevant for our discussion on responsible innovation. It could potentially offer a solution to our dilemma.

She argued that we have a second-order duty to avoid moral dilemmas known as ' "Ought implies Can"': One ought to act in such a way that, if one ought to do X and one ought to do Y, then one can do both X and Y

Figure 2-10: Ruth Barcan Marcus

The principle entails a collective responsibility to create the circumstances in which we as a society can live by our moral obligations and our moral values.

New technologies as potential solution

One way in which we can do so is by developing new technologies. See also our first case study about smart meters. Technical innovation can entail moral progress by creating new opportunities to meet conflicting values ate the same time. This is the essence of RI.

Of course, not all instances of technological innovation entail moral progress.

One reason why technical innovation does not necessarily result in moral progress is that it may result in a 'technological fix,' i.e. a technical solution to a problem that is social in nature. For example, world hunger is not primarily a problem of production capacity but rather a problem of distribution of food, corruption, income, disasters and land, which is far less amendable to technical solutions.

We should also be aware that technical innovation not only enlarges the range of options but that new options also bring new side-effects and risks. This may introduce new value dimensions that should be considered in the choice situation and these new value dimensions may create new forms of moral overload. Nuclear energy may help to decrease the emission of greenhouse gases and at the same time provide a reliable source of energy, but it also creates long-term risks for future generations due to the need to store the radioactive waste for thousands of year. Technical innovation may also introduce difficult ethical choice in situations in which there was previously no choice. An example is prenatal diagnostics. This technology creates the possibility to predict that an as yet unborn child will have a certain severe disease with a certain probability. This raises the question whether it is maybe desirable or allowed to abort the foetus in certain circumstances. This choice situation is characterized by a conflict between the value of life (even if this life is not perfect) and the value of avoiding unnecessary suffering.

Implications for the Responsibility of Engineers

This higher order moral obligation to see to it that can be done what ought to be done can be construed as an important aspect of an engineer's task responsibility. This is also known as a meta-task responsibility.
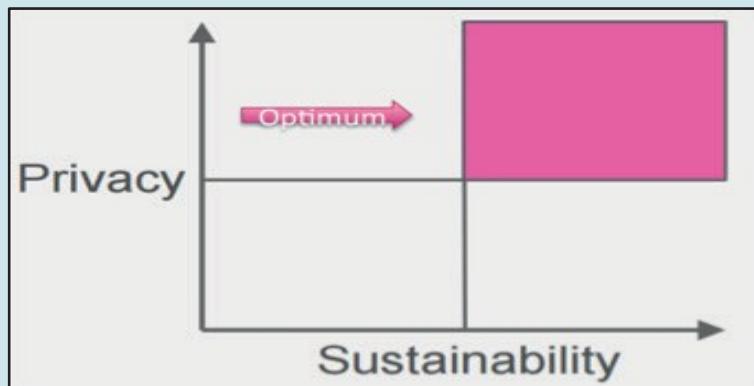
An interesting way to fulfil this responsibility is the approach of Value Sensitive Design. We will discuss this in chapter 9.

Case study #1: Smart meters and conflicting values as an opportunity to innovate

Let's look at the example of smart meter design. Smart meters are a good idea because they help us in becoming more sustainable, but people have raised concerns about their impact on privacy.

We have to design meters that satisfy all the functional requirements - thus serving the goals of sustainability - but in addition, they need to respect our privacy pertaining to household electricity consumption, and by extension, knowledge of our daily comings and goings.
The dilemmatic structure of our problem in smart metering is as follows:

Figure 2-11: Meeting different requirements

The purple area in the figure above is the area that we are interested in for our ideal moral solution. Here we can satisfy both of our values, above a certain reasonable threshold level. A first-generation smart meter may neither get us the desired level of privacy nor the desired level of sustainability. The smart meter 2.0 may give us one, but not both. The smart meter 3.0, which is what we are ideally looking for, is designed to accommodate both of the functional requirements in order to make energy use more efficient, while also protecting personal data. It gives us privacy and sustainability. In this respect, innovation in smart metering is exactly this: the reconciliation of a range of values, or moral requirements, in one smart design, some of which were actually in conflict before.

Similarly, if we would like to benefit from RFID technology (enabling to automatically identify and track tags attached to objects) in retail, but fear situations in which we might be tracked throughout the shopping mall, it has been suggested we can have it

both ways. A so-called "clipped chip" in the form of a price tag with clear indentations would allow customers to tear off a piece of the label, thereby shortening the antenna in the label so as to limit the range in which the label can transmit data.

There are more examples which illuminate how we can take moral obligations seriously - towards customers, future users, future generations, the climate and even flora and fauna. We can confront difficult moral choices not by compromising on our value commitments or doing more philosophical homework, but by changing the world through applying creativity, knowledge and skills.

Moral dilemmas can help stimulate creativity and innovation, and innovative design may help us to overcome problems of moral overload**.**

Case study #2: Medical ethics in the age of AI and big data

This case study is based on a talk *given by Jeroen van Hoven at the WHO Global Summit 'Bioethics, Sustainable Development and Societies' (Dakar, Senegal), 23 March 2018.*

Introduction
As discussed in this chapter, the core idea of responsible innovation is that we try to accommodate as many of our moral values as we can by design, by tweaking the world, by innovation, by creativity. This also applies to medical ethics in a world of big data and AI. We should to accommodate our moral concerns AND make use of big data and AI in health care. There is no guarantee that this will always be possible. But because the stakes are high, we have the obligation to explore whether there are suitable solutions.

Digital technology and its impact on health care.
Digital technology affects health care in all its dimensions: research & development, clinical practice, policy, innovation, entrepreneurship, insurance and financing. It changes everything. It not only enables new practice, but is a constitutive technology. It is obvious that data and AI can reduce costs in health care, improve patient safety, empower patients, improve the quality of diagnosis, therapy and patient journeys, and create more efficient billing and logistics. Smart phones and watches with health apps, wearables and so-called digi-ceuticals are part of an Internet of Things revolution that is well underway in the health sector. Wearable devices can be used to detect arrhythmia, predict Parkinson's disease (via the accelerometer in the phone), and measure a range of biomarkers such as blood sugar, blood pressure, fat percentage, oxygen and stress levels. They can diagnose skin cancer and retina damage and assist in the management of eating disorders, phobias, depression, chronic pain and PTSD. They can even gauge the risk of suicide on the basis of social media posts – looking at the time of day of the post, the number of human faces it contains and the colours.

Big tech companies are moving into health care and biomedicine on a large scale. There is a lot of money to be made and big data to be harvested. Big data in turn will drive the development of more powerful machine learning and AI, which leads to a superior position in the market. IBM applies its Watson technology in oncology and uses health data collected via Apple ResearchKit and HealthKit. The Alphabet conglomerate focuses on AI for health via its London-based partner Deep Mind Health. It has closed deals with NHS and individual hospitals to gain access to large amounts of patient data from the UK and is trying to achieve breakthroughs comparable to beating human world champions at chess and go. Alphabet already claims to be able to predict the death of patients in hospital much earlier than

clinicians using traditional methods. Microsoft has just started a new health care division in Cambridge, looking at medical algorithms. Apple has chosen the hardware route and focuses on smart phones and wearable sensors as medical devices in collaboration with Stanford University. Amazon has teamed up with Warren Buffet's investment company Berkshire Hathaway and investment bank JP Morgan Chase to move into health care.

Serious concerns

However, there are serious limitations to a purely data-driven approach and the glorification of statistical correlation. In medicine – and in other fields of great social importance – the data-driven approach, while it can be clinically useful and morally responsible, needs to be complemented by theory-driven approaches which aim at uncovering causal mechanisms.

Attitude problems

Another problem is caused by the tech companies' attitude to health care. The digital industry and Silicon Valley approach to health care is a solutionist approach, which focuses exclusively on problems for which we have nice and clean technological solutions at our disposal. David Lazer has called their approach 'Big Data Hubris': the idea that there are simple digital solutions to complex problems in the very complex world of health care, with its very complex institutional settings, multiple stakeholders, plurality of moral values and great cultural diversity. This is culpably naïve.

Moral problems

There are not only *epistemic* failings in the digital usurpation of the health domain; there are also *moral* concerns. We know that there are race and gender - and many other - biases in health care data and they may become entrenched in algorithms, or even be built into medical systems with the conscious aim to deceive, in order to save money or make profits. US-based company Aspire Health, for example, tries to save money in palliative care by estimating which patients will die soon. Moreover, algorithms in decision support systems affect the fiduciary relationships between doctors and patients. Furthermore, there have been massive breaches of security and privacy in health care in the last decade. Deep Mind Health has been reprimanded by the Information Commission of the UK for its processing of NHS patient data. Deep Mind replied that it had "underestimated the complexity of the NHS and of the rules around patient data, as well as the potential fears about a well-known tech company working in health."

It is all about trust.

But the key question of course is: why would we trust Facebook, Uber, Google, Amazon and Microsoft with all of our sensitive medical data? They can't even fix basic problems regarding fake news, data security, filter bubbles and bias, nor can they prevent the data of 50 million users being abused to run political campaigns. Big tech is essentially about quarterly revenues. These companies come to health care with a Silicon Valley approach to innovation: innovate in the grey zone, move fast, break

things first and apologize later. This is not a very helpful approach in health care.

This poses one of the most important and hardest problems of the twenty-first century: trust, or rather, the lack of trust. We could all benefit if only we could trust others with our data. If we cannot trust, or misplace our trust, the cost will be enormous. Can we trust big tech and their acolytes and subsidiaries with our health data? Can we ever be sure that their services will not be solicited by foreign (failing) states? Can we be sure that they (and our data) will not merge with companies and data bases in the hands of oligarchs who do not feel constrained by the rule of law?

It is against this background we need to situate the discussion about sharing and using identity relevant data. Ethics in the digital world should be about designing things, systems, devices, algorithms, governance systems, protocols and combinations of them. If we do not consciously and carefully design to safeguard our shared moral values in the age of technology, then our conceptions of privacy, accountability, democracy, autonomy, safety and security will simply not survive. Medical ethics needs to shift gears in the age of big data and AI if it wants to save the lives and dignity of human beings, as well as stay relevant. We cannot compromise on ethics, human well-being, human dignity and public interest under the pressure of profit maximization. It is too late to insert ethics when the rubber hits the road.

Part III:
Institutions and Values

Figure 3.1: Institutions and values

3. Institutional context of innovations

3.1 Introduction

The institutional context in which a technology is being developed and implemented is very important, because values are not only embedded in technology, but also in the institutional context.

Our main focus in this section is on technological projects with a spatial impact. Examples are infrastructural projects, such as the construction of roads and dikes, or energy projects such as wind farms and even natural gas production. Many such projects have had to deal with public acceptance issues. The public have historically opposed the construction of railways, transmission lines, carbon capture and storage (CCS) projects, waste facilities, etc.

Substantive and procedural values

One of the claims of responsible innovation is that, if these projects are designed in such a way that they are more acceptable and sensitive to the values at stake, this will increase public support for such projects. However, value-sensitive design of technology alone will never be sufficient to develop projects that are acceptable or accepted. This has to do with the fact that besides the so-called substantive values – i.e. values that relate to the technology itself, such as safety or efficiency – *there are also procedural values that determine the acceptability of a technology*.

Procedural values refer to the way decisions are taken and projects are being executed in a particular policy environment. Literature in the field of Science and Technology Studies (STS) shows how responses to new technologies are largely determined by the process through which the public are informed and involved.

This means that the acceptability of a new energy project is determined not only by the characteristics of the technology itself, but also by the characteristics of the decision-making procedure. Values such as transparency, fairness, and procedural justice are of the utmost importance. The importance of procedural values suggests that value-sensitive design for responsible innovation requires a broader scope than just the technical design of technology.

Institutions and their values

Values are not only present in technology, but also in the rules and regulations under which these innovations are developed and introduced. Therefore, it makes sense to extend the scope of our discussion beyond technology and include institutions as well. By institutions we mean the 'rules of the game' that can both constrain and facilitate certain behaviours. In literature, these rules are often referred to as institutions. The **following definition** by Jeff Hodgson is quite illuminating:

Institutions are systems of established and embedded social rules that structure social interactions.

Institutions can be both formal and informal. Examples of formal institutions are laws, standards, regulations and contracts. Informal institutions could be customs, traditions and routines. Both formal and informal institutions embody certain values.

This is most obvious for formal institutions. For example, the law prescribes that project developers must conduct an environmental impact assessment (EIA) of their planned project. This assessment is done to safeguard the value of environmental health and safety. It is interesting to note that in controversies, institutional rules such as environmental impact assessment often become hotly contested. People do not always agree with the scope of the assessment, for example. By focusing on a particular set of values - environmental health and safety in this case - some of these values are prioritized in the decision-making process at the expense of others like affordability or accessibility. This shows how values embedded in institutions are intertwined with the acceptability of technologies.

Informal institutions are perhaps less tangible, but they equally embody certain values. Routines, for example, represent ways of doing things that, by their repeated enactment, don't require much mental effort. This makes efficient behaviour possible, but there is a downside in that they implicitly favour certain unspoken values over others.

The use of jargon is a good example. The repeated use of particular words or abbreviations may turn into an efficient jargon that facilitates easy and quick communication among peers. However, it functions also as a mechanism that excludes (lay)people who are unfamiliar with that jargon. This may therefore hamper the involvement of public in decision-making. The resulting risk is that certain public values end up not being represented.

Accounting and designing for public values

If we want to design for values, this means that we should not only think about the design of technology, but also about how institutions can be designed or re-designed in order to accommodate divergent values. It is the task of the analyst to identify values that are (deeply) embedded in both formal and informal institutions, as well as the (potential) conflicts between these values. This implies the study of a broad empirical domain: legal frameworks at different territorial levels, but also strategies, cultures, and routines in a variety of segments of civil society, industry and policy.

The institutional context is not static nor fixed, but rather changes with time and place. This means that the acceptability of a technology - or what is perceived as acceptable - also changes over time, and depends on the context in which technology is developed and implemented. For instance, in the Netherlands, the value of flood safety is currently being reformulated, as a reaction to both changes in the perceived threat of floods and the degree of acceptance of high dikes as the primary means of protection against them. This suggests that neither values nor the way they are translated can be taken for granted. Indeed, values emerge and transform during the development and implementation of technology.

If we want to design for values, this means first and foremost that we cannot rely solely on an ex-ante assessment of the relevant values. Rather, it requires ongoing and continuous assessment of public values, in order to make sure that newly emergent values can also be accounted for. Secondly, it means that a design can only be value-sensitive when it is adapted to the context at hand, in terms of space, time, culture etc. There is no such thing as a fixed blueprint for value-sensitive design of a particular technology. If we really want to design for values, this means we have to go out and talk to people in order to find out what the technology means to them, how it affects them, what is at stake for them, how they want to be involved or not, etcetera. It requires the use of methods highlighted by the social sciences.
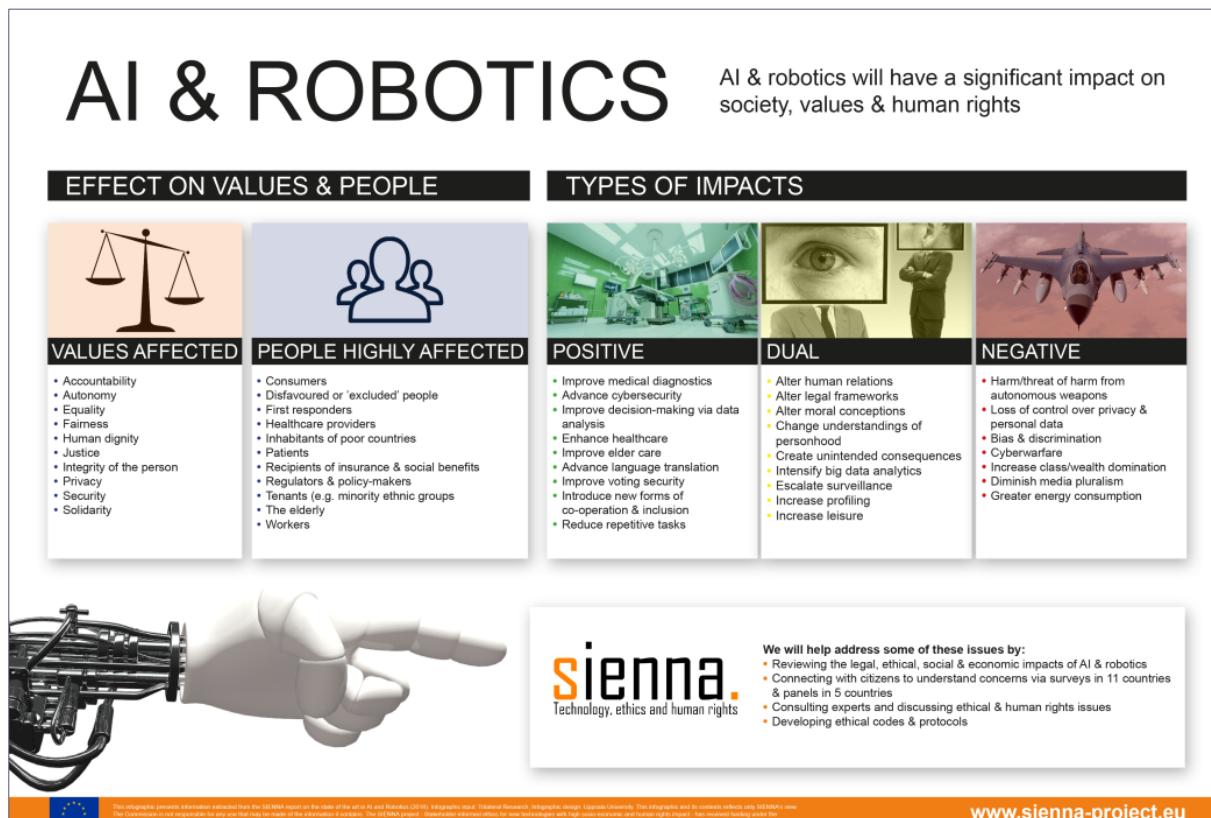
Figure 3.1 Pros and cons of AI and robotics

Understanding the values of developers and policymakers

So far, we have talked about the public and its values. But in order to comprehensively understand how to implement value-sensitive design (VSD, see the last chapter of this book for more details), it is also important to consider the values and beliefs of technology developers and/or relevant policymakers. This is because we know that the way the public responds to technology depends heavily on the way technology developers or policymakers communicate with them.

Let's consider a quick example. A label that project developers often use to describe public opposition to new technologies is NIMBY, an abbreviation for "Not In My Backyard". The NIMBY label claims that people oppose new technologies because they put their own private short-term interests - for example a quiet and aesthetically pleasing living environment - before collective long-term interests, for example a secure energy supply from wind turbines, that may not always be an aesthetically pleasing view.

This jargon, however, is strongly linked to the deep-rooted belief that the public is ill-informed and risk averse. Such beliefs shape how project developers interact with the public. If a project developer thinks that the public is ill-informed, in his communication he will probably focus on providing technical facts and explaining the safety of the project.

Yet, as we saw earlier, the public may be more concerned about procedural issues, such as fairness and transparency, or the distribution of costs and benefits. These concerns are not addressed by providing more information on just the technology and the associated risks. This mismatch, based on assumptions on both sides, frustrates the communication process. This may lead to the paradox that efforts to prevent opposition by providing "the hard facts" may actually

provoke even more public opposition, since the public feels its concerns - and thus its values - are not taken seriously.

Designing for values therefore means that we, as technology developers, need to think about and reflect upon our own beliefs and values in order to investigate how these assumptions may steer our interactions with other stakeholders. It is imperative to accept this need for reflexivity, accepting that there is a range of values and problem definitions at stake in case of technology.

Accounting for institutional values in innovation

By now it should be clear that value-sensitive design is about more than just the technologies themselves. It is equally about the (re-)design of institutions. The four main action points when designing for and around institutional values are listed below.

1. Value-sensitive design is about technology and institutions;
2. Value-sensitive design requires ongoing and continuous assessment of public values;
3. Design should be adapted for specific contexts, not based on a standard blueprint;
4. Design should include reflection by designers, technology developers and policymakers.

The above shows that in many cases there are many institutions involved and many different types. Oliver Williamson has identified different categories of institutions in his *Four-Layer Model of Institutions*, as illustrated in the figure below.
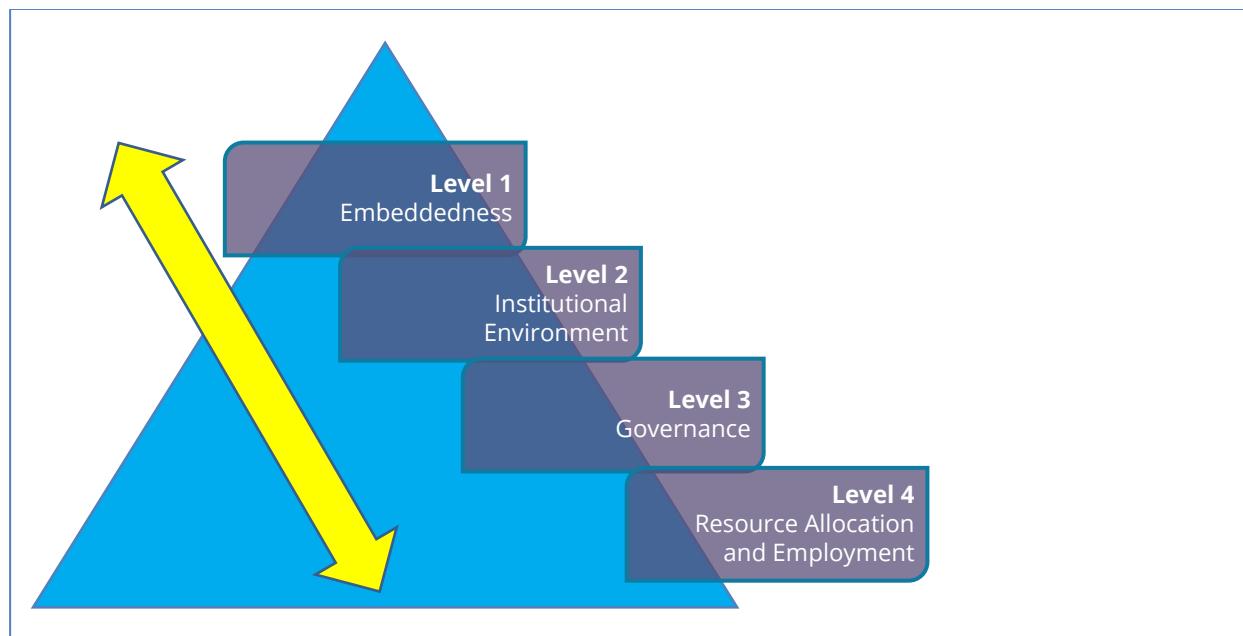
Figure 3.2 Four-Layer Model by Oliver Williamson

We will explain these categories below.

- ***Level 1: Embeddedness***
  Let us start with the upper layer, L1, which is called Embeddedness. This refers typically to informal institutions, customs, traditions, norms and religions. As mentioned earlier, fishermen who traditionally fish in certain areas, or citizens who are used to camping in certain places every year, do so (even) in the absence of specific rules and regulations that formally substantiate these rights. This is called "*Embeddedness*"

The frequency of change of these informal rules is typically very low. Williamson indicates that change occurs once in one hundred to one thousand years; this means these traditions are very deeply embedded in the behaviour of people (actors). The purpose, is often non-calculative. These informal rules evolve spontaneously and they are very difficult to plan. Equally, trying to influence these informal rules can be a difficult task. On the other hand, this layer is very important if we are looking to pursue responsible innovation.

- **Level 2: Institutional environment**
  Layer 2 is the *Institutional Environment.* These are the formal "rules of the game". Examples include the constitutions of sovereign states, or in this case, the energy policies on which the establishment of the offshore wind farms is based.

  The frequency of change at this layer is about once in ten to one hundred years. So, these formal rules are still quite stable. We can also identify specific objectives behind the institutional environment. One objective from an economic point of view might be creating the right institutional environment, which Williamson refers as first-order economizing. Once we have established which institutions would best serve the purpose of stimulating the development of offshore wind energy, we can design specific formal rules for establishing wind farms.

- **Level 3: Governance**
  Layer 3 is *Governance,* or the play of the game. Given the embeddedness and institutional environment, what kind of contracts or legal organizational forms can actors choose to get the governance right and to also realise their objectives? What kind of contracts and organizational structures best serve the individual objectives of firms and other actors? This kind of consideration is what Williamson calls second-order economizing. These governance structures change perhaps once in one to ten years, so the frequency of change is much shorter here than in the upper layers.

- **Level 4: Resource Allocation & Employment**
  Finally, Layer 4 is representative of a continuous change of rules and regulations called *Resource Allocation and Employment.* These are the daily routines of stakeholders and actors to get the marginal conditions right. These constant interactions individually and collectively shape how the institutions work in practice. This is referred to as third-order economizing.

Applying the Four-Layer model of Institutions

These are the four different layers or categories for institutions that we can identify. A very interesting aspect of these different layers is indicated in the scheme shown above by the top-down and bottom-up arrows. The yellow arrow in figure 3.3 means that the different layers should not be analysed in isolation.

For example, if there are certain rules or informal institutions in a country or in a region, it is important that these rules and informal institutions are protected by the institutional environment. This means that the formal rules of the game are to a certain degree based on the informal institutions. If this were not the case, we would have a serious problem, because the formal institutions would not be credible or relevant to that community. We need to align the informal institutions to the formal rules, and if we go further down this layered scheme, we can also argue that governance and resource allocation need to be aligned with each other.

So, these different layers of institutions are structured by a certain logic, and they need to be built up in a specific way, otherwise we would have disturbances. But note that there are also dotted arrows going bottom-up in the diagram. This indicates that there is also a reverse influence of the lower layers on upper layers. When the resource allocation or governance changes, this may influence formal rules and informal institutions.

If we consider the energy sector as an example, we see that currently a lot of attention is given to decentralised energy production and consumption, even down to the household level. Solar panels on the roofs of homes are not only used by households themselves, but the surpluses are fed back into the grid. We might consider that these initiatives on the local level warrant a change in the governance of the energy sector.

Similarly, any change of governance requires that the institutional environment is adapted to these new practices of producing and consuming energy, which in turn might also influence informal institutions. Households might consider that the production of electricity is no longer a utility which the state should provide, but something they can take care of by themselves. That would initiate a change in informal institutions and the values that are associated with the production and use of energy.

This interrelation between these different layers of institutions is very important if we are considering responsible innovation. It's not only a top-down activity, there is also a bottom-up development from individual users towards change of governance, the institutional environment and the embedded institutions, values and customs. In this respect, the institutional context is crucial to understanding and shaping the success or failure of new innovations.

Case study #3: Wind energy in the North Sea

Let us consider offshore wind energy, which can be considered an innovation of the electricity system. Below, we see a small map of the Dutch part of the North Sea.
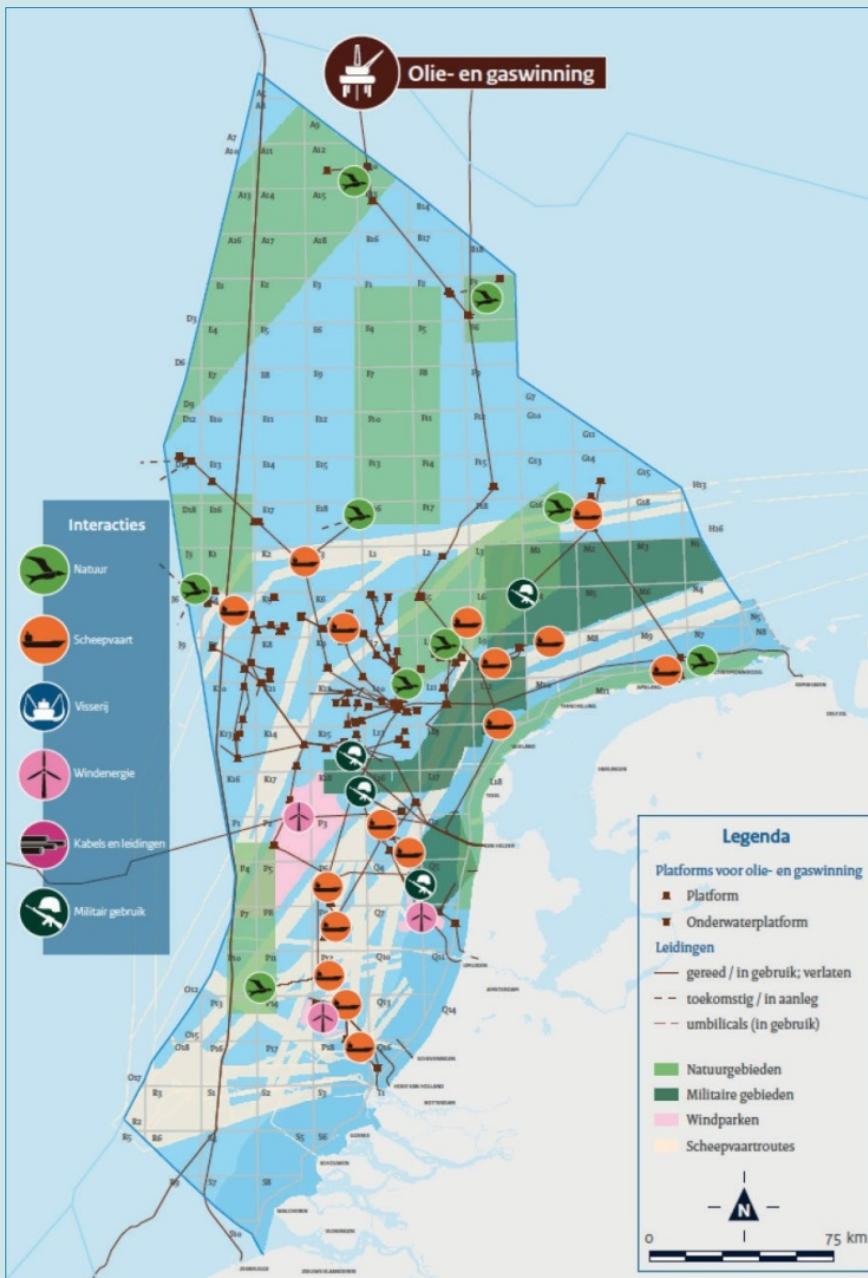

Figure 3-4: Map of the Dutch North Sea and its various uses

As illustrated by the different lines and dots, the North Sea is not just an empty space, but is used for very different purposes, like recreation, naval transport, fishery, cables or even military uses. Wind energy production is increasingly being introduced in that area. Since the North Sea has many different users, these users naturally need to adhere to certain social rules. This is simply a matter of practicality in most cases: it would not be useful to conduct military exercises in areas very close to naval transport routes or close to wind power farms.

Formal and informal rules

We need to dedicate certain areas for certain purposes in order to ensure that all these different stakeholders can make full use of the North Sea, while not infringing on the needs of the other users. In other words, we need social rules to designate how to use this space for different purposes. These social rules can either be explicitly established or embedded implicitly.

If we want to use the North Sea for the production of wind power, we need to expressly dedicate certain areas of the North Sea to the production of wind power. If we want to use the North Sea for military purposes, we need to explicitly assign
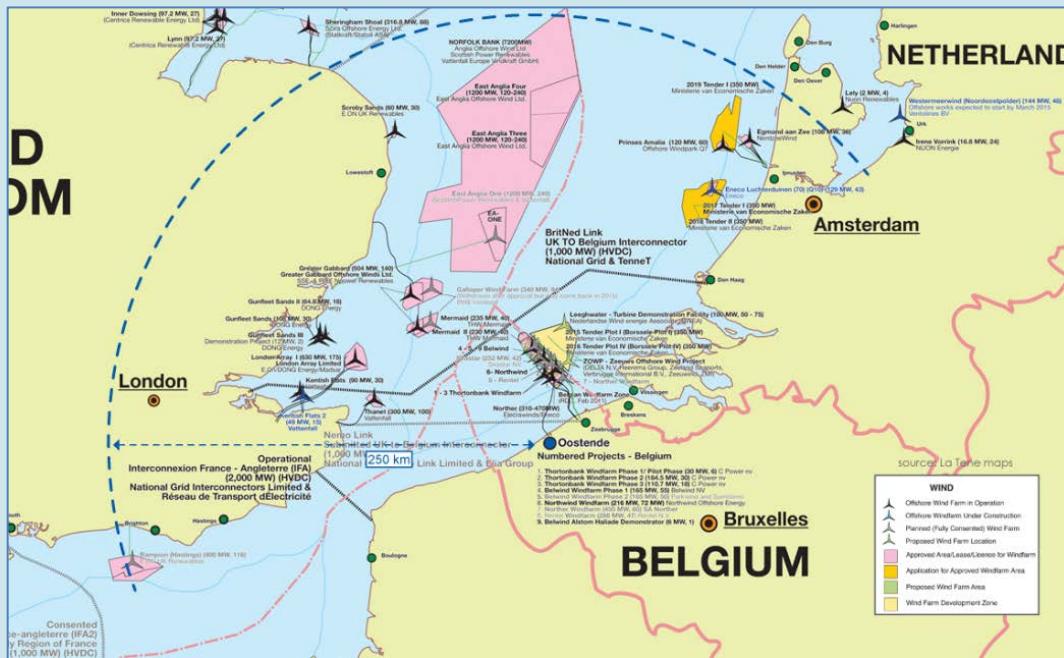
Figure 3-5: Map of planned wind parks in the North Sea

certain areas where the military may exercise. This means we need formal rules to dedicate these areas for specific purposes.

But there might also be unspoken social rules. Fishermen in the North Sea might have a long tradition - going back generations - of fishing in specific areas for herring or other fish. There may even be informal norms that certain areas are used for recreational purposes. Such social rules are not based on formal regulations, but instead they are embedded, that is to say based on traditions which people have held for long periods of time.

Building on this understanding of social rules, let's go back to the possibility of building offshore wind turbines in the North Sea. We might have to alter existing rules which are in favour of naval transport or fishery, and re-allocate certain areas for the generation of wind power. But it's not only about allocating a specific area for wind power, because other stakeholders might be impeded by these new rules; we need to understand how and rectify this, if possible. For example, if we assign certain areas for wind power, fishermen may no longer get good catches in these areas, naval transports need to find other routes and military exercises would be prohibited close to this area. This means we are not talking about single regulations, but looking at a system of different social rules, which in turn both prescribe and describe social behaviour.

Case study # 4: Self-Driving Vehicles

Over the last years, newspaper articles about Google developing technologies for self-driving vehicles were followed by heated social and scientific debates about such vehicles. The focus of the debate on autonomous vehicles (AVs) - which may be partially or completely automated - generally revolves around their impact on congestion, travel times and safety - that is to say, questions of utility. Ethical issues are discussed much less often. This section aims to ask some critical questions about the ethical issues surrounding AVs.

But what are the important ethical concerns about AVs? Below we will highlight some:

- **The problem of "many hands"**
  The first area of concern is a problem we have already encountered a few times: the problem of "many hands", resulting from several different actors playing parts of varying influence in the design and deployment of these vehicles. This raises concerns regarding the accountability of any one actor. Imagine an AV causes an accident due to a failed sensor, partly due to bad weather.

  Who is responsible? The driver? The car manufacturer? The company providing sensors? The company providing the key software? Perhaps the dealer doing vehicle maintenance? Or even the road authority which allowed these vehicles in the road despite the bad weather?

  Even if one actor is identified to be legally accountable, that does not mean the other actors are not involved, and their culpability is far from settled. Let us assume for the sake of argument that the human driver is held responsible. We can expect that a debate will inevitably arise about his/her responsibility, because the driver feels there was no wrongdoing on their part.

- **The "trolley problem"**
  We can also ask how even the most complex algorithmic intelligence might deal with the "trolley problem". What should the AV do if there is an oncoming vehicle on an impact trajectory, and the only options are to a) crash against that vehicle, endangering both drivers or b) make a sudden turn that will inevitably endanger a pedestrian nearby? AVs might very well face situations like these where a choice between two alternative accident scenarios needs to be made. Whatever the choice, that outcome would be based on the instructions it has been programmed with.

  From a **consequentialist's perspective**, hitting, or even killing, the pedestrian would be the preferred option, because only one person will be at risk, rather than two if the two cars were to crash into one another. But from a **Kantian perspective**, this may be not the preferred option. Of course, this is not to mention that pedestrians might change their behaviors in the (ubiquitous) presence of AVs.

- **Distribution of utility**
  A third issue is the potential trade-off between travel times, safety and sustainability. Any optimization of the system from one of these three perspectives may not result in equal outcomes. Taking a safety perspective for instance, it is preferable to maintain longer

distances between vehicles, but this arrangement would induce higher fuel consumption due to higher air resistance. Also, the utilized capacity of the road would not be optimal, possibly resulting in more congestion and longer travel times.

Or let us assume AVs can drive short distances at 160 km/h without any risks. This may result in shorter travel times, but at the same time increased CO_2 emissions. Or consider that AVs may in time become so convenient that they become preferable to public transportation even over long distances, potentially increasing emissions, but also indirectly inducing urban sprawl and increased land demand. Of course, we should not forget that trade-offs of this variety exist even now, in the current status quo.

- **Economic disparity**
  Fourth, there is the question of fair distribution when it comes to financial or economic considerations. At least initially, AVs will be more expensive than regular cars. Experts project that the cost of an AVs may be at least € 10,000 higher than that of comparable normal cars. We could argue that this is not a problem considering the benefits. However, if the road authorities were to allocate dedicated road space - say one lane of the highway - specifically for AVs, this would effectively reduce the available road capacity for normal vehicles, possibly leading to more congestion. Moreover, we could argue that since only affluent consumers would purchase expensive AVs, such a scheme would indirectly benefit only people in higher socioeconomic brackets, at the cost of those in lower brackets.

- **Decrease in demand for public transportation**
  AVs could make individual car ownership more attractive. The logic is as follows. One of the competitive advantages of public transport for individuals is that they can work, read or sleep while commuting, for example by train. On the other hand, personal cars provide the option of an on-demand mode of transportation. However, AVs can essentially provide the best of both worlds, serving both as personalized on-demand transportation and freeing the driver from actively having to drive.
  In the long term, there could be a large-scale shift from public transportation to AVs, in turn exacerbating issues such as congestion on highways, pollution, emissions etc. Moreover, the decline of demand for public transportation could hurt the population from lower socio-economic brackets disproportionately, since this is the group that depends the most on public transportation and moreover, cannot afford AVs in the first place.

But of course there can also be major advantages, like:

- **Increased accessibility**
  Firstly, in rural areas and in large urban sprawls, people without personal mobility options face de facto social exclusion due to poor access to education, economic opportunities, medical services and even social privileges such as maintaining contact with distant family and friends. Public transport may be too expensive in some cases, or simply not available. In cases like these, AVs - perhaps through schemes like car-sharing - could provide essential transit alternatives, reducing levels of economic, educational and social exclusion

- **Increased safety**
  In the long run, AVs could significantly improve safety not only for AV users, but also for non-users, such as pedestrians, (motor)cyclists and other drivers. Another overlooked area where we would see immediate and significant safety benefits would be the decrease in incidents due to driving under the influence of alcohol or other drugs.

- **Lower pollution and emissions**
  Overall energy use and emissions may also be reduced, perhaps directly, due to better design and more precise handling of AVs, and indirectly, as people shift to AVs that are known for their fuel-efficiency. After all, what sense is there in having a powerful but inefficient car?

Transition issues!

It is also important to pay attention to the transition period between traditional cars and AVs. Experts think that, immediately after the initial market introduction of AVs, the capacity of roads might decrease, as well as road safety. This could be expected due to 'growing pains', failures in technology, or a period of real-world learning. Eventually though, AVs would be more efficient, solve highway capacity concerns and improve safety. In other words, the initial years of AVs could decrease the performance of the transportation system, but due to learning effects and increasing market penetration of AVs, the system could improve over time, eventually exceeding

the performance of the status quo. In the meantime, of course, there will be an inter-temporal ethical issue.

Moreover, we could argue that the relative safety of AVs as compared to human drivers would introduce an interesting trade-off when it comes to car insurance. Over time, insurance premiums for AVs - if they consistently have fewer and less severe incidents than human drivers - could become lower. This would effectively make AVs a preferred investment in the long term, pricing out traditional cars on purely economic grounds.

Embracing cautious optimism

Perhaps the key point of this case study is to warn against overly conservative approaches to complexity and innovation, especially when it comes to technologies like AVs. For example, let's assume we have AVs only, and no traditional cars at all. If someone were to suggest introducing traditional human-driven cars (as we know them) and ban AVs, we could easily make a case against such a move, citing ethical consequences like lower safety of drivers, pedestrians and others, lower efficiency, higher degrees of exclusion for poorer groups in society, higher emissions and energy consumption, and so on.

As such, even as we debate the complex and thought-provoking ethical aspects of AVs, there is no reason to conclude *a priori* that the introduction of AVs is undesirable from an ethical perspective.

Figure 3-7 self-driving bus

Part IV: Management and innovation
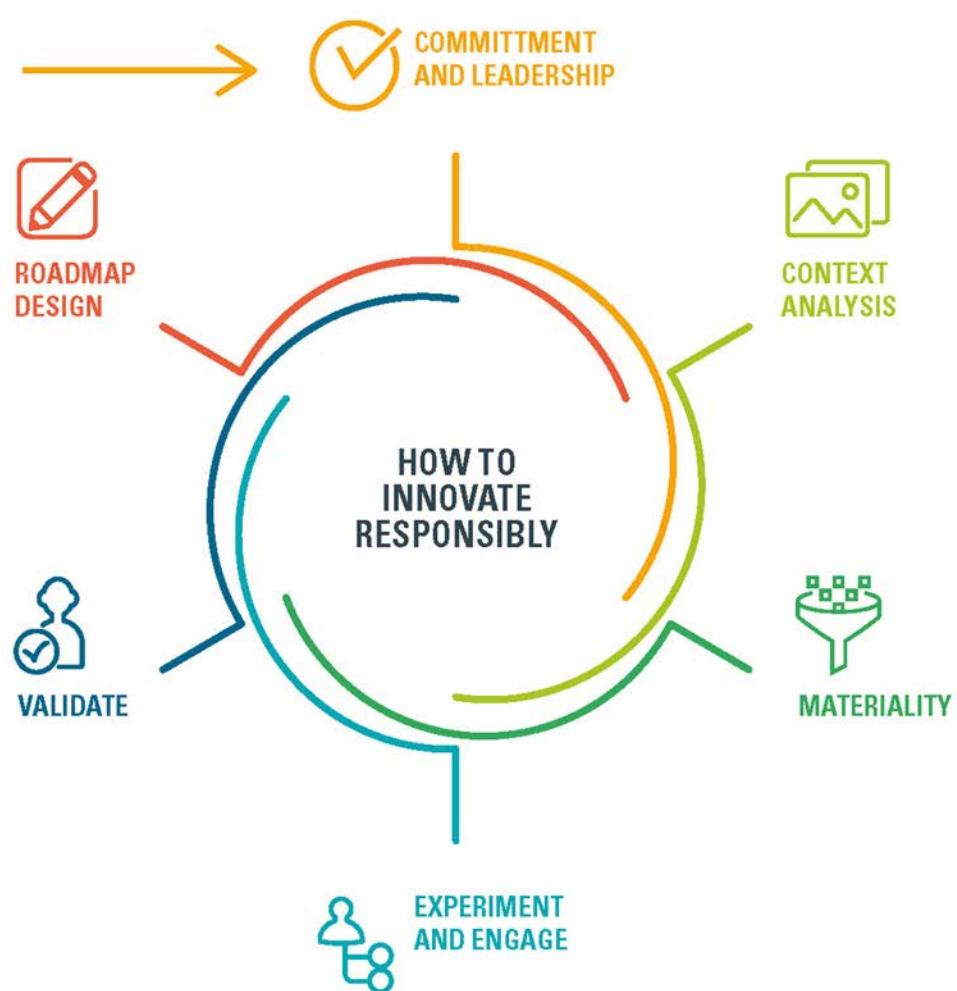
Figure 4.1 Process for companies to develop an innovation strategy based on RI, see chapter 6

4. Innovation and business

4.1 Incremental and radical innovation

Do you recall the definition of innovation we discussed in Chapter 1? We defined it as follows:

"Innovation is an activity or process which may lead to previously unknown designs, pertaining either to the physical world (such as buildings or infrastructure), the conceptual world (e.g. conceptual frameworks, mathematics and logic, software etc.), the institutional world (such as social and legal institutions, procedures and organizations), or combinations of these, which when implemented, expand the set of options we have to solve problems."

A taxonomy of innovation

Now let us focus specifically on technical innovation. There are different kinds of technical innovation. One often-made distinction is that between product and process innovation. A product innovation is an improvement in the product design. A process innovation, on the other hand, pertains to a change in the production process itself. A new feature on a mobile phone could be a product innovation, but a new type of machine to assemble mobile phones more efficiently would be a process innovation.

Another distinction that is often made is that between incremental and radical innovation. Innovations can be radical in a number of ways: they can be based on new operational principles; they can be based on new scientific knowledge; they can offer new functionalities; reach out to
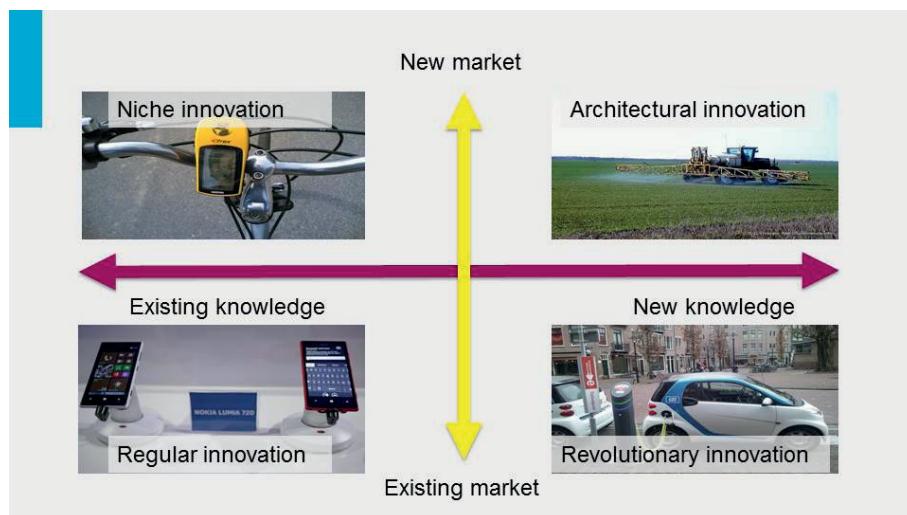
Figure 4-2: Taxonomy for innovation by Abernathy and Clark

new user groups; or may serve new types of values.

Here, we will rely on **taxonomy for innovation** developed by Abernathy and Clark in 1985. It classifies innovation along two axes. First, it asks whether the innovation is based on existing

knowledge or if it requires new knowledge. Second, it asks whether the innovation is intended for current users or for completely new users. Combining these two axes leads to the quadrants shown below.

Let us briefly discuss the different kinds of innovation.

- **Regular or incremental innovation** builds on existing knowledge and aims at existing customers. A typical example is a new model of a mobile phone, which are developed and updated each year.
- **Niche innovation** builds on existing knowledge, but reaches out to new customers or markets. A typical example is a GPS device especially for cyclists.
- **Revolutionary innovation** is aimed at existing customers, but based on new knowledge. A good example would be electric cars.
- **Architectural innovation** is based on new knowledge that opens up new markets and reaches new customers. Typical examples of architectural innovations are the Ford model T, the television, the first fighter jets, fertilizers, the internet, smart grids and cities, nanotechnology and so on.

The link between radical an responsible innovation

Architectural innovations have some identifiable characteristics. Firstly, they only occur once in a while. Secondly, they create a basis for a range of more incremental innovations. Thirdly, they are typically initiated by outsiders - that is to say, new companies or companies established in other domains - because they typically destroy existing knowledge and market relations. A prime example is Apple entering the mobile phone market with the iPhone.

Now, we can ask: does responsible innovation require radical innovation? To answer this question, let us refer again to the definition of responsible innovation given earlier.

Responsible innovation is innovation which - when implemented - expands the set of options available for solving a moral problem.

By this definition, all four types of innovation could possibly expand the set of options. All types of innovation can therefore be responsible. Nevertheless, responsible innovation will often require radical innovation. Why is this

To ensure responsible innovation, we need to take values into account in the design process. Often, we will need to take into account values that were not addressed before. Taking these new values into account often requires new knowledge. For instance, if we want to take the privacy of smart meters into account, we need knowledge about what privacy is. We also need knowledge about how to translate privacy into the design of smart meters.

Taking new values into account may also have an effect on the relation with users. It means an extension of the functional requirements met by the previous design. As a result, responsible innovation may mean opportunities to engage new markets and add new functionalities. This suggests that responsible innovation will often be similar to architectural innovation. To see whether this is really the case, more empirical research is needed. But if it were true, it would have some interesting implications. For example, market-leading companies would not always be best suited to introduce responsible innovations; instead, such initiatives may typically come from outsiders or newcomers.

Ethical considerations of radical innovations

This raises the question whether radical innovations introduce new ethical issues. We will argue that indeed they do. This is because radical innovation is not (just) about doing things in a new

way, but is about doing new things in general. Think of the internet, smart phones, air transportation or prenatal diagnostics. All of these technologies create new possibilities to act, and as such, they raise new ethical issues. For example, consider the privacy questions raised by the internet. Or think of prenatal diagnostics: suddenly, we have the ability to predict how likely it is that a child will have a certain syndrome, possibly one without a cure. This information raises completely new ethical questions for parents and for society. Parents see themselves faced with the choice to have an abortion or not. Society at large should discuss whether it is a problem that more foetuses with these syndromes are aborted – should we aim for eradicating these syndromes or not?

For existing technologies, there are often moral customs and rules. In case of incremental innovation, these rules and customs are usually still adequate. In case of radical innovation, the same rules and customs are often insufficient to address the new conditions.

A good example is the introduction of the jet engine in civil aviation. This was a radical innovation. Not very long after the jet engine had been introduced, two such aeroplanes - with the name Havilland Comet - crashed. The problem was not with the engines themselves, but the fact that jet-powered planes flew much higher than other planes at the time. Therefore, the cabin had to be pressurized to make flying comfortable for passengers. As a result, some points of the fuselage were subject to greater stresses than before, which led to metal fatigue and ultimately to disaster.

This brings us to a cautionary statement about radical innovations: responsible innovation often requires radical innovations, that in turn often raise new ethical issues.

4.2 Determinants of innovation

Building on the knowledge of what innovations are, let us now examine what factors determine whether a particular technological innovation is successful. In order to do so, we first have to say something about who the actors that innovate are and what their motivations are. Next, we have to understand how we can scale up innovations; thirdly, we should focus on determinants or incentives that influence the innovation performance of private, profit-oriented firms.

Innovative actors and their motivations
So, who are the actors that innovate? As innovation is a human activity, the straight answer should be: the individual inventor. The individual inventor is a creative person who is stimulated by intrinsic motivation - which is to say, his or her drive to innovate is a personal interest in specific technological problems. Combined with their personal ability or creativity, they solve technological problems with new approaches and/or answers. One example of such an individual inventor is Thomas Edison, who invented the light bulb. Another example is Rudolf Diesel, who invented the diesel engine.

In order to scale up the production of innovations, it is necessary to put a number of creative people together in an organisation and structure the whole innovation process in such a way that their creativity can be used in an optimal way. The advantage of this is that organisations generally have more resources than any one individual inventor and, therefore, organisations can be used to stimulate or scale up innovations. Examples are public organisations, such as universities like *Delft University of Technology*; but also consider that most innovations take place in private, for-profit firms, such as Apple, IBM and Philips, as well as smaller and less well-known firms, especially start-ups.

Economic determinants of innovation

Let's now look at the economic determinants of innovations in private profit- oriented firms. The process of innovation is highly uncertain. Translating a creative idea into something novel and useful can be very costly, while the benefits are highly uncertain. This originates from the fact that novel ideas should first prove themselves useful, before customers will start to buy them.

Many studies show that the chance of translating a new original idea into a successful commercial product is less than 0.1 %. This means that only one in more than one thousand new ideas actually becomes a successful commercial product. In other words, it is highly uncertain whether innovative activity leads to higher profits. A number of determinants of successful technological innovations can be identified. We can start with the external factors: these are factors outside the firm, such as the technical, economic and legal environment. On the other hand are the internal factors, which play a big role inside the firm.

A first important factor is the technical environment in which a firm operates. This is the industrial sector to which the firm belongs. For example, a firm in the aerospace industry operates in a dynamic technological environment, in which staying ahead of your competitors is more intense and necessary than in other sectors, such as the textile industry.

This brings us to a second important external factor: the economic environment as described by competition or market structure. The Austrian-American economist Joseph Schumpeter was one of the first scholars who investigated the impact of market structure on innovation. His central question was: in which kind of markets would firms achieve the highest innovation performance? Two competing explanations exist.

- First, innovations will mainly be generated in markets with many intensively competing small enterprises, that are forced to innovate in order to stay ahead of their competitors.
- The second explanation claims that markets with less competition will see more innovation. The reason is that big firms such as Philips, Unilever, etc. have many resources available, so that they can be involved in uncertain innovation processes without immediately going bankrupt when an innovation fails.

Empirical studies are rather inconclusive. It seems that the technical environment is an important factor. For example, the present-day software industry is dominated by big firms such as Google, Apple, Facebook and Microsoft. These giants in the software industry have many resources available for innovation, but this does not necessarily mean they have it easy, nor are their continued profits guaranteed. There is vigorous competition in this technologically fast-changing environment.

In another example, the oil and gas sector is also dominated by major players, such as Shell, ExxonMobil etc. Although firms in this sector do innovate, the competition between them is much less vigorous, because the technological environment in which they operate is changing at a slower pace than in the software industry.

Collaboration is a third determinant of innovation. In the last twenty-five years technological innovations have become increasingly complex, fast-changing and much more international than before. The consequence of this development has been that it becomes harder and much more costly for any individual firm to innovate successfully. In order to find sufficient new ideas, firms have to go beyond their own borders and collaborate with other actors, such as suppliers, customers and universities, in order to increase their innovation performance. This could be a collaboration in a so-called technology cluster, which is an arrangement where firms sharing a common technology (for example, software or biotechnology) engage in buyer, supplier and complementary relationships for production; these firms also do collaborative research. The

reason this works is that complex knowledge is often tacit - which means that it is not always explicitly documented, but resides is in the heads and (unspoken) actions of the engineers or developers in that area. This requires frequent and close interaction among people.

A good example of this kind of collaboration is the concentration of semi-conductor and software firms in Silicon Valley. Software developers of different firms in Silicon Valley regularly meet each other, share a few drinks and exchange views, based on their knowledge. Other examples are meetings of watchmakers in Switzerland or fashion designers in Milan.'

Being in a technology cluster has a number of advantages. First, it can lead to technology spill-overs; thus, the benefits of the R&D of an innovating firm spill over its boundaries and increase the benefits of another innovating firm. A second advantage is that the local labour force is technologically well-educated. Yet another advantage is that suppliers and distributors, as well as other supporting firms, such as accountants, lawyers etc., will be at hand, which increases the chance of commercial success for the innovations.

A fourth external determinant of innovation is the *legal environment in which the innovating firm is operating.* Innovative persons or firms often spend a lot of money on developing ideas and
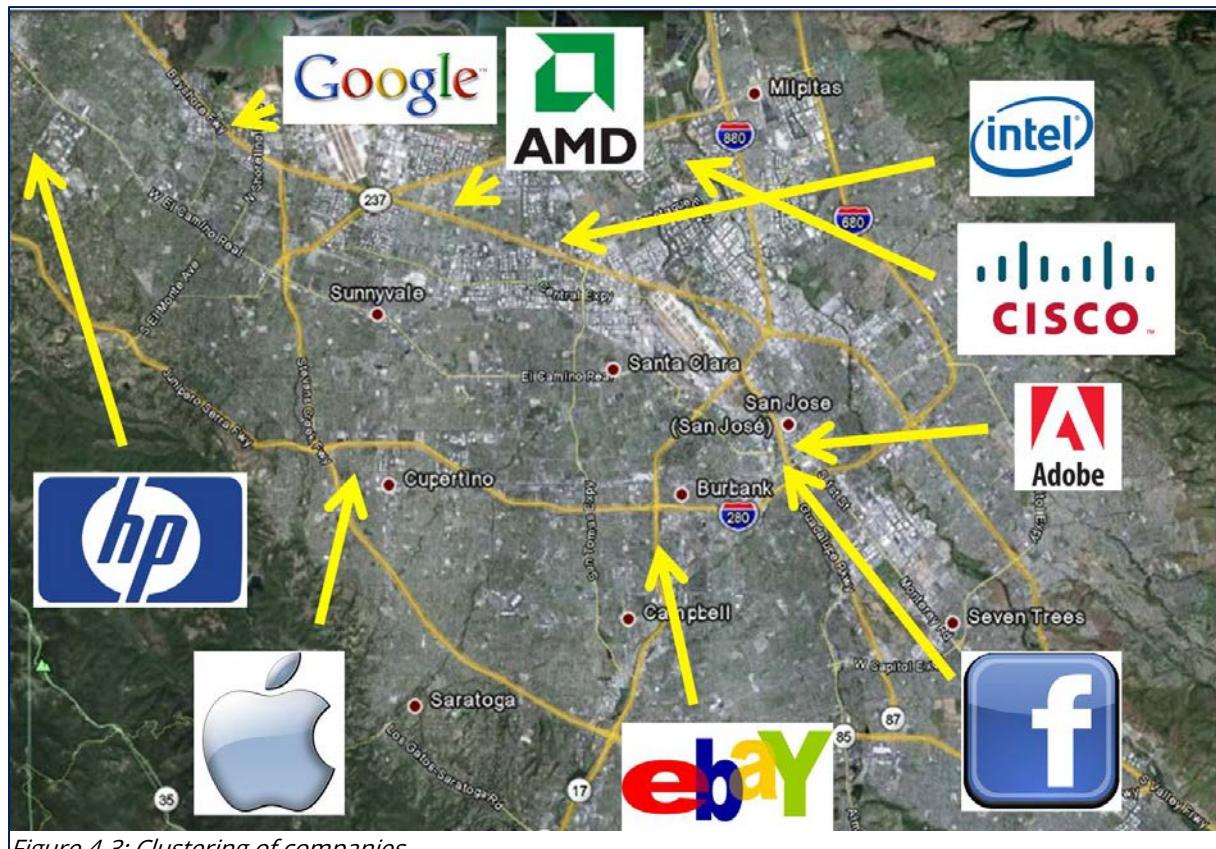
Figure 4-3: Clustering of companies

transforming them into concrete applications. The high development costs should be earned back after the launch of the product. At that moment, there is a risk that other individuals or firms with the right technological knowledge may perform what is known as reverse-engineering. This means that they investigate how the novel products or systems work, figure out what they consist of, and examine their .

Then they produce and sell their own version of the innovative product or system, but without incurring the high development costs that the original inventor experienced. If that happens, the profit and therefore the pay-back opportunities of the original inventor decline substantially.

Intellectual property rights primarily exist in order to avoid this. An example of intellectual property rights are patents, as designed by patent laws. The original inventor gets protection for a period of twenty years, during which his innovation cannot be produced or sold by someone else, unless the original inventor is financially compensated through so-called licenses. In effect, the original inventor receives a temporary monopoly, which guarantees him the possibility of a stream of revenues to earn back his enormous development costs. The incentive for governments to provide legal protection through patents is that they want to encourage original inventors to continue being innovative, thus driving industries and the economy forward.

Finally, the fifth determinant, the main internal factor, has to do with the *organization of the innovating firm itself*. The American economist **William Baumol** emphasized that the power of innovating firms in a capitalist society lies in the routinization of innovations. In order to routinize innovations, firms that are continuously looking for novel ideas and transforming them into useful products or systems, have to work within standard procedures. However, procedures and creativity are two opposing forces. Creativity is intrinsically unpredictable, whereas procedures are developed to make innovative outcomes less unpredictable.

That means that a centralized hierarchical organisation, with strict top-down management in which total control is considered key, is not the best organization to stimulate the production of innovations. On the other hand, a fully decentralized organization, in which each employee is creative and can do what he or she wants, is also not workable.

Hence, ambidextrous organisations are often considered the best of both worlds. These are organisations in which different units can have different organisation structures. For example, decentralized units aimed at generating novel ideas co-exist with more centralized units that try to translate the best ideas into concrete products and sell them successfully.

Let's now focus on one kind of organisation with a typical institutional profile: a company. Companies generally operate in a larger institutional environment, in highly competitive settings, in which they try to be successful, i.e. profitable. In order to manage innovation, while acting as an entrepreneur at the same time, it is necessary to understand both the creation of innovation and subsequent diffusion process in companies. The way innovation has to be managed, and the way entrepreneurship is involved the process, strongly depends on our perspective of these processes.

4.3 Management of innovation

Management of innovation in companies

We have seen earlier how the different types of innovation are distinguished. For now, we will particularly focus on product innovations, i.e. innovations based on new technology. Examples of technology-based product innovations are communication appliances such as telephones and television, materials such as Kevlar or Glare, and medicines such as Prozac or Aspirin. At the time of their introduction, these were radically new technology-based innovations. How can a company profit from with this kind of discovery? We can distinguish two different perspectives on innovation and diffusion processes.

Innovation as a simple project

In the first perspective, the whole innovation process is seen as a project - a new product development project - that starts when a new technology becomes available. The project ends when the new product is ready for production and distribution and its marketing is prepared. Subsequently, there is the market introduction phase, which is when the diffusion process starts.

In this perspective, different types of management are required to complete the process successfully. For example, R&D management is required to develop the technology and R&D continues to be involved in the subsequent product development phase. Project management is
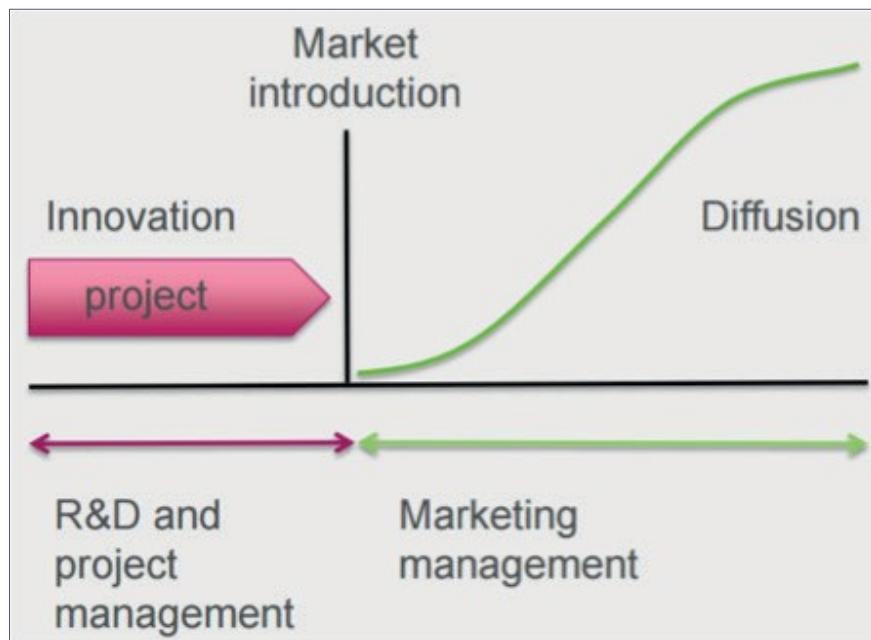
Figure 4-4: Innovation as a simple project

required to manage the new product development trajectory. Marketing strategies are required to prepare a market introduction plan and manage the subsequent diffusion process, as can be seen in the figure below.

From the 1980s onwards, mainstream innovation management handbooks presented innovation as a project. From this perspective, the management efforts require close interaction between marketing and R&D teams. Which of the two competencies is in the lead depends on the type of company and the particular industry; the leading department usually provides the main project manager.

The success of these joint efforts is reflected in a large increase of sales or in a steep diffusion curve after introduction. If you track the diffusion of telephones and televisions, for example, you will find an almost perfectly shaped diffusion curve. Indeed, these products were quite successful in the market, and this perspective was rather astute in describing the trends and prescribing steps for managing such innovations. In innovation management literature, we find a considerable body of knowledge on the marketing/R&D interface. This line of thinking continued until the turn of the 21st century.

Innovation as a complex process

In the second perspective, the innovation process is not seen as just another new product development project. The process is rather more complex. We can distinguish four aspects that make innovation more complex than a product development project.

- Firstly, technology development and product development usually proceed in parallel. Usually the first products are unreliable and the technology needs to be developed further in order to enable the development of reliable products. Jointly developing a product and the required technology is not just a single project, but more of a complex program of highly inter-related - and therefore iterative - projects.
- Secondly, many companies, or networks of companies, compete with each other by working in parallel on technology and product development. Sometimes these findings are patented and subsequently used by other consortia or networks of companies. In that case, the innovation process is not just a project; in an era of open innovation, innovation is an inter-linked process of many separate projects.
- A third reason why the innovation process is not just a project is that a product cannot be introduced out of the blue. The initial market usually lacks complementary products and services (e.g. an infrastructure), or cost-effective production facilities are not yet available. Materials such as nylon, and strong fibres such as *Dyneema*, were developed long before their large-scale production was even possible. Sometimes consumers do not really understand the product. As a result of all these elements that might be lacking, market introduction becomes a trial-and-error process in which development proceeds. even as the product is already diffusing in the market. Thus, market development and product/technology development proceed in parallel.
- Lastly, in some cases the basic underlying scientific principles behind a technology only become clear long after the technology has been successfully deployed for years. Sailboats for example were built for thousands of years before we understood the scientific principles that enable their movement. Airplanes too were used for years before we understood and really mastered the principles of flight.

The modern innovation process

In the updated perspective of the innovation and diffusion process, various types of management are required to complete the process successfully. On top of R&D management, project management and marketing management competencies, new types of competencies are also required.
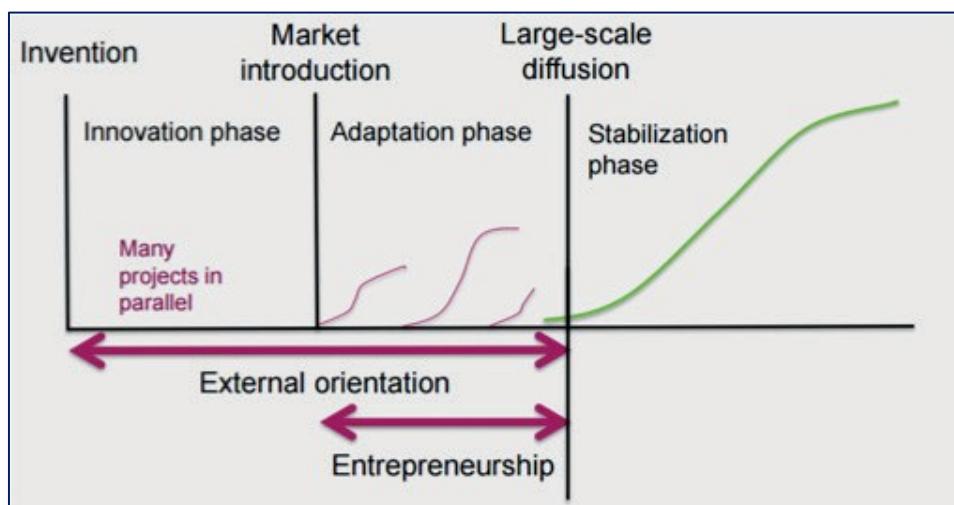

Figure 4.5: The modern innovation phase

Firstly, companies need to adopt an external focus, in order to align the development of their products with the development of related and/or complementary products and services by partner companies. This external orientation is also required in order to track the technology and product development activities of rival companies and to track the latest market developments.

Secondly, entrepreneurial competencies are required to develop a market. The second phase - the adaptation phase - is usually when many companies go bankrupt or leave the market before they crash and burn. It is essential for entrepreneurs to have a vision for their product, combining the technological functionality with a (latent) market need and designing a product that fulfils that need. They also must create a sustainable business model to commercialize the product. Entrepreneurs can also decide the best timing to introduce the innovation and where to introduce it; in other words, they also provide the niche strategy.

This has some implications for responsible innovation. In the adaptation phase (the experimental and entrepreneurial phase, in which several product versions can be introduced in multiple market niches) accidents can happen and unexpected side-effects may emerge. A responsible approach is required here.

The mainstream application in the stabilization phase is sometimes hard to predict, and so are the consequences of the use of the product in this application. Again, responsibility is necessary to manage the potential trade-off between profits and consequences.

Case study #5: The development and diffusion of television

Let us consider a typical historical case, frame it as an innovation and diffusion process and try to conclude what this implies for the types of management and entrepreneurship that are required to complete these processes successfully.

The invention of the principal technology behind television can be dated as early as 1925-1930. However, product development did not start immediately; it took almost a decade before the first televisions were introduced. Apparently, it takes a couple of years (a decade on average) to turn an invention into the first product. This is the first so-called innovation phase in the larger process.

When the television was first introduced, it was not the appliance that we know today. At first, televisions were introduced in Germany and the UK circa 1939 as a kind of semi-public service for bars. Instead of a large-scale diffusion after market introduction, only small-scale diffusion in specific niches could be seen. A similar pattern of diffusion can be found for almost all new radical high-tech product innovations. This phase of initial small-scale diffusion of the different product versions in small market niches is referred to as the adaptation phase.

Only from the 1950s onwards (more 20 years after the original invention) the large-scale diffusion of televisions began in earnest. This last phase is referred to as the stabilization phase.

From analysing the innovation and diffusion of more than a hundred cases of radical high-tech products introduced between the year 1850 and 2000, we can conclude that the television is in fact a typical and average example in terms of the time between invention and large-scale diffusion.

Case Study #6: Coolants

The history of cooling is almost as old as the history of humanity. In prehistoric times, men heaped ice in deep caves to keep food cold, while in the Middle Ages niches in wooden walls had a cooling effect. Cellars and cold rooms were ways to store food in wooden barrels or clay pots. From the mid-19th century onwards, natural substances such as ammonia, carbon dioxide, isobutene and others were used as refrigerants for industrial and commercial applications. With the discovery of chlorofluorocarbons (CFCs) in 1929/1930, a new age began, as CFCs were soon also used as refrigerants. CFCs reached significant market penetration worldwide; in the 1960s household refrigerators and shortly after home freezers became common goods in Europe´s homes.
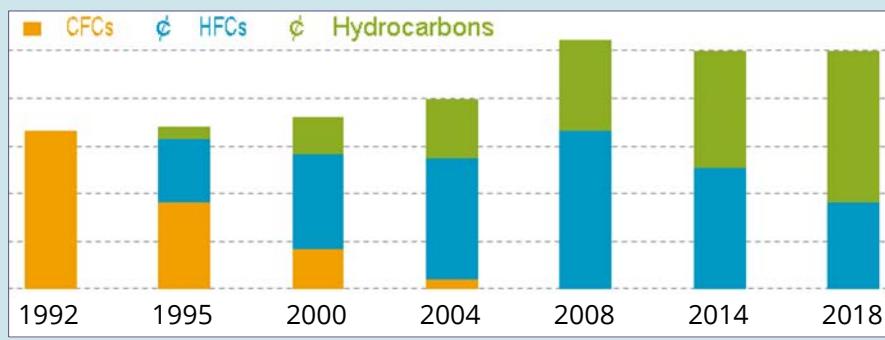
Figure 4-6: Global market penetration of GreenFreeze technology

It wasn't until 1973 that the destructive impact of CFCs on the atmospheric ozone layer was detected and further investigated. However, scientific evidence that linked ozone depletion and emissions of CFCs was denied and it took another 15 years until a political agreement on the phase-out of production and consumption of ozone depleting substances (ODS) through international action was reached in the Ozone Protocols of Vienna (1985), Montreal (1987), London (1990) and Copenhagen (1992).

Mainly the **Montreal Protocol**, which entered into force on 1 January 1989, pushed industry for research. Hand in hand with political negotiations, the chemical industry started working on the identification of alternative refrigerants. Hydrofluorocarbons (HFCs), in particular HFC-134a, were soon identified and tested as the most promising CFC substitutes during the late 1980s, since they did not harm the ozone layer. There was a general conviction that HFC 134a was going to be the new universal coolant. However, HFC-134a still had a huge global warming potential (GWP). It is actually a powerful greenhouse gas that contributes significantly to climate change.

Thankfully, Greenpeace in the nineties successfully initiated an alternative coolant known as "GreenFreeze". This alternative has now been adopted in many countries and by many companies. GreenFreeze uses hydrocarbons, entirely free of ozone-depleting and global warming chemicals. This new technology rapidly spread to

other European countries and soon GreenFreeze revolutionised the world-wide refrigerator industry. Today there are over 700 million refrigerators using this "Made in Greenpeace" technology.

This success did not come naturally because - although isobutane contributes far less to the greenhouse effect - it is flammable. This was the main reason why the fridge industry opposed it initially. In fact, existing technical codes even banned the use of flammable coolants like isobutane.

Still, GreenFreeze became accepted because of a number of circumstances as identified by **GIZ, Germany**:
  • A new CEN standard was adopted quickly, allowing for the use of hydrocarbons in domestic refrigerators. This had to do with the re-interpretation of safety as a value.
  • At first, the safety of GreenFreeze was interpreted in terms of the inflammability of this type of coolant. Later, it was understood in terms of the ignition and explosion risk of a fridge. Flammable coolants turned out to be not so dangerous as generally thought. One reason was that nowadays fridges contain only small amounts of coolants. (Note: When the original standards for inflammable coolants were formulated, fridges still contained much more coolant, because they had a much lower efficiency.
  • The demand for household appliances was forecasted to grow globally, as positive economic growth in many countries was fuelling domestic investments;
  • The technical know-how of the engineers, paired with the willingness of management to take the risk and convert to the brand-new, largely unproven hydrocarbon technology;
  • Early on, ministries and government agencies supported research and introduction of hydrocarbon technology. Political measures, such as information and awareness raising (conferences and studies) and the award of the Blue Angel ecolabel facilitated the market uptake of the GreenFreeze technology;
  • The global diffusion was further promoted by the large-scale German governmental engagement
  • In the early 1990s, consumers had a high level of environmental awareness and were generally well informed about the depletion of the ozone layer. They therefore were willing to accept eco-friendly appliances and to contribute to environmental protection;
  • Greenpeace linked up with the relevant stakeholders.

5   Frugal innovation

5.1 What is frugal innovation

We have seen how companies deal with innovation in order to capitalize on new technologies, so that they can enter new markets and make more profits. Now, let us look at a specific form of innovation associated with global development: frugal innovation. Frugal innovation is a new global phenomenon; in order to understand it, let's look at the dictionary definition of 'frugal'.
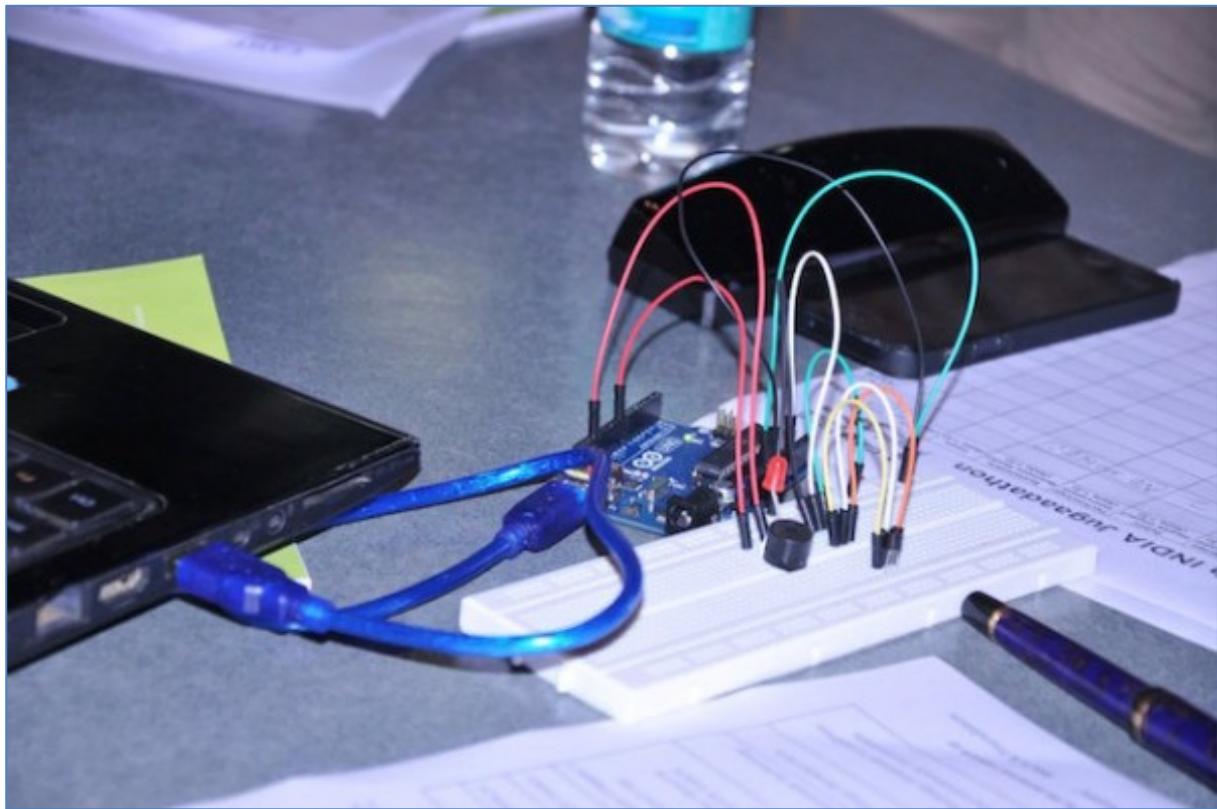
*Frugal is **defined as** "economical in use or expenditure; prudently saving or sparing; not wasteful; entailing little expense; or requiring few resources". Note that frugal does not mean a poor-quality, off-the-mark, improvised solution; it's not just about making existing products cheaper. Instead, frugal innovation is innovation aimed specifically at serving the needs of some of the world's poorest people.*

Frugal innovation is a new phenomenon in global development. It is usually defined as stripping down and/or re-engineering products and services, thus reducing complexity and costs, to offer quality goods at very low prices to the people in who are at the "Bottom of Pyramid" (BoP) - i.e. the almost four billion people in the world, who have to live on less than US$2 a day.
A recent comparison of product prices has shown that frugal innovations can lower the price of a product from between 50% to 97% according to Rao in his article **'How disruptive is frugal'** .

From an economic perspective, frugal products and services seek to minimize the use of material and financial resources in the entire value chain with the objective of substantially reducing not just the price point, but the complete cost of ownership and usage of a product; all that while fulfilling or even exceeding pre-defined criteria of acceptable quality standards. Equally, from a functional perspective, frugal innovations often - considering the clients - should be able to cope with trying everyday conditions, like dust, heat or power failure. Therefore, the design - and the mind-set of the designers - has to take this into consideration. Design has to serve users who face extreme affordability constraints, in a scalable and sustainable manner.

Generally, we can distinguish two versions of frugal innovations. In the first type, an existing product, service or system is stripped from its luxury attributes, while its basic technical functionalities remain intact in order to guarantee it will function optimally. Without these luxury attributes, prices go down dramatically and hence the product or system is affordable even for low-income groups in developing countries. An example is the Nokia 1100 cell phone. This phone was targeted at low-income users in developing countries who do not require advanced features beyond making calls and sending SMS text messages.

Figure 5-1: Frugal innovation hackathon in India. Note that Frugal innovation is there known as Jugaad

The second version of frugal innovation means creating new products or systems that originate from demand by potential customers. An example is a frugal thermometer, developed by the Centre of Frugal Innovations in Africa (CFiA) - a collaboration between Leiden University, Delft University of Technology and Erasmus University Rotterdam. The thermometer has specific characteristics, zo that even illiterate people can use it in a responsible way. For example, the body temperature can be measured by a scan of the forehead. The temperature can be read in colours: red means 'go to the physician', green means 'everything is fine'.

5.2 The case for frugal innovations

There are some misconceptions with regard to the development and production of frugal innovations. In traditional innovation and strategic management literature, the focus is primarily on studying the determinants and impact of innovations in high-income markets. In these markets, high profits per unit - or margins - can be pursued, presenting a profitable opportunity for companies.

The general view is that low-income groups cannot generate comparable or even substantial profit opportunities. This notion is correct when talking about innovations specifically tailored for high-income markets. But it is not correct when we speak of frugal innovations.

Still, many Western multi-national firms show a number of strategic misconceptions with regard to development and production of frugal innovations:

- They believe that limited purchasing power of "Bottom of Pyramid" consumers cannot be translated into profitable opportunities due to low prices.
- They think that there is no room for high-technology firms in BoP markets, as customers in these markets use simple products that are produced with low-technology production processes.
- They are afraid that serving BoP markets would be seen as exploitation of the poor.

The world has changed in the last twenty-five years due to globalization and liberalization of international trade and capital flows. Particularly the high growth rates in emerging markets in the past two and a half decades have led to a new bracket of customers that exerts new demands. In these countries, the number of people belonging to the middle classes is increasing. At the same time, there are still some four billion potential BoP customers. This is more than half of the world's population, mostly living in developing countries, particularly in Africa.

Multinational firms can contribute to economic and social development in developing countries by serving the hugely untapped potential of BoP customers, which would create opportunities for profit-making as well as to economic and social development. One channel along which economic development can be stimulated is that frugal innovations for these customers have the potential to change disorganized and fragmented local markets into an organized private sector market where products can be supplied at much lower costs than currently.

One example of multinational firms providing frugal innovations to the BoP markets is General Electric Healthcare, which has produced an electrocardiogram for use in rural areas in India. This is an environment characterized by lack of electricity, scarcity of trained medical personnel and poverty. The ECG costs about US$1000, which is just a tenth of the price of ECGs developed for the US market.

5.3 The link between frugal innovation and responsible innovation

There are two elements of responsibility when speaking of frugal innovation. Firstly, the double-digit growth rates of emerging economies in Asia, Africa and Latin America increase the desire for a higher standard of living. If new customers in the expanding middle classes would consume the same kind of products as customers in high-income countries do, the pressure on the world's natural resources would inevitably increase. This cannot be sustained indefinitely. Therefore, regulations encouraging more sustainable development are being implemented by governments and international organisations.

At the same time, customers in the lower and lower middle classes, and particularly those in the BoP, can only afford frugal products at low prices. Here we see a tension between the cost of increasing social and sustainable regulations (raising costs) and the immediate needs of the world's poor. This requires design processes that are different than those we are used to in high-income countries.

Figure 5-2: An optical instrument which allows anyone, not just experts, to diagnose malaria at a very early stage, using a simple pin-prick blood sample.
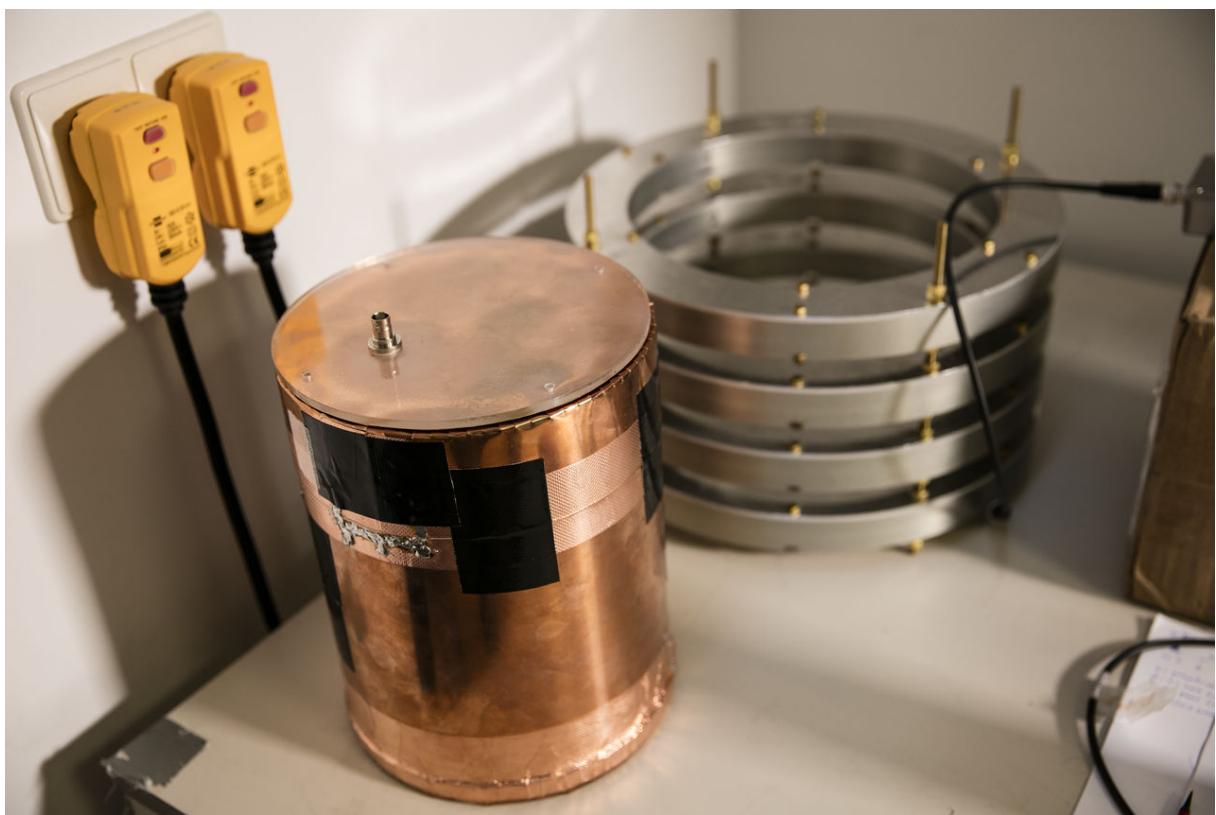
Figure 5-3: Elements of a frugal MRI.

The second element of responsibility has to do with the business model that can provide a link

between profits and local economic development. Traditional product management has a product-centric approach. In the case of frugal innovations, a completely new business ecosystem should be designed. Such an ecosystem means that the innovating firm has to collaborate with external partners, such as governments and NGOs, but also local entrepreneurs.

Local entrepreneurs especially can be very important for two reasons. Firstly, they can be a clear distribution channel for frugal innovations, particularly relevant for BoP customers living in remote rural areas. Secondly, they are much closer to the cultural and local preferences of potential customers, and hence provide an important input in the early phases of the design process at the innovating firm. Note that BoP customers may live in extreme resource-constrained environments. Working around these constraints offers an opportunity for new frugal innovations, which could lead to a responsible, "inclusive" contribution to local economic and social development. Local entrepreneurs can be a very relevant medium to transfer relevant knowledge of this environment to the innovating firm.

5.4 Innovation and social standards

Frugal innovations are not automatically responsible innovations. We also have to pay attention to the issue of social standards, and see how they co-determine when frugal innovations are also responsible innovations. Here, we will argue that frugal innovations are not responsible innovations when the social standards applied in the production processes are either too low or too high.

Social standards, sometimes also called labour standards, have two main elements:.
- The first is ensuring decent working conditions for labourers, such as a reasonable minimum wage and proper health and safety precautions. In a factory context, this includes simple things like making sure the fire extinguishers actually work. On farms, this responsibility could mean providing protective clothing for workers dealing with chemicals.
- Secondly, we have to ensure labourers have so-called "enabling rights", which includes freedom of association and the right to collective bargaining. A social or labour standard needs to ensure that workers can form an association, and that union leaders are not simply fired, or worse, mistreated by employers.

This are easier said than done, since a key complicating characteristic of social standards is that we usually cannot "see" them. Take for example child labour. We cannot deduce from looking at a T-shirt or drinking a cup of coffee whether the processing of that product involved child labour. We call this a "credence" good, which means that we need to put our trust in those who monitor these production processes.

This monitoring can be done by government agencies like a labour inspection organization, and sometimes this is done by companies who wish to operate at a higher level of responsibility, by monitoring their own social or ethical standard. It can also be done by non-governmental organisations (NGOs) like FairTrade, who monitor co-operatives of small farmers to ensure they are not using child labour, for instance.

There are of course additional challenges associated with monitoring as well. While some governments more effectively monitor social standards compared than others, all governments

face challenges with regard to production processes in the informal economy, where many relatively poorer consumers buy most of their products.

How social standards impact frugal innovation

After introducing the idea and practice of social standards, let us return to the main argument of this section: frugal innovations are not always responsible innovations. Social standards play an important role in explaining this point. When frugal innovations are based on 'stripping' existing higher-value products, one of the first things that producers may sacrifice are social standards, like minimum wages for workers, or they may cut back on health and safety considerations in order to reduce costs. For example, frugal innovations produced in the informal economy may not protect workers against exploitative working conditions. In such situations, where social standards become too low, frugal innovations cannot be seen as responsible innovations.

Of course, this is not a simple yes or no issue, but a matter of trying to ensure as decent working conditions as possible. What this means is that we cannot only look at the technological or ethical dimensions of the product as such, but also need to consider under what social conditions these frugal innovations are produced.

Unfortunately, there is also the possibility that social standards are too high. Many examples exist of large firms that successfully lobby with national governments and international agencies to create entry barriers for new firms. These incumbents try to protect their vested interests and block new firms with new ideas from entering the market, using (among other means) their higher social standards as an argument to protect their dominant market position. This type of protectionism is heavily criticised by firms from emerging economies, who find it difficult to get access to European and US markets.

In principle, higher social standards are a good thing; after all, who could be against higher wages or better health and safety conditions? But it becomes a different matter altogether when established large firms can use such standards to create the impression that their way of doing things is the only legitimate way of doing business, effectively creating barriers to entry for new firms and thus obstructing innovation. This also means that too high a bar for social standards may hamper frugal innovations, as it obstructs innovation, especially types of innovation that try to significantly reduce costs without sacrificing user value.

Caveats for frugal innovation

To conclude, frugal innovations are not responsible innovations, when social standards are set either too low or too high. When social standards are set too low, this can easily lead to exploitation of workers and therefore to irresponsible innovation and production processes.

But social standards can also be set too high. In that scenario, large, established firms use the argument of unnecessarily high social standards to block the entry of new firms into their markets - and not for ethical reasons.

This means that frugal innovations are more likely to contribute to inclusive development when social standards are set as high as possible in order to ensure decent working conditions (and not higher), and as low as necessary to allow for new innovation opportunities (but not lower).

Another dimension that frugal innovations have to satisfy in order to qualify as responsible innovations is whether they have the potential to include poor consumers and producers in the

ensuing economic growth and development. Here, we will explore this issue and how to achieve systematic, inclusive growth.

5.4 Innovation and inclusive development

The need for inclusive development

What do we mean by inclusive economic growth and development? Inclusive growth means that there are sufficient opportunities for everyone to participate in the growth process, and at the same time making sure that benefits are shared across the community. To be inclusive, growth should benefit everyone, while reducing the disadvantages faced by the poor, both in terms of benefits enjoyed and especially in terms of access to opportunities for participation.

Today, the majority of people living on less than US$ 2 day live in two regions: Southern Asia and Sub-Saharan Africa. Nearly two thirds of these people, that is, the extreme poor, can be found in five countries: India, China, Nigeria, Bangladesh and the Democratic Republic of Congo. However, since 1990, GDP growth rates have been quite high in these areas, as shown in the following table.

Region / country	Average GDP growth p.a. (%)	
Period	1990-1999	2000-2013
World	2.7	2.7
Sub-Saharan Africa	1.9	4.9
South Asia	5.5	6.5
China	9.6	9.8
India	5.8	6.8

Table 5-1: Average GDP growth (World Development Indicators, 2014)

This economic growth has resulted in a steep decline of the number of people in extreme poverty, that is to say, the people who live on less than US $1.25 a day. The decline in the number of people living in poverty, below US $2 a day, is less spectacular though. In 2011, 2.2 billion people were living in poverty, compared to 2.6 billion in 1981.

In many developing countries, we furthermore observe a widening gap between rich and poor, and between those who have and those who do not have sufficient opportunities. It means that access to good schools, healthcare, electricity, clean water and other critical goods and services remains elusive for many people in developing economies.

These trends in growth, poverty and inequality highlight the character of current development. In many countries, many people are still excluded from the fruits of economic growth and development. This exclusion takes two forms:

On the one hand are those who have gainful employment or access to land, but are often still exposed to highly variable or declining real incomes. On the other hand are those who are wholly outside the sphere of income-generating activities: the unemployed and the landless.

How can high economic growth rates go hand in hand with a slow decline in poverty numbers and increasing inequalities? Among others, the dominant trajectory of innovation is one of the causes. This trajectory is characterized by its capital-intensive nature, scale intensity, dependence on high-quality infrastructure, reliance on skilled labour and the usual product portfolio, which is aimed mostly at the needs of the middle and upper class. Taken together, such innovation trajectories systemically disadvantages the poor, both as consumers and producers. It also excludes large segments of the population from productive employment. In short, the dominant innovation trajectory is a partial, but important contributor to the persistence of global poverty.

Achieving inclusive development with frugal innovation

This brings us to the question: can frugal innovations make a difference? Are frugal innovations more inclusive towards poor people than the dominant innovation trajectory?
Let us first try to answer this question with regard to poor consumers. Two issues are important to consider here. First is the identity of the consuming unit, and second the demand characteristics of poor consumers.

People with very low disposable incomes have less capacity to buy goods and services individually. Typically, when poor consumers purchase a product or service, this will be a household purchase (for example, one mobile phone for the whole family), a purchase by several households or a single purchase for an entire village or community organisation (an oxen plough, a water pump, a weather station and so on).
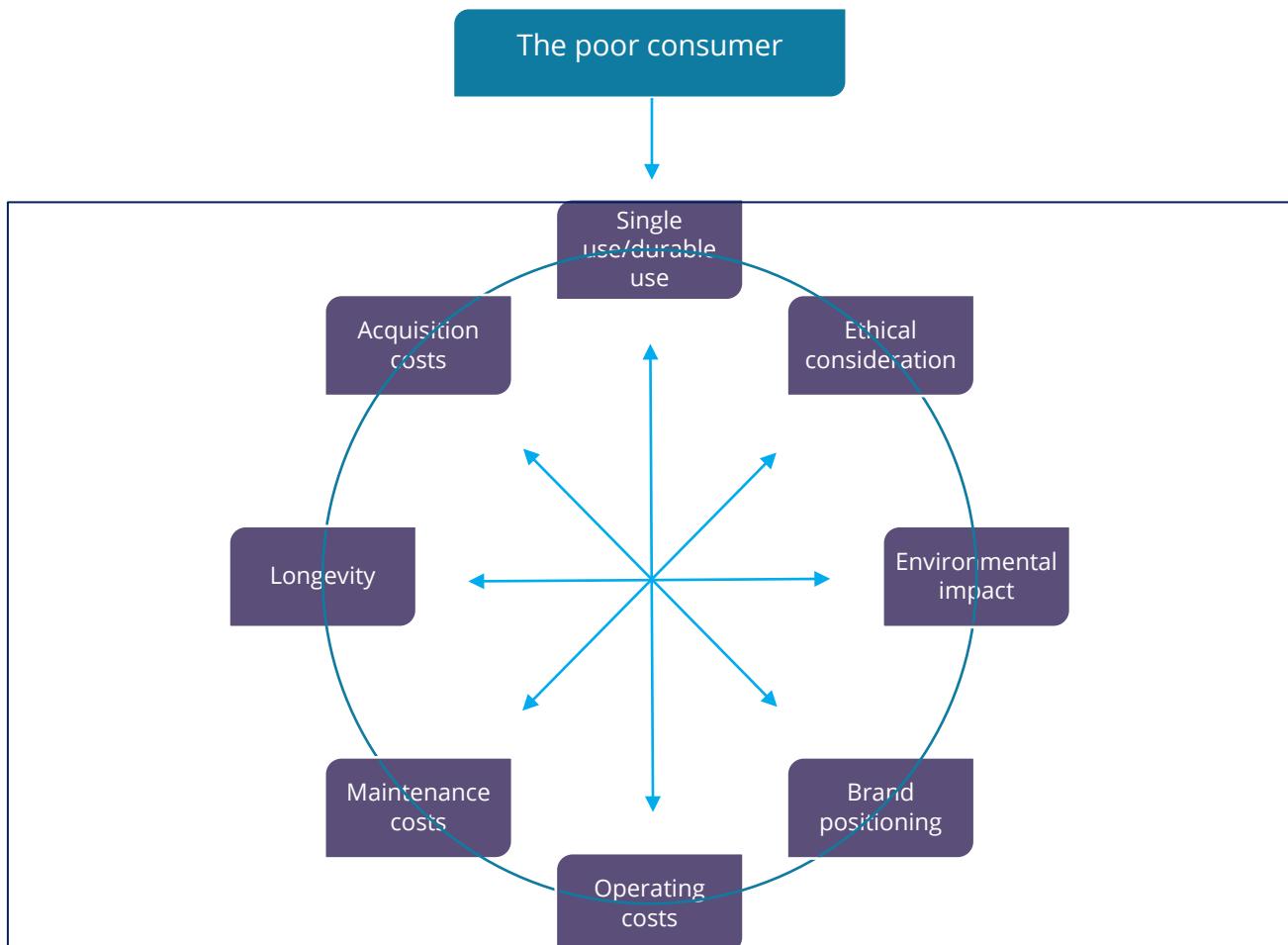
Serving poor consumers

Where frugal innovations aim to lower the acquisition cost of products and services, the more likely consumption decisions will be made at the individual or household level. This means that more people can afford the product or service, and therefore the frugal innovation will be more inclusive. An example is the OMO washing powder sachet for washing in cold water. Providing a small portion allows more poor consumers to buy washing powder at low cost.
Frugal innovations are also more inclusive if they take into account the demand characteristics of poor consumers. The figure below depicts nine product characteristics which may reflect the choices of poor consumers.

These characteristics are whether the product is for single or repeated use, the acquisition cost, longevity, costs of maintenance, operating costs, brand image, impact on the environment and the extent to which the product or service has characteristics which reflects local, environmental and ethical considerations. Frugal innovations typically reflect a characteristic that match low consumer incomes: frugal products and services have low acquisition costs, through which they become affordable for poor consumers.

So far so good for poor consumers. But, as we saw in the previous section, making products or services available to poor consumers may come at a price: the products may not be recyclable, and/or may embody low ethical, security, labour and environmental standards. This means that the inclusiveness of frugal innovations may be at odds with other dimensions, which would make them responsible innovations.


Figure 5-3: Product characteristics

Serving poor producers

We may also ask if frugal innovations are more inclusive towards poor producers than other types of innovation. The majority of poor producers can be found in the so-called informal sector. Poor producers generally have micro, small or medium-sized enterprises, and often they have to use their own or family labour, as very little capital is available to them. Production is generally on a small scale, using unskilled or semiskilled labour.

As a general rule, therefore, inclusive innovations should involve the generation of processes which lend themselves to ownership by small-scale or collective producers, using relatively labour-intensive techniques and utilising unskilled labour. Do frugal innovations fit in this category?

The answer is: not necessarily.

	Poor consumers	Rich consumers
Poor producers	• Informal sector furniture • Informal sector clothing • Informal sector equipment	• Unskilled labour in assembly of iPhones • Small-scale farmers involved in export of flowers
Rich producers	TNCs producing products for the BoP, for example, related to energy use, electronic devices, health and hygiene	Luxury products such as automobiles and watches

Table 5-2: Innovation for poor producers and innovation for poor consumers: some examples.

Consider the chart above. Frugal innovations are mostly to be found in the top left and the bottom quadrant. Typically, we find that multinational or transnational companies (TNCs) are an important driver of frugal innovations. However, it can be questioned whether poor producers are included in the value chain.

There might be inclusive effects, for example in the decentralized marketing and distribution of these products, which can create employment for poor traders, as well as through the employment of unskilled or semi-skilled labour in the production stage. However, the role of local producers will generally be limited, unless they are able to become part of the value chain of the multinational company. They might do so by becoming part of the marketing and distribution network of the multinational, or by becoming a local source of input and information.

This type of innovation is different from frugal innovations which originate from the informal sector itself. In these cases, local producers are involved in the design, production and marketing of these innovations - there is local ownership, as it were. But the spill over may be quite limited. Poor designers and producers face many constraints which prevent them from upscaling and/or linking their activities to other actors in the local or national economy.

Overall, frugal innovations are not by definition inclusive innovations, in the sense that they allow poor producers to 'lock in'. **Redding (2002)** defines the technological lock-in as an extreme example, "when agents continue to employ an existing technology, even though more productive ones exist".

However, poor producers have some comparative advantages to multinational companies when it comes to the design and production of frugal products and services. For example, they know better the demands and preferences of local poor consumers, and they are less vulnerable to reputational damage which may arise from neglecting or not meeting high standards. We need more empirical research to assess to what extent frugal innovations can be inclusive innovations for poor producers.

5.5 Conclusion

We have shown that frugal innovations are not by definition responsible innovations, unless we satisfy the dimension of inclusiveness. Like in the case of social standards, various criteria have to be met for frugal innovations to be inclusive for poor consumers and producers.

With many frugal innovations still designed, produced and marketed by multinational companies, inclusiveness is not necessarily guaranteed. Bottom-up frugal innovations may allow for higher inclusiveness, but poor producers of frugal innovations still face various constraints that limit the upscaling and creation of spill-over effects into the local and national economy. Still, the example of the TAHMO weather stations (see case study # 6) shows that frugal innovations can have a huge potential to be inclusive and thus serve as responsible innovations.

Case Study #7: TAHMO weather stations

The history of cooling is almost as old as the history of humanity. In prehistoric times,

Let us now look at an example of frugal innovation, and the different considerations behind it: the Trans-African Hydro-Meteorological Observatory or TAHMO weather stations. The TAHMO weather stations project is a frugal innovation, since it is a simple concept that tries to replicate the functionality of high-technology sensors in weather stations at relatively low prices, specifically for the region of Sub-Saharan Africa.

The original goal of these frugal weather stations is simply gathering weather and water data. Consider the map below from the World Meteorological Organization, showing the operational weather stations that feed our global weather predictions.

Figure 5-4: Operational weather stations around the globe

Blue stations are functioning at 100%, the rest at a lower percentage or not at all. As you can see, Sub-Saharan Africa is particularly sparsely equipped. This negatively affects the accuracy of weather predictions and the management of water resources. The idea is to leapfrog and make Africa the best monitored continent through a network of 20,000 stations.

Figure 5-5: Insects in moving parts

Figure 5-6: Web of caterpillars

Maximizing functionality and minimizing costs

Researchers from Delft University of Technology, together with researchers at Oregon State University, are trying to build a self-sustaining observation network. Each TAHMO station is a stripped version of an existing product (weather stations as we know them), using cheap sensor technology in order to achieve frugality. For several reasons, we cannot use standard equipment. The costs of a typical weather station are anywhere between US$5,000 and US$15,000. This would be prohibitively expensive for Sub-Saharan Africa. Moreover, such stations demand specialized technicians for continued maintenance. This is why the project aims to install low-cost, robust weather stations that hardly need maintenance.

There are many considerations to take into account in the harsh environment. There should, for example, be no moving parts. As you can see in this picture from Ghana, insects tend to build nests in and around such moving parts, thereby rendering them worthless. For example, a standard weather station has a well-ventilated, screened housing for temperature and relative humidity sensors. When the researchers opened such a housing in Ghana, they found a web of caterpillars around the sensors.

The researchers also want to reduce the costs of the stations by using mass-produced sensors instead of specialized ones. For example, the ZyTemp TN9, which is normally used in non-contact medical thermometers, can accurately measure long-wave radiation at a fraction of the costs of an official radiation sensor.
Another example is the measurement of rainfall, probably the single most important weather variable in the African context. Ideally, one would not only want to know the amount of rainfall, but also the distribution of raindrop sizes. The latter could be important in erosion studies for example. Normally, instruments that accurately measure the size distribution of raindrops cost upwards of US$10,000.

After trying several materials, the researchers found a simple piezoelectric element, which can be found in any smoke alarm and costs about US$1. This element produces an electrical signal when it is mechanically excited. In other words, when a drop falls on it, a signal is produced which can be captured and recorded. The bigger the drop, the bigger the signal. In the calibration curve below, you can see that there is a very nice correlation between drop size and signal strength.
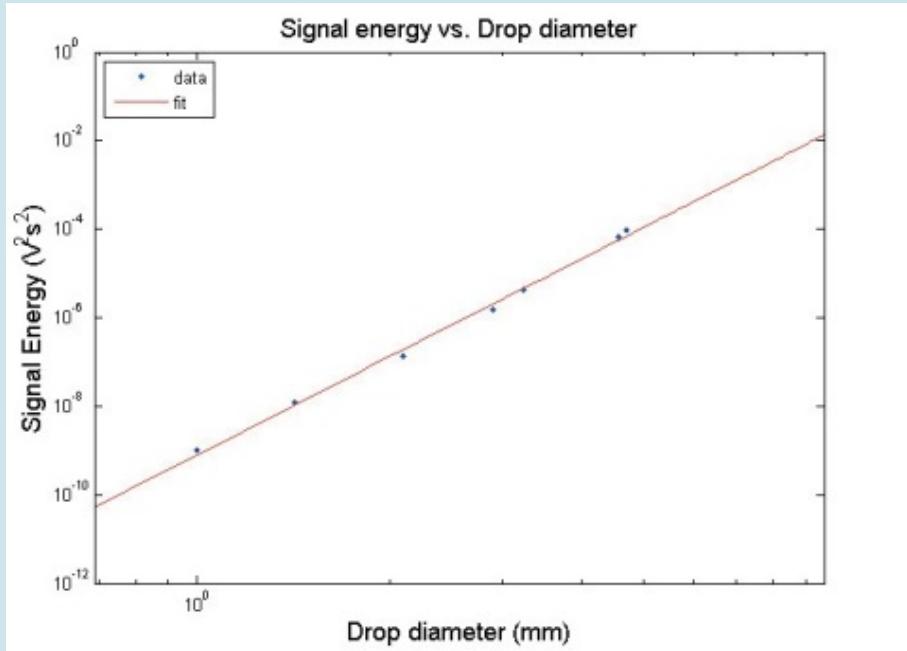

Figure 5-7: Correlation between drop size and signal strength

Leveraging educational networks for support

A second important feature of the TAHMO project is the educational angle. Weather stations typically need fences and dedicated caretakers. One idea is to link up with local schools. By placing the stations at schools, we provide them with protection against theft and vandalism. In return, the schools will have access to the data and to a complete set of educational materials.

The early pilots in Ghana aimed to determine what needs to be done to include weather and water stations in the school curriculum. The project included a school-to-school program in which richer schools in the developed world paid for two weather stations - one to be installed at their own premises and the second one at a relatively poor school in rural Africa. This was followed by a series of lectures on climate, water, weather and the exchange of information between sister schools. The first exchange took place in 2014 between schools in Idaho and Kenya.

Another component of the project was a sensor design competition with African universities, where teams were tasked with designing new sensors along the TAHMO design criteria. This resulted in 23 design submissions. One example was an idea from Nigeria to weigh the desiccant used to protect the electrical circuit. The weight would reflect the relative humidity of the air naturally. It is an interesting example of how to leverage items that are already in use for alternative uses. In this case, we would have an extra data point on relative humidity, other than the sensors. Another example was the idea by Gilbert Mwangi and Ken Odhiambo from Kenya,

which attempted to determine wind speed and wind direction by measuring the movements of a flag.

The thirteen teams with the most interesting designs received a maker package, which included general electronics, such as an Arduino micro-controller and many other tools, to actually build their designs.

Business models for the TAHMO project

At the same time, we see that scaling up is an important issue, and that appropriate business models are required. Special attention is also paid to the development of business cases. Many people tell us that gathering data on weather and water is something the government should do. That may be true, but over the past decades we have only witnessed a decline in environmental monitoring networks around the globe. Data-gathering is not something with which politicians can win over the hearts of voters. The aim of the TAHMO project therefore is to develop public-private partnerships and business cases that are financially compelling. All along the value chain, from weather station installation and operation to data analysis and forecasts, people need to have some incentive to continue to operate the TAHMO network.

The initial financial numbers are significant, but not staggering. It is probably necessary to start off on the basis of government grants and subsidies. However, to continue beyond the grant period, the TAHMO project needs to be financially self sustaining. The potential is there. In the United States, it is estimated that the economic value of weather data and predictions is about US$31 billion per year. In Africa, we would only need to capture a very small percentage of this value in order to maintain the program.

One possible business case would be commodity traders. To know the status of a growing crop of cotton or cocoa would provide important financial advantages with respect to hedging. A small fraction of these advantages would suffice for the upkeep of the TAHMO project.

Very promising also is index-based weather insurance, whereby farmers can use the weather information for taking business decisions; for instance, deciding when to sow crops. Similarly, insurance companies may be interested in these data, since they provide them with better estimates on crop failure, which underlie their calculations for the crop insurances they sell. Thus, farmers insure their inputs and insurance companies pay out when there is not enough rain. Forecasting rainfall could be a service provided by nearby weather stations. The project is now partnering with the insurer Kilimo Salama in Kenya, which leverages the possibilities of mobile phone networks to sell insurance and organize payments.

TAHMO might be able to make Africa leapfrog in the field of weather and water monitoring. By combining innovative design with education and business, the TAHMO network could provide excellent information services.

The targeted consumers consist of various entities. At the individual level, the weather stations have the objective to reach local farmers. Through mobile information services, the station can provide timely, reliable and locally relevant weather data that will enable, for example, local cocoa farmers in Ghana to better manage their limited resources, make more efficient use of the

available water and invest in their farms. Other consumers may be larger co-operatives of poor farmers.

Other - not necessarily poor - consumers may be insurance companies, local governmental bodies and NGOs, which may need the data of the weather station to serve poor clients better and reach them with new services tailored to poor consumers (like the weather-based insurance discussed earlier).

For the moment, possible gains for local entrepreneurs in West Africa lie not so much in the production of the weather stations but rather in the extra employment that the weather stations create for processing the data and the marketing services linked to the weather stations. This may be within banks, insurance or micro-credit providers, but also within ICT companies that are needed to communicate the information. This type of job creation may not reach poor people in all cases, because it requires semi-skilled and skilled labour.

So, the most likely inclusive effects of the weather stations will be that poor farmers can profit from easy-to-access weather data and can thus improve their farm management. Moreover, through new initiatives like weather-based insurance, they could become less vulnerable to income shocks.

6. Implementation of RI by companies: new standard

6.1 Introduction

Research and Development constantly lead to the development of new technologies that can have significant impact on people's everyday life, the communities and territories, the whole economy and society. The more these technologies enable improvements or change paradigms, the greater can be their impacts. Responsible Innovation provides a way to address the needs and concerns of people and society and to develop processes, products and services aiming to positive societal impacts, guiding innovation towards sustainable development goals.

For companies, RI can be helpful to anticipate social or market trends or requirements, technological scenarios, possible regulatory changes and thus inform the overall business strategy and help to save money and time. RI is considered crucial also to build trust and legitimacy.
Early identification of societal needs and concerns helped in designing products that will be better aligned with societal expectations and thus could gain more acceptability by the end-users and by other actors such as partners in the .supply chain actors, regulators, authorities or certification bodies.

However, there have been still limited initiatives looking at implementation (embedding) of RI by companies.

The EU project PRISMA (Piloting Responsible Research & Innovation in Industry [**https://www.rri-prisma.eu**]) helped eight companies to implement RI in their innovation and social responsibility strategies.

Note: The essential difference between RI and CSR (Corporate Social Responsibility), is the focus of RI on the ethical and social impacts during the research and innovation process, from the early stages to prototyping and go to market. CSR is a broader concept.

6.2 Roadmap

Note that in the remaining part of the chapter we will use the term RRI (Responsible Research and Innovation) instead of RI although the meaning is similar in this case.

Based on the experience with these eight pilots - active in different sectors and technologies, - PRISMA developed practical guidelines – **a RRI-roadmap** - for companies aiming to strengthen consideration of ethical, legal and social impacts (ELSI) aspects in their technology and product development roadmaps.

This roadmap is now the starting point for a new European standard for RI by companies. It is aligned with existing standards on social responsibility, risk management, quality and innovation management. For this reason we will discuss the approach in detail.

The roadmap describes the *process* to develop a RRII-roadmap so that it can be context-sensitive, as the results depend on the sector, technology, type of company and business.

Note that the roadmap is aligned with a number of existing standards, i.e.
•	ISO 26000 (Social responsibility)
•	ISO 5600 (innovation Management)
•	ISO 31000 (Risk management
•	ISO 9001 (Quality management)
 (see **www.iso.org** for full details).

Principles for RRI
The roadmap is furthermore based on the following pillars for RRI:

:Principles for RRI implementation	Action lines
Reflection & Anticipation	Integrate analysis of ethical, legal and social impacts (ELSI) since the early stages of product development
Inclusiveness	Perform stakeholder engagement to inform all phases of product development
Responsiveness	Integrate monitoring, learning and adaptive mechanisms to address public and social values and normative principles in product development.

Table 6-1: RRI Principles

These principles are further described below:

- *Reflection:*
  Scrutinize each activity, commitment and assumption in order to connect them with a moral value system and the good practices of science, taking into account the limits of knowledge and that a particular framing of an issue may not be universally held.

- *Reflexivity:*
  It is intended as an institutional practice. It can also be intended as a public matter and

people external to the organization can be part of reflexivity actions. Reflexivity is important also with respect to the other phases of the product value chain or other functions inside the organization (besides the R&D), that could be affected by an R&I action or result.

- *Anticipation:*
  Systematically extrapolate all the plausible scenarios for the application of the R&I results; identify in these scenarios the possible risks, opportunities, uncertainties, critical issues, and draw possible ways to prevent, manage or exploit them. Anticipation isn't only intended to prevent undesirable events, but also to shape desirable futures and organize activities and resources towards them. When describing desirable futures, anticipation should be realistic and avoid to overestimate the benefits of the innovation.

- *Inclusiveness:*   i
  Introduce participatory approaches in the R&I processes from the very early stages, in order to engage people interested with the innovation process or results. Inclusion is referred to the engagement of both internal and external stakeholders. Inclusion is also connected to the other dimensions of RRI, because the reflexivity, anticipation and responsiveness can be improved by a broad participation of different stakeholders.

- *Responsiveness:*
  Change the direction of the innovation process to answer to stakeholder and public indications, needs, and values or to react to changing circumstances. It could be necessary also to adjust innovation actions when recognizing insufficiency of knowledge and control, or in response to new knowledge, perspectives or regulatory requirements. The entire R&I processes should be shaped to be as responsive as possible.

The Roadmap has 6 'process steps'. See the next table:

	Step	Goal
	1. Top management commitment and leadership	Ensure endorsement of the organization toward RRI values and approach
	2. Context analysis	Analyze the organization, the R&I product(s) and technologies to focus on; Identify ethical, social and legal impacts of the product and stakeholders of the product innovation eco-system
	3. Materiality	Identify and prioritize: drivers and challenges for RRI; risks and barriers to overcome; stakeholders to work with; significant RRI actions to pursue

	Step	Goal
	4. Experiment & engage	Perform exploratory/pilot RRI actions, engaging with stakeholders to inform the RRI roadmap
	5. Validate	Evaluate impact of the roadmap on both the product development and the organization (Key Performance Indicators)
	6. Roadmap design	Consolidate and visualize the long-term RRI strategy, covering all the R&I value chain (time to market) and product life-cycle.

Table 6-2: List of methodological steps for the roadmap design

We will now highlight each step below.

✓ *Step 1: Top Management commitment and leadership*

A pre-requisite for RRI implementation is top management commitment. This commitment is necessary but not sufficient to achieve RRI intended outcomes, as the top-down approach should be integrated with a bottom-up approach, involving other roles providing leadership. Top management shall demonstrate leadership and commitment with respect to the RRI by:

- Ensuring that the RRI roadmap, related actions, objectives and vision are established and are compatible with the values and identity and stakeholders the organization is referring to.
- Identifying and sustaining the motivation for the company to engage with RRI.
- Ensuring that RRI principles are integrated into the organization's management systems and governance to ensure that the RRI achieves its intended outcome(s)
- Ensuring that the resources needed for both the roadmap design and its future implementation are available (also on the long term).
- Communicating the importance of effective RRI, supporting the application of the guidance provided in this document.
- Supporting other relevant roles for RRI implementation, for example supporting RRI promoters.

✓ *Step 2: Context analysis*

RRI is connected to a broad spectrum of factors related to the type and management policies of a company, the technology and products it works on, the sectors and markets, the pertinent regulatory frameworks and stakeholders involved. For an effective and efficient RRI uptake, it is essential to identify strategies and practices that fit with the realities and constraints in which the organization operate.

- The ethical, legal and societal impacts, and as well as the technical, strategic, organizational, economic impacts concerning the RRI product.).

- The specific technologies and products, and related R&I projects, on which to focus the RRI roadmap design ("RRI product").
- The development stages of the RRI product, from the start of the analysis to the expected time to market of the product.
- The stakeholders interested/involved in the development of the RRI product throughout the innovation eco-system, including an initial understanding of their needs and perspectives (based on desk analysis).

✓ *Step 3: Materiality analysis*

A key aspect of RRI is anticipation. Identify materiality aspects of the RRI product and the organization early on in the R&I value chain is essential to anticipate impacts, and thus have time to change and adapt the process to ensure creation of value (e.g. maximize positive impacts and minimize negative ones).

The goals of this phase are thus the following:

- Identify relevant ethical, social and legal impacts of the RRI product, and describe them in terms of drivers (creation of value, positive impacts), and challenges (of the organization in achieving the impacts).
- Identify the risks and barriers (uncertainties) to address in order to achieve the impacts. Scientific, technical, strategical, organizational, economic, ethical and social aspects should be considered in determining risks and barriers.
- Select stakeholders within the innovation eco-system of the RRI product to engage with.
- Select significant RRI actions that can contribute to achieve impacts and as well address risks and barriers.
- Set an initial vision of the roadmap, addressing drivers and challenges

An example of questions to deal in a materiality analysis, are the safety and privacy issues related to operations of autonomous vehicle devices. *See also case study # 4.*

What are the safety concerns (both actual and perceived by stakeholders) related to the different conditions of work of the device? How to manage the data the device could or have to collect during its operations? Is the collection of these data critical from a social, ethical and legal point of view? What could be the ethical issues related to autonomous decisions that these vehicles might have to take during their operations? All these aspects are relevant, but depending from the specific device (e.g. autonomous cars or drones), the technology (e.g. the type of data collected, the way these are managed by the device, etc.), the use scenarios (e.g. use of the device in buildings, cities, farms, etc.) and the stakeholders

✓ *Step 4: Experiment and engage*

Stakeholder engagement is one of the pillars of RRI, and it is as well essential in order to validate the materiality analysis and the design of the roadmap. On the basis of the previous steps, it is possible to identify one or few RRI pilot actions that the organization should

perform in order to ascertain the appropriateness and the feasibility of the RRI roadmap. Thus, in this phase the following aspects are addressed:

- **At least one inclusiveness action is performed**, involving stakeholders within the innovation eco- system in discussing and analyzing key ethical and social impacts of the project and in reviewing the draft roadmap.
- **Additional RRI actions are performed**, as a way to practice and pilot activities planned in the roadmap.

As acknowledged by the experience in the social responsibility field (**e.g. ISO 26000, see https://www.iso.org/iso-26000-social-responsibility.html**), the identification of material issues to address is not a simple exercise. While the methods developed from the perspective of economic and financial materiality capture only those relevant areas that impact performance or risks in the short term, from the perspective of RRI the time frame shall consider not only short-term impacts and effects, but also ones in the medium to long term, including both tangible and intangible aspects.

It is important that the views of the stakeholders are always considered and appropriately integrated into the reflections internal to the organization. The stakeholder's analysis involves the identification of relevant groups, organizations and people, their perspectives and relevance. Having this in mind, stakeholders can then be mapped, using one of the many tools available for this purpose. An example of stakeholder analysis is presented in the box. Examples of tools for this purpose are the materiality matrix and the interest/influence grid (see appendix).

The materiality analysis started in this phase, is then complemented by the other phases.

The main objective in this phase is to create a dialogue with stakeholders of the innovation eco-system (as selected in the materiality analysis) to discuss their views and perspectives on the RRI product and its ethical, legal and social impacts, and on the specific elements included in the roadmap. Examples of suitable methods include focus groups, plenary sessions, multi-stakeholder workshops, world- café, and fish-bowl exercises.

The outcome of this phase is a complete materiality analysis, in terms of significant ethical, social and legal impacts to address, and stakeholders of the innovation eco-system and a consolidated version of the roadmap.

ACTIONS	BENEFITS
Set and implement a communication and dialogue strategy on ELSI	Ø Strengthen relations and trust with all stakeholders, networks building
Work with business and social stakeholders sharing values and create positive ethical networks	
Co-design product through dialogue with policy actors and authorities and normative bodies (EU, regional and local)	Ø Reconcile opposing views and bridge opposing values
Organize public dialogues, build/use public platforms for expressing needs and concerns	Ø Creation of new values
Connect to or organize Living labs and social experimentation, using participatory methods	

ACTIONS	BENEFITS
Build user-based communities of practice	Ø Anticipate potential regulatory change
Promote initiatives for social inclusion, provide consumers an official role in the innovation process	Ø Increase product quality, desirability and acceptability
Promote capacity building with vulnerable stakeholders in the value chain	

Table 6.3: Example of inclusiveness actions for stakeholder engagement

✓ **Step 5: Validation**

The success of RRI up-take is strongly context-dependent and is affected by several factors, as underlined in the context analysis clause (e.g. company size, complexity of the organization, features of the technology, the level of innovation and the associated risks). RRI actions could have both tangible and intangible impacts, spanning from long-term strategic factors at the company level (e.g. company reputation) to short-term factors in product development (e.g. alignment with user needs and stakeholder values).

Thus, in this phase the following aspects are important:

- **Identify what needs to be measured and monitored**, selecting criteria to perform evaluation of impacts of RRI actions.
- **Select the methods for measuring, monitoring, and evaluating the impacts** of the roadmap on the RRI product and the organization.
- **Evaluate (at least qualitative) the impacts of the RRI action**s defined in the roadmap, focusing on the added values both tangible and intangible, based on the selected criteria.
- **Explore whether and to what extent the roadmap could be embedded** in the usual innovation, risk, quality, social responsibility policies of the organization. This includes identification of Key Performance Indicators to measure the impact.

In the next table we will give you some examples of responsiveness that address social values in product development.

	ACTIONS	BENEFITS
1	Integrate user-centered design, user innovation, flexible and adaptive design, co-creation	Ø Create value, increase the social value/impact of R&D
2	Screen suppliers for positive practices	
3	Put in place procedures for investigating reports of concerns or misconduct	
4	Ensure non-discriminatory recruitment processes	Ø Build corporate image and reputation
5	Employ adaptive risk management	
6	Embed ethicists in the R&I process	
7	Establish of an ethical, social and legal monitoring board	
8	Include ELSI criteria in internal procedures for R&D project quality monitoring	Ø Compliance with qualified norms and standards
9	Ensure ethical management of research data and **FAIR data management**)	
10	Perform regular ethical review and get ethical certification (by independent bodies)	Ø Facilitate the access to

11	Obtain social accountability and quality certification at company and supply chain level	financial support
12	Post-marketing monitoring of ELSI impacts	
13	Include ELSI of R&D and Innovation products in the CSR/sustainability reporting	
14	Support and invest in sustainable supply chains	
15	Select funding mechanisms based on ethics/responsibility requirements	

Table 6.4: Examples of responsiveness that address social values in product development

✓ *Final step: Roadmap design*

Based on the outcomes of the above-mentioned steps, a RRI roadmap is designed to guide an organization to put RRI in practice RRI implementation. So, as we have seen, this means:

- *Anticipation & Reflection:* Integrate analysis of ethical, legal and social impacts since the early stages of product development
- *Inclusiveness:* Perform stakeholder engagement to inform all phases of product development
- *Responsiveness:* Integrate monitoring, learning and adaptive mechanisms to address public and social values and normative principles in product development.

Monitoring of the actual impact is of course key. The project developed a list of 10 key performance indicators. See the next table for an overview and a further description:

	RRI KPIs	**Examples of Quantitative parameters**
1	Awareness of moral values	Number of training sessions/meetings per year to learn and reflect on moral values connected to innovation strategy and core business
2	Awareness of ethical issues	Number of training sessions/meetings per year aiming to reflect on integration of social and ethical values into specific R&I/R&D projects
3	Does the company embed moral values in its innovations?	RRI principles formally integrated into the company's mission and vision (e.g. ethical code of conduct)
		Number of R&I/R&D projects per year where moral values are actively and included into innovation strategies and technological design
4	Does the company (actively) anticipate social effects of its innovations?	Number of R&I/R&D projects per year where internal/external stakeholders were involved from the early stages in product development
		Number of consultancy initiatives with other innovators and external advisors to discuss and identify social impacts of R&I/R&D projects.
5	Stakeholder engagement	Number of stakeholder engagement initiatives organized per year by the company
		Number of R&I/R&D projects per year where active stakeholder engagement is foreseen into R&I/R&D plans
		Number of R&I/R&D projects per year where engagement with end-users has been performed
6	Gender Diversity	Percentage of men and women involved in R&I/R&D function/teams in the company

7	Transparency and accountability about RRI-relevant choices	Formal communication strategy established at company level to ensure most relevant RRI choices are explained in key company documents and/or the website
		Number of patents per year aiming to integrate non-financial values.
		Number of open access publications
8	Learning mechanisms to address public and social values in product development	Number of user-centered approaches per year formally integrated into the company innovation model (e.g. user-centered design, co-creation)
		Number of user experience tools per year carried-out to respond (new) societal demands and developments
9	Learning mechanisms to address public and social values in product development	Number of R&I/R&D projects per year addressing socially/ethically-oriented products/services
10	Active monitoring of RRI impacts	Percentage of R&I/R&D projects per year that apply impact analysis strategies (e.g. risk management, ethical/social impact analysis, etc.)
		Formal external auditing procedures (at least yearly basis) in place to monitor non-financial values of the company

Table 6.5: Key **performance** *Indicators for RRI to monitor the roadmap*

5.3 Template for RRI-Roadmap

The steps mentioned above will result in the following schematic overview:
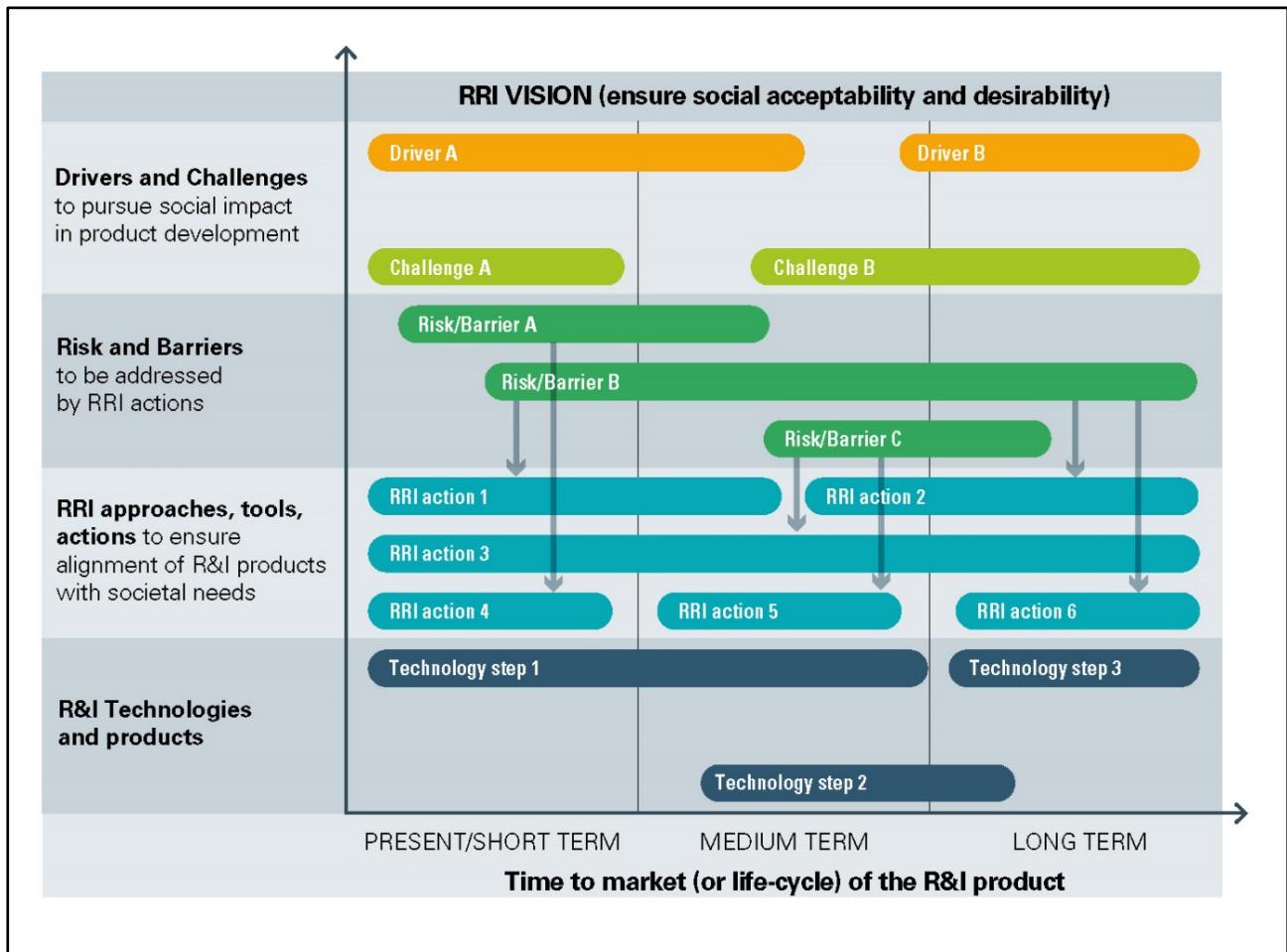
Figure 6-6: Template RRI- roadmap

The template has 4 " lines of action" (which are aligned with the 6 steps mentioned above):

1. The definition of the drivers and the challenges, based on consideration of the significant ethical, social and legal impacts, and strategic, organizational and economic issues at stake, for both the organization and the specific RRI product
2. Identification of the risks and barriers addresses by the RRI actions
3. Identification of an action plan to implement RRI all along the steps for product development, core part of the roadmap
4. Identification of the innovative technologies that enable to address the objectives of the research and innovation (RRI product).

6.4 SWOT analysis for RRI implementation

As indicated above, RRI within companies may require quite some effort and is thus not self-evident. This is a SWOT-analysis, also made by the PRISMA-project,

	Strengths	Weaknesses
Internal organisation	- Create value - Motivate workers - Offer competitive advantage - Strengthen relations with all stakeholders - Increase trust among stakeholders - Increase the social value/impact of R&D - Strengthen quality of innovation at industrial level - Ensure compliance with qualified norms and standards - Identify new market needs - Potential to communicate benefits and risks of products - Increase transparency in product development	- Limited awareness and skills on the RRI concept - Additional bureaucratic burden, lack of resources (particularly for SMEs) - Low perception of tangible impact on product development - Lack of integration of RRI across the company functions - Internal boycott from some functions in the company - Difficulties in measuring associated costs - Adding excessive extra costs to product development - Intellectual Property Rights - Misuse of the concept (checkbox exercise)
	Opportunities	Threats
External organisation	- Improve product quality, desirability and acceptability - Improve product sustainability, safety and reliability - Increase customer satisfaction - Improve effect on quality of life and health of customers – by addressing existing social needs - Improve efficiency (e.g. use of resources, decision- making process) and cost reduction on a medium/long term - build corporate image and reputation - Improve market penetration, profit - Facilitate the access to financial support	- Difficulties in engaging with stakeholders - Possible slowdown or even premature stop of innovation - Few practical examples available from industry (case studies, applications) - Lack of engagement along the value and supply chain - Lack of endorsement by partners and suppliers - Seen by stakeholders as a "window dressing" exercise - Lack of incentives (at policy and regulatory level)

Table 6-7: SWOT analysis

Part V:
# Risk assessment and safety

Figure 7.1 The Bhopal disaster, a gas leak incident on the night of 2–3 December 1984 at the Union Carbide India Limited pesticide plant in Bhopal, India. It is still considered to be the world's worst industrial disaster.

7. *Understanding risk*

7.1. Risk, Uncertainty and Ignorance

So far, we have seen what RI is and why it should play an important role in the development and diffusion of new technologies. Let us now look at risk, uncertainty and ignorance in technology and how they can be mitigated. Our running example will be anthropogenic climate change as induced by the burning of fossil fuels.

When we talk about climate change as risk, we need to be clear how we use the term "risk". When we use the term colloquially, we use risk in statements like "smoking increases the risk of cancer", or "this hole in the ground is a risk". While in the first sentence we can replace risk by likelihood, probability, or even possibility, this does not make sense in the second sentence. Here risk is synonymous with actual harm or immediate danger. Now, in science or in philosophy, the term risk always comprises both meanings: that of certain harm and of probable harm.

The difference between risk and uncertainty
Sometimes we may assign probabilities to indicate uncertainty about the harm's occurrence. It is the natural, social or engineering sciences that provide the probability that a certain valve in a nuclear power plant will break, for example. However, what constitutes harm always derives from a normative concept which goes beyond the sciences and requires some ethical expertise. For example, to understand why climate change is actually a harm, we need a normative concept that tells us why this is the fact, and also why we need to care about the environment at all. So, risk is per se an interdisciplinary concept and this has important consequences.
Firstly, anthropocentric ethics tells us that climate change in itself is not a harm, but rather, the implications of climate change may be dire for human beings. These implications are actually not modelled through climate models in the narrow sense, but require so-called welfare economic impact models, made by economists. Hence, better political decisions that recommend how to react to the threat of climate change may not require better climate models, but better climate impact models - an area of research currently not nearly as active as climate models.

Secondly, the ethical, i.e. normative, evaluation should, at least in parts, precede the empirical, scientific prognoses that analyse uncertainty about a certain harm. In cases where uncertainty can be quantified in terms of probability, risk is often defined as mean harm - that is to say, harm times its probability of occurrence.

Without these probability estimates for the occurrence of harm, we say in technology assessment that reacting to climate change does not constitute a decision under risk, but one under uncertainty. For a decision under risk, we know all possible outcomes of the decision - like choosing not to mitigate climate change - and we can assign meaningful probabilities to these outcomes. Uncertainty, however, refers to situations where we know the full probability space, but cannot assign probability to all outcomes.

The difference between uncertainty and ignorance

In technology assessment, we further distinguish decision under ignorance, when not even the probability space is known. Such situations recently became famous, as former US Defence Secretary Donald Rumsfeld termed them "unknown unknowns", or what Nassim Taleb calls "black swans".

Mitigating climate change is an example of a decision under uncertainty, whereas the introduction of CFCs in the 1970s is an example of a decision under ignorance. At the time of the market release of CFCs, their damaging effect on the ozone layer could not have been known.

Dealing with risk, uncertainty and ignorance

This distinction between risk, uncertainty and ignorance commonly rests on a certain interpretation of probabilities as relative frequencies. These objective probabilities are common in technology assessment and also in many aspects of engineering and science.

There are advocates of a more subjective view, in which probabilities are grades of belief, instead of relative frequencies. This is sometimes referred to as the **Bayesian approach**. In theory, this would blur the distinction between risk, uncertainty and ignorance. In practice, however - for climate change in this case - assigning subjective probabilities remains difficult, as we cannot update our beliefs and assigning a priori probability distributions in a Bayes formula is difficult.

So, how do we deal with risk and uncertainty?

For risk, we may use what is known as maximizing the expected utility analysis, or, formulated negatively, risk minimization or risk analysis. This is simply the utilitarian paradigm of the greatest good for the greatest number of people. However, since we do not know the exact outcome, we can only maximize the expected good or utility, or negatively, minimize the expected damage, i.e. risk. A typical example of such an approach would be policies concerning nuclear power.

When no suitable probability estimates are available like in the case of climate change, this approach is of course not applicable. We may fall back onto a more elementary decision approach that does not require any probability estimates. The most prominent example of such an approach in environmental and engineering ethics is *the Precautionary Principle*. The Precautionary Principle is used in various ways, but most of them are a variant of the following two versions.

The first version is cited here from the **Declaration on Environment and Development in 1992**: "Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation". This version is considered the weak formulation, because it advises us to take into account any possible implications of technologies where full scientific certainty is not available. However, it does not say exactly how to deal with such uncertain situations. Still, we may be able to perform a risk analysis, in which we can at least make sure to consider uncertain effects as well.

The second version reads: "In its simplest formulation, the Precautionary Principle has a dual trigger: if there is a potential for harm from an activity and if there is uncertainty about the magnitude of impacts or causality, then anticipatory action should be taken to avoid harm". The version - also known as the strong formulation - does advise us on how to act. It even tells us that when the harm is uncertain, anticipatory action should be taken to avoid it. So, no matter how unlikely a negative impact is, and even when we do not know how severe this could be, we need to take action to avoid those negative outcomes.
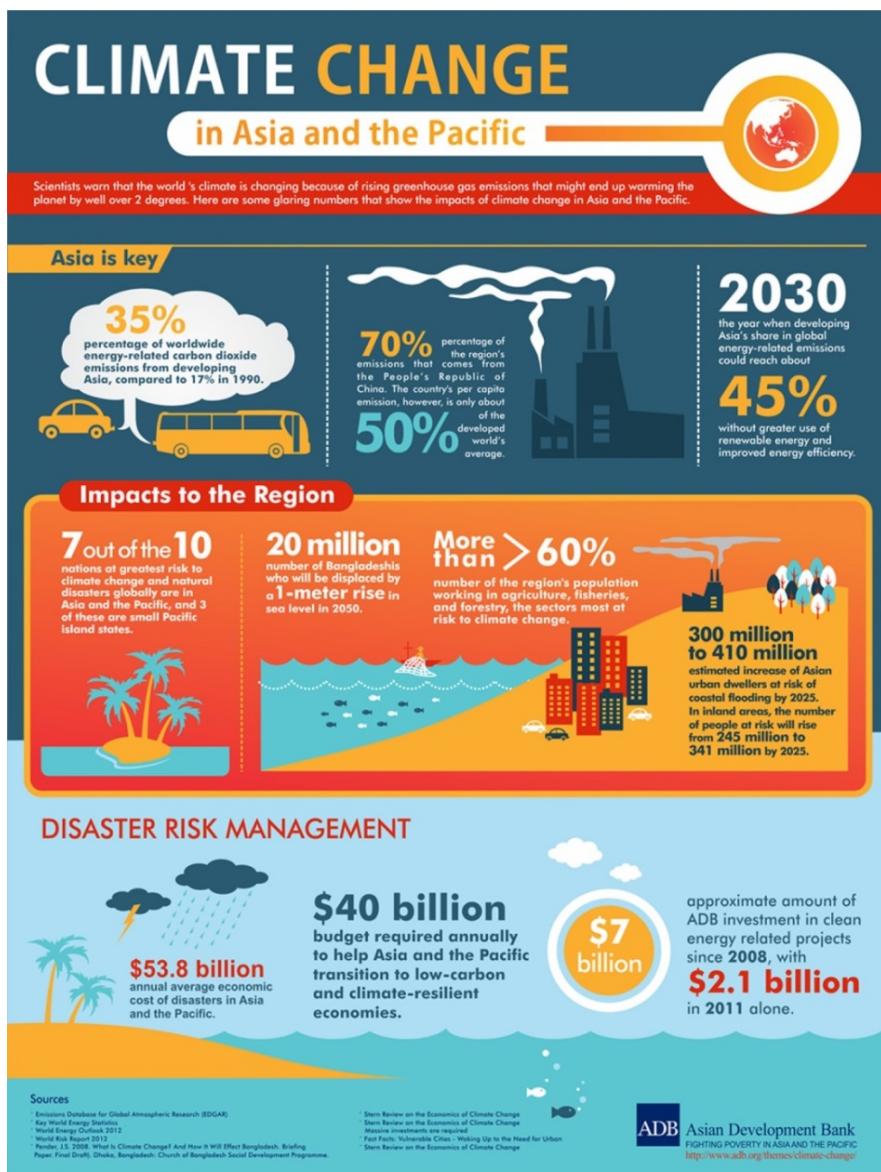
Figure 7-2: An example of the impact climate change may have

The Precautionary Principle and moral overload

Applying this formulation to the issue of climate change means that we need to mitigate any negative impact climate change may have. Therefore, it comes in handy that reducing anthropogenic greenhouse gas emissions is not very costly, as some economic assessments suggest. For example, in the Stern Report from 2007, we find that an annual investment of only 1% of global GDP is needed to avoid the main damage caused by global warming. This amounts to about US$450 billion per year.

It is hard to grasp such a big number, but we can perhaps compare it to other figures. Consider that 'only' US $1.3 billion per year is needed to fulfil one of the UN's Millennium Goals, to provide 80% of the rural population of Africa with safe water and sanitation. This comparison shows that simply applying the Precautionary Principle falls short of adequately accounting for this comparison. If we invest 1% of global GDP per year to avoid climate change, we must accept that that money is not there for other goals. How can we decide?

This is not a question that can be answered in a short chapter, but one that needs political discussion, and more than that: it needs an interdisciplinary approach to risk and uncertainty, in which not only the harm, but also the likelihood of its occurrence needs to be taken into account – whether this likelihood can be quantified or expressed in terms of a probability.

7.2 Extreme uncertainty of unknown unknowns

In 1943, Thomas Watson, chairman of IBM, said: "I think there is a world market for maybe five computers." Of course, he was thinking of large mainframes like ENIAC. He could not have known that in just half a century, there would be PCs, laptops, tablets, smartphones et cetera, and that nearly everyone would have their own computer, or even more than one. Nevertheless, this anecdote shows that it is hard to predict the future.

The Collingridge dilemma
Let us start with the famous Collingridge dilemma, which observes that:

> "In the early phases of technological development, technology can still be changed, but the effects of technology can be hard to predict. In the later phases, we see the opposite, where the effects are clear, but technology is already embedded in society and therefore much harder to change".

Most current approaches to the Collingridge dilemma focus on anticipation: an attempt is made to make technology more predictable.
We will discuss here two ways of anticipation, first using a risk approach and then using the Precautionary Principle.

The risk approach proceeds as follows. First, we determine the risks of a new technology. Then, we decide whether these risks are acceptable. Risk is here objectively understood as likelihood times severity. However, the problem is that we often do not know the probabilities, which results in uncertainty. Sometimes we do not even know all possible consequences – and so end up in ignorance. As a consequence, we cannot actually determine the risks.

An alternative approach is the Precautionary Principle. As we already stated, this principle argues that when an activity poses a threat to the environment or to human health, precautionary measures should be taken, even if some cause-and-effect relationships are not fully established scientifically. Note that this principle does not require the establishment of probabilities; it can therefore deal with what we have called uncertainty.

Drawbacks of the Precautionary Principle
This principle has two drawbacks. First, it might give conflicting advice. Consider the following case. We want to apply the Precautionary Principle to the capture and storage of carbon dioxide. In the Netherlands, it was proposed to store carbon dioxide below the town of Barendrecht, close to Rotterdam. However, this proposal led to heated opposition.
If we apply the Precautionary Principle, one might say: yes, we should capture and store carbon dioxide, because it contributes to the greenhouse effect, which is a clear harm. But we could also say no - by applying the same principle - because if carbon dioxide escapes from the storage facility, it might be dangerous as well. Both perspectives refer to possible but uncertain dangers and therefore we cannot make a decision on the grounds of the principle alone.
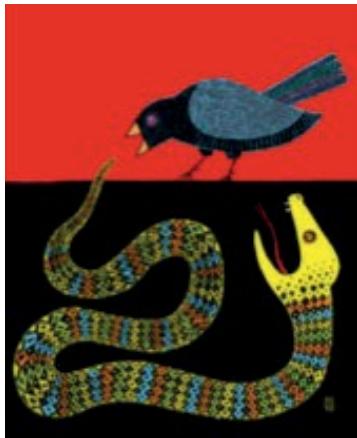
Figure 7-3 : Protests against CO_2 storage: do we know the risks for the long term

The second problem of the Precautionary Principle is that it cannot deal with ignorance. Ignorance may lead to 'unknown unknowns'.

This is nicely illustrated in this image.
This Image above suggests that there will always be surprises and unexpected developments when we introduce new technologies into society.

Figure 7-4: Unknown unknowns (painting by Robert Meganck

The **European Union expert group on science and governance** expressed this in 2007 as follows: "We are in an unavoidably experimental state. Yet this is usually deleted from public view and public negotiation. If citizens are routinely being enrolled without negotiation as experimental subjects in experiments which are not called by name - then some serious ethical and social issues would have to be addressed."

Therefore, we propose to conceive the introduction of new technology in society as a social experiment. This will lead us to ask the question: under what conditions are such experiments morally acceptable?

We can now ask how we can achieve responsible innovation. What kind of approaches can we take and how should we use the tools available? This line of questioning falls within the field of technology assessment. Responsible innovation approaches in general have evolved from the

larger practice of technology assessment.

7.3 Technology assessment

Forerunners of responsible innovation

There are two main forerunners of technology assessment: *ELSI* and *impact assessment.*

- **ELSI** stands for Ethical, Legal, and Social Implications; this program officially started in 1990 as a part of the Human Genome Project. It was aimed at identifying the ethical, legal and social implications of the mapping of the human genome. Five percent of the annual budget of the project was allocated to address the ethical, legal and social issues arising from the project.

- **Impact assessment** is another important forerunner of responsible innovation. Its history goes back to the late 60s and early 70s. It aimed at identifying the future consequences of a current or proposed action. There are many kinds of impact assessment, some of which are legal requirements in some countries, before certain projects can be carried out. Impact assessment can include environmental impact assessment and risk assessment, but also health impact assessment, social impact assessment and gender impact assessment, among others.

Technology assessment (**TA**) is also a form of impact assessment. It can be described as an attempt to objectively predict the social consequences of new technologies in order to provide input for policy making by the government. In the United States, the Office of Technology Assessment (OTA) was established in 1972 and served as an official body until 1995. Its purpose was to provide the US Congress with an objective analysis of complex scientific and technical issues. Although the OTA has been disbanded now, several countries still have a similar organization.

Although technology assessment started off as an attempt to objectively predict the consequences of technology for policy makers - no small task on its own - it has evolved even further over time. It now ranges from the objective prediction of expected consequences to the anticipation of possible consequences, across governments, companies and research organizations; from reactive to proactive approaches, even influencing R&D and design. So, by virtue of its broad scope, TA encompasses many of the values behind RI.

Types of technology assessment

There are large number of TA approaches. Here, we will briefly elaborate three approaches:
1. *Constructive Technology Assessment (CTA),*
2. *Midstream Modulation*
3. *Network Approach for Moral Evaluation (NAME).*

Ad 1. Constructive Technology Assessment (CTA)

Let us first look at the approach of constructive technology or CTA. CTA was developed in the Netherlands in the 1980s by Arie Rip and Johan Schot. The aim of CTA is to reduce the (human) costs of learning by trial-and-error. It aims to do so by anticipating future developments and their impact. The aim is also to feedback these insights into the design process of technology.

CTA has some specific goals as well: firstly, learning about the social consequences of future developments; secondly, reflexivity - which implies awareness of the other actors - and the thirdly, anticipation of possible technological developments and their possible social consequences. In striving for these goals, CTA also aims at broadening technological development by including more aspects and involving more actors.

One of the tools used in CTA is the building of scenarios. The aim of such scenarios is not to predict the future; rather, the aim is to anticipate. Such possible futures help to avoid worst-case scenarios. It also helps to develop strategies that are robust for various possible futures.

Ad. 2 Midstream Modulation

A second method is **Midstream Modulation**. This method was mainly developed by Erik Fisher in the United States. The method is directed at research laboratories where new technologies are developed. The aim of Midstream Modulation is to enhance the responsive capacity of laboratories to the broader social dimensions of their work. The term midstream is used to stress that the method focuses on modulating (changing) R&D practices. The reason for this is that, rather than making an upstream decision on what research to fund or make downstream decisions about how to use particular technologies, guiding the R&D process is seen as more preferable. The method has mainly been applied to nanotechnology, an emerging technology that proceeds by manipulating the properties of materials on the nanoscale - 10 to the power of minus 9 meters.

The National Nanotechnology Initiative in the US pays attention to ELSI and to what it calls responsible nanotechnological development. Similarly, the Dutch Nanonext program also pays attention to risk assessment and technology assessment concerning nanotechnology. Midstream Modulation implies the inclusion of a humanist perspective, perhaps by involving social scientists or ethicists at the work floor in research laboratories. This humanist field agent undertakes the following activities: participant observation, asking laboratory peers thoughtful questions, discussing various issues and giving feedback with different perspectives.

Ad. 3 Network Approach for Moral Evaluation (NAME)

A third approach is the network approach for moral evaluation, or NAME. This approach was developed here in Delft University of Technology. It starts from the assumption that innovation takes place in the context and presence of social networks. Such networks consist of companies, research laboratories, universities, users, suppliers, customers etc. All of these actors play a role and influence the development and diffusion of innovation in some way. The idea behind NAME is to trace these network dynamics in order to discern moral issues. The approach also consists of network norms to judge such networks. The two main network norms in the NAME approach are, firstly, learning and reflexivity, and secondly, openness and inclusiveness.

With respect to learning and reflexivity, one can make a distinction between first-order learning and second-order learning. First-order learning is about how to achieve goals; for example, learning how to improve a technology, whereas second-order learning is about what goals to achieve - e.g. what values should be incorporated in technological development and design. Openness and inclusiveness can also be further defined. Openness means that it is possible to reformulate the central issue of the network. Inclusiveness means that all actors and relevant considerations are included in a network.

Finally

As we have seen in chapter 6, RI has four main components.

They describe that responsible innovation should be:

1. *Anticipatory:* It should anticipate possible social consequences of new innovations;
2. *Reflective:* It should reflect on underlying purposes, motivations and potential impacts, and on what is known and what is uncertain;
3. *Deliberative:* It should include a wide range of stakeholders and perspectives;
4. *Responsive:* It should influence the direction of technological development and design by responding to social and ethical concerns.

We can clearly see that all these four components are inspired by earlier Technology Assessment approaches.
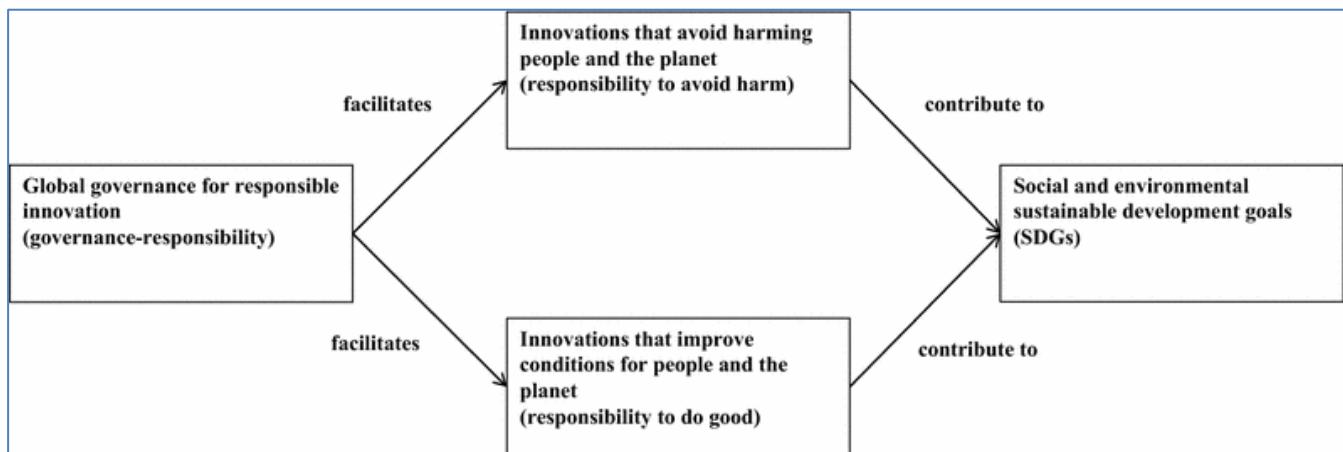
Figure 7.5 Governing responsible innovation for sustainable development

Case Study 8 #: The debate on nuclear energy

The history of cooling is almost as old as the history of humanity. In prehistoric times,
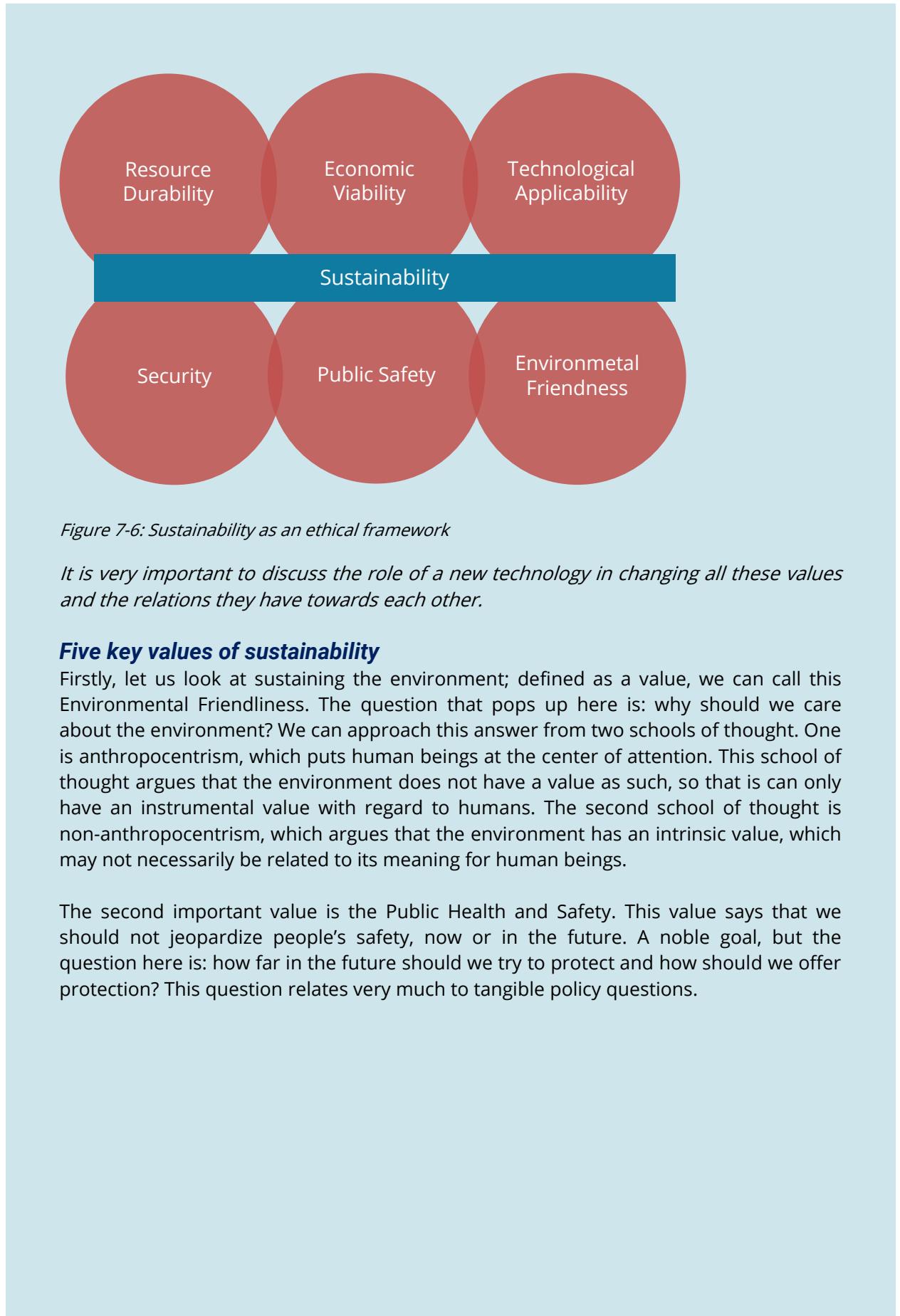
To put what we have just discussed into real-world perspective, we will discuss the case of nuclear energy production and the values at stake in the production of nuclear power. First, we will present an analysis of several values at stake when we are producing nuclear energy in what is called a nuclear fuel cycle. This is called an ex-post analysis, an analysis of an already existing technology.

We can also do an ex-ante analysis, which is an analysis of a technology that does not yet exist. This is actually the more important analysis with regard to responsible innovation, as it tries to accommodate important values prior to and also during the development of new technology.

Sustainability as an ethical framework

We must first be very clear about the definition of sustainability and, in that definition, there are several values at stake. We will argue here that sustainability is to be considered a moral value next to five other values. Each of these values has a temporal and spatial dimension. What are these values? In the discussion on sustainability and ethics, the very first question that we need to answer is the question of: sustaining what?

We distinguish here between two different aspects. Firstly, sustainability could relate to sustaining the environment and mankind's safety; as such, we are talking about the environment and about public health and safety. But sustainability could also relate to sustaining human well-being. Here, we speak of resource durability and the economic aspects of a new technology. Again, each of these values will have a spatial and temporal dimension.

```
[Diagram: Six red circles arranged in two rows of three, with a horizontal teal bar labeled "Sustainability" across the middle]

Top row:  Resource Durability | Economic Viability | Technological Applicability
                        Sustainability
Bottom row: Security | Public Safety | Environmetal Friendness
```

Figure 7-6: Sustainability as an ethical framework

It is very important to discuss the role of a new technology in changing all these values and the relations they have towards each other.

Five key values of sustainability

Firstly, let us look at sustaining the environment; defined as a value, we can call this Environmental Friendliness. The question that pops up here is: why should we care about the environment? We can approach this answer from two schools of thought. One is anthropocentrism, which puts human beings at the center of attention. This school of thought argues that the environment does not have a value as such, so that is can only have an instrumental value with regard to humans. The second school of thought is non-anthropocentrism, which argues that the environment has an intrinsic value, which may not necessarily be related to its meaning for human beings.

The second important value is the Public Health and Safety. This value says that we should not jeopardize people's safety, now or in the future. A noble goal, but the question here is: how far in the future should we try to protect and how should we offer protection? This question relates very much to tangible policy questions.

Let's consider the example of the Yucca Mountain Repository, the world's first and biggest nuclear waste repository, built in the United States over the last couple of decades (see image below).

When introducing radiation standards for the Yucca Mountain Repository, the US

Figure 7-7 Yucca Mountain Nuclear Repository

Environmental Protection Agency (EPA) presented several standards. The EPA argued that for the next 10,000 years, we should offer exactly the same level of protection. Practically speaking, for the generations living during the next 10,000 years, that is about 15 millirem per year. (REM is a unit for measuring the health impact of radioactivity, that is, radiotoxicity.) Beyond that period, EPA guarantees a much lower level of protection. Consider that the first proposal was 350 millirem and only later, after a lot of public debate, the figure was adjusted to 100 millirem per year. This means that, beyond 10,000 years, the level of protection is six times less than what is offered to the present generation. This goes back to the fundamental question: how far in the future should we care and for how long can we offer the same protection?

The next issue is Security. In nuclear energy discussions, we make a distinction between safety and security in the following sense. Safety is about unintentional harm, while Security relates to intentional harm. When we discuss security, we talk about sabotage - the possibility of making a dirty bomb for instance - and about non-proliferation. Non-proliferation relates to issues like the manufacturing of a bomb - a device that could be used for destructive purposes - or the dissemination of knowledge that can contribute to the manufacturing of such a bomb. Therefore, safety and security are two separate

notions, discussed separately in nuclear energy debates.

The next value relating to sustaining human wellbeing is Resource Durability. This
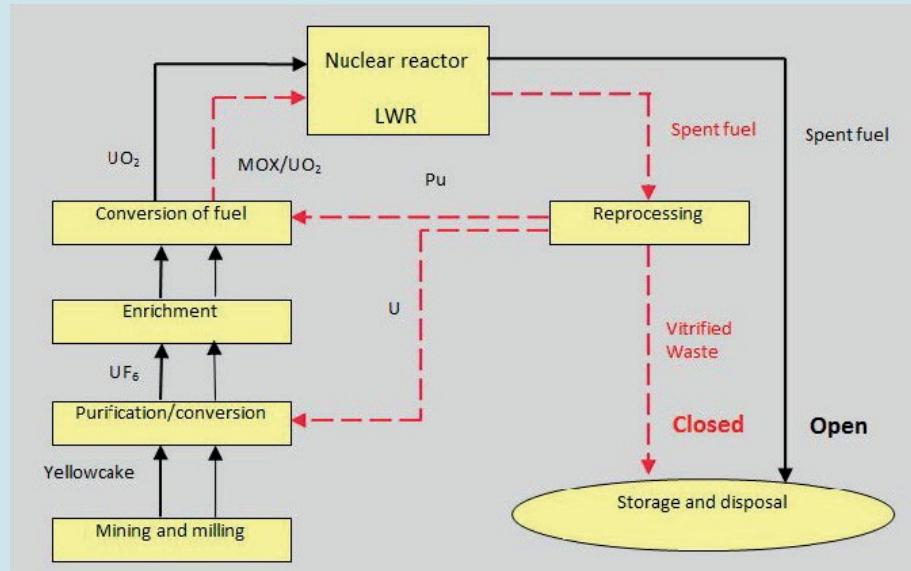
Figure 7-8: Two methods of nuclear power production

involves the availability of natural resources. Durability is a very common understanding of sustainability. Of course, we cannot stop using non-renewable resources immediately, and there will have to be a transition period. So, here the moral question at hand is: to what extent can we offer compensation for the resources that we have used, which future generations will not be able to access?

The last key value we will discuss here is Economic Viability. For an energy source to be sustainable, it needs to be economically durable. This again raises many moral questions. Durable for whom exactly? Whose interests are at stake and whose interests do we need to seriously take into account in our moral analysis? Are future interests as important as present ones? And if not, how do we value future interests compared to the current interests?

In economic studies, the notion of future evaluation becomes important. The value of future interests would be discounted for a certain percentage against the present value. Discounting is a very important aspect of cost-benefit analysis (CBA) and is performed in the interest of future generations.

Open and closed nuclear fuel cycles
Let us now look specifically at the key reaction in nuclear energy production, the nuclear fuel cycle. Discussing the fuel cycle is necessary in order to understand what options are available to deal with nuclear waste. In the picture below, we see the two dominant fuel cycles currently in use.

The black arrows denote what we call open fuel cycles, commonly used in the US, in Sweden and some other countries. Uranium is first mined and milled uranium, then

purified, converted and subsequently enriched. The processed uranium oxide passes through the nuclear reactor. What comes out of the reactor is called spent fuel. Spent fuel could be considered as waste, but due to its high radioactivity, it needs to be disposed of underground for a period of between 200,000 years to a million years.

However, spent fuel can also be recycled in a closed fuel cycle. We call the process of recycling spent fuel reprocessing. The greatest benefit of reprocessing is that the still usable materials, uranium and plutonium, can be salvages by extracting them and re-inserting them into the fuel cycle. A second important benefit of recycling or reprocessing is that we can drastically reduce the waste lifetime, from a minimum of 200,000 years to about 10,000 years. However, reprocessing is a chemical process that also introduces waste. More importantly, the plutonium extracted in the reprocessing process is a type of material that could easily be used for manufacturing a nuclear device. This means that reprocessing causes a very important security risk.

	Resource durability		Environmental friendliness		Economic lability		Public health safety		Security	
	Short	Long	Short	Long	Short	Long	Short	Long	Short	Long
Open	+	-	+	-	+	-	+	-	+	-
Closed	+	+	-	+		+	-	-		+

Table 7-1: Relating values to fuel cycles

In short, we can relate each of these values mentioned above to the underlying nuclear fuel cycle: the open or the closed cycle.
The argument presented here is that the open fuel cycle is particularly good for the present generation, because it introduces the least amount of burdens on the present generation.
The closed fuel cycle on the other hand is better for future generations, because it reduces the waste lifetime significantly. However, the closed fuel cycle brings various additional risks, notably the security risk, and also the safety risk of the reprocessing plants for present generations.

Safety in the design of nuclear reactors
In responsible innovation, we can try to anticipate and accommodate the values at stake during the design phase itself. Let us look briefly at the nuclear reactor, focusing on the history of nuclear reactor design - particularly on the notion of safety as a leading criterion in that design. There are other values at stake as well, because safety is not the only important value. Sometimes, we have to design for conflicting values,. Again, a main issue of responsible innovation is to understand these values and to address these conflicts prior to developing new technologies.

Safety is one of the most important design criteria for nuclear reactors. After every notorious nuclear reactor accident, safety rises again as an imperative condition. For instance, consider the nuclear accident in Harrisburg, Pennsylvania - the famous Three Mile Island (TMI) accident.

Probabilistic Risk Assessments are used in order to reduce the probability of a meltdown in a reactor. These risk assessments were actually introduced a couple of years before the Three Mile Island accident. Probabilistic Risk Assessment tries to map events that could contribute to a meltdown, and it assigns action points to prevent or mitigate those events, with higher priority given to higher probability events. Eventually, we assign a probability to the meltdown outcome as the final event and try to reduce that probability.

The Probabilistic Risk Assessments - made in 1975 by the Rasmussen Group - anticipated that the risk of meltdown of a reactor would be 5×10^{-5}, i.e. one every 20,000 reactor-years. Be aware that this figure is not actual years, but rather, years of reactor operation, hence reactor-years. If there are 500 reactors, there could be one accident every 40 years. That is, or was, how the argument went back then, and it was deemed a fairly acceptable risk.

However, it was decided to adjust the reactors, because considerable growth was anticipated. From 500 reactors at a time, production was scaled up to an expected 5,000 reactors, which meant ten times more reactor-years. In turn, this meant that any accident, should it occur, would be ten times more likely. So, going from 500 reactors to 5,000 reactors - based on the same calculation - would imply that an accident would occur once every four years. This was absolutely unacceptable, and motivated serious change in the design of nuclear reactors.

Two different approaches for making reactors safer. Firstly, we can make incremental changes to the safety. This means that we take the current design as the point of departure, and then add safety features or remove unsafe elements. The second approach is a radical approach of change in the design. This means we start from scratch, redesigning with safety as the leading criterion.

The paradox of designing for safety
In nuclear reactor design, we refer to different generations of nuclear reactors. The first generation (Generation I) are the prototypes, which do not exist anymore. The second generation (Generation II) of nuclear reactors are the reactors which are right now, all around the world. Beyond Generation II, we talk about Generation III, III+ and Generation IV reactors. Some Generation III reactors are already operational right now, while Generation III+ and IV are still being developed.

As mentioned earlier, there are two different approaches to improving safety.
- One is incremental improvement of safety, indicated in green;
- The other is a radical change in design, indicated in blue.
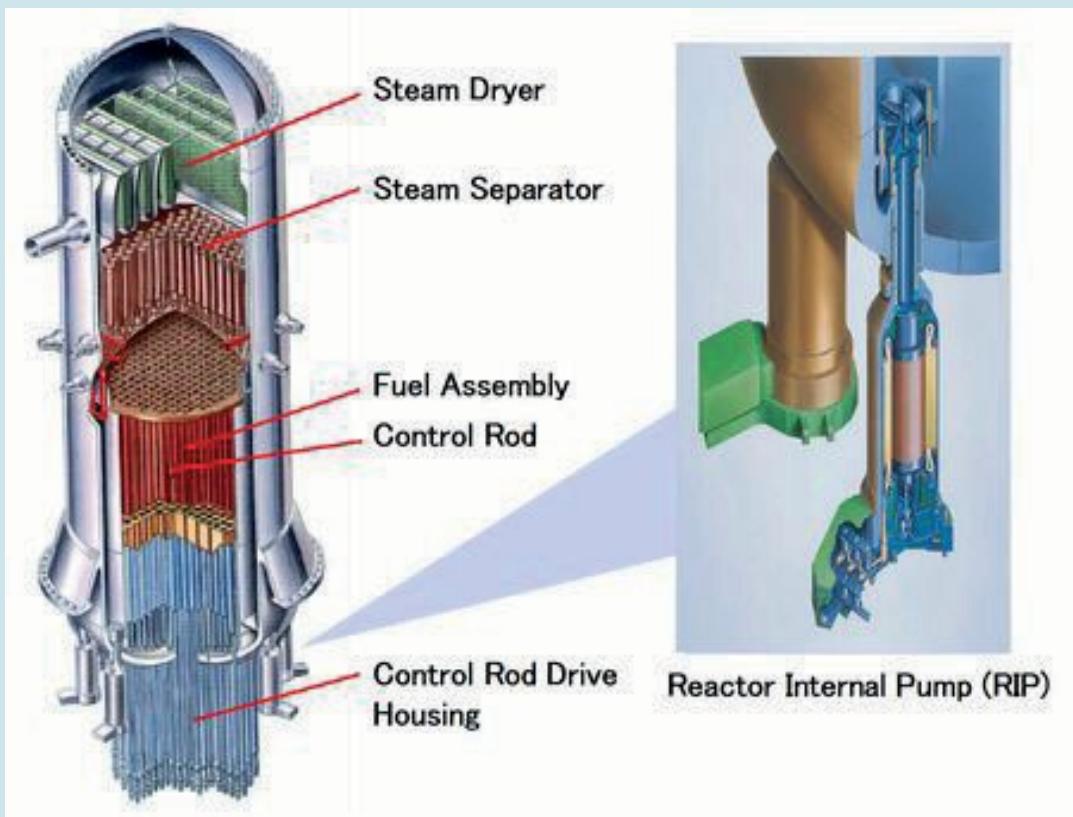
For radical design change, *not* only safety is relevant; there are other values at stake which we should design for. Generation IV reactors are supposed to be highly economical, and to enhance safety, minimize waste and be proliferation-resistant; they are in effect designed to accommodate a multitude of values.

This means that the paradox of reactor safety is the following. Most of the reactors operational right now are the Generation II reactors - namely, Boiling Water Reactors (BWR) and Pressurized Water Reactors (PWR - these are also referred to as Light Water Reactors). Light Water Reactors, especially the PWRs, were originally designed for submarines, but subsequently were scaled up and used for commercial nuclear productions.

When PWRs were later adapted for even bigger commercial reactors, the scale-up implied that safety would inversely decrease with size. Hence, various safety features were added, such as valves, pumps et cetera. However, these additional features made the design more complex and this complexity again exposes the reactor to additional risks. This is the paradox of safety.

Values and innovations in nuclear reactor design

Let us look at Generation III reactors, which represent an evolutionary design based on the Boiling Water Reactor (BWR) of Generation II. The primary reason for designing the Advanced Boiling Water Reactor (ABWR) was to make the BWR safer.

Figure 7-9 Reactor vessel of an ABWR. On the right side, the internal pumps are placed under the reactor, so less piping is needed. This means the complexity decreases, and hence the reactor is

safer. There is also redundancy, a key feature of safety.

There were many additional safety features, like ten separate internal pumps at the bottom of the reactor vessel and thick fibre-reinforced concrete containments. Through these features, ABWRs substantially reduce the risk of meltdown.
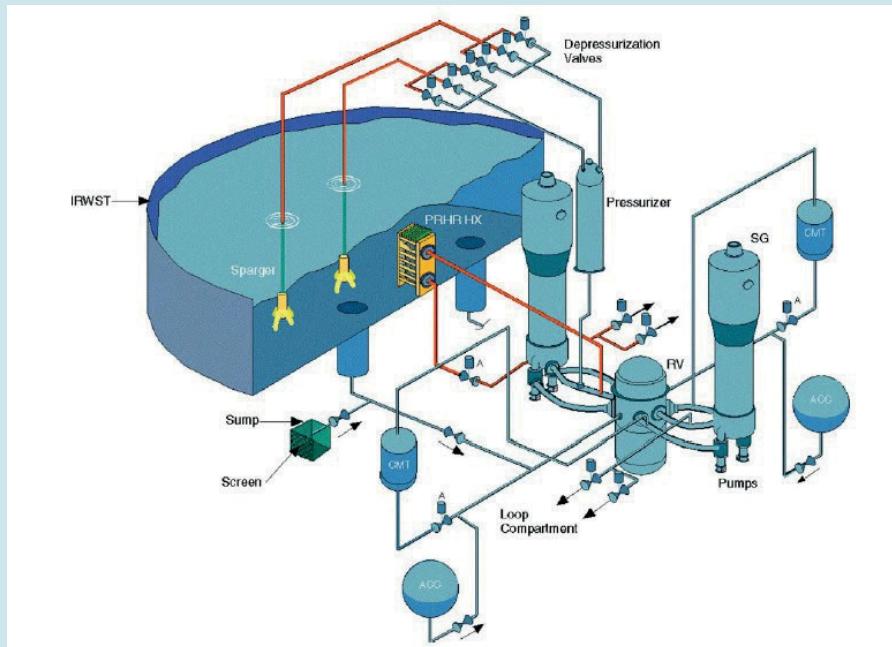
Figure 7-10: Passive core cooling system of the AP1000
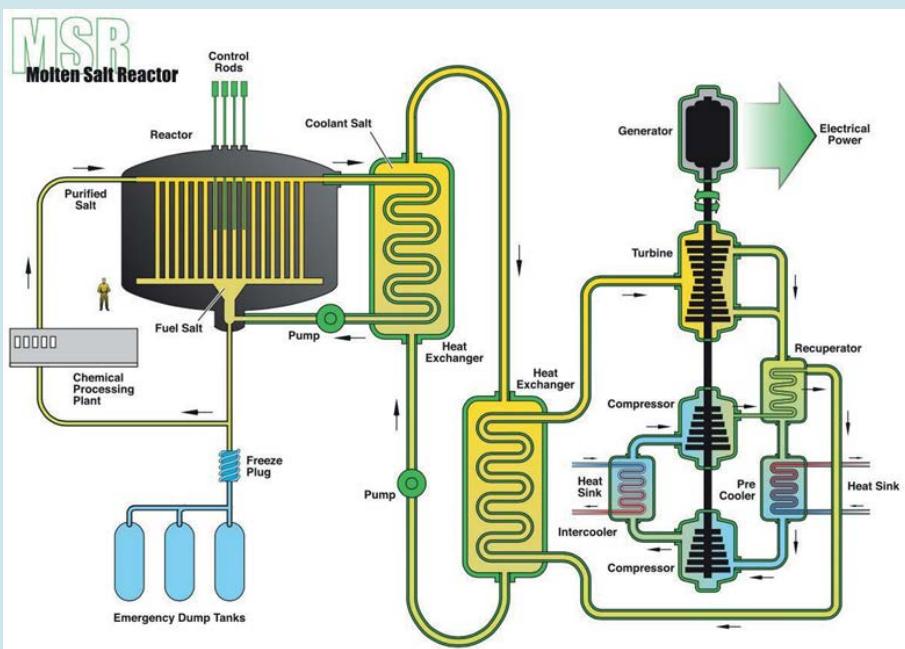
Figure 7-11: Molten Salt Reactor

Responsible compromises for nuclear power generation

As we have seen, each of these reactors (and their constituent innovations) help us realise certain values, while compromising other values. When designing reactors, we are designing for a variety of values. The table below shows a quick comparison.

	PBMR	GFR	MSR
Safety	++	-	0
Security	+	--	-
Resource durability	-	+	++
Economic viability	+	0	-

Table 7-2: Trade-offs in Reactor Design

The PBR is the safest choice, because it is physically incapable of a meltdown. Meanwhile, the MSR is optimal if we seek to maximize resource durability, since it offers the possibility of using thorium, which is found in greater abundance in nature than uranium.

We will attempt to summarize this long but insightful case study. First, we saw that sustainability is best understood in terms of several moral values - safety, security, economic viability, resource durability and environmental friendliness - and these values have both a spatial and temporal dimension.

When conceptualizing responsible innovation, we need to first have an ex-post analysis of what is at stake, ethically speaking. To that end, we discussed the technical intricacies of the nuclear fuel cycle. Naturally, this ex-post analysis is the first step towards an ex-ante analysis.

We argued that, before opting for a specific type of nuclear power generation method, we need to first assess each type of fuel cycle and understand the advantages and disadvantages of various reactor designs. Each of these reactors could help us realize certain values, but would inevitably compromise on some others. For example, the safest nuclear reactor is not necessarily the most efficient or sustainable one. These trade-offs need to be known and addressed in order to enable the responsible innovation of nuclear reactors.

Case study #9: When Big data meets Big brother

The text below is – besides on the weblecture based on an essay by Dirk Helbing and Jeroen van den Hoven titled: ' Digitale Digitale Demokratie statt Datendiktatur'.

Introduction
Through profiling and big data analysis, the internet knows us probably better than we do ourselves, including our weaknesses. Cookies collect all our clicks when we use the internet. Most of this information is collected without our knowledge and consent. It is practically impossible for the normal consumer to use the internet without being digitally exposed. By now, much more information has been collected about each of us than the Stasi or secret services of totalitarian states ever could. Our privacy has become a commodity. How long can this continue?

Experts agree that the technical possibilities of combining big data, artificial intelligence, smart devices, the internet of things and quantum computers have now exceeded the scenarios described in *George Orwell's 1984* and *Aldous Huxley's Brave New World*.

What are the consequences of extended government, corporate and other organizational access to knowledge about individuals and the possibilities to prediction human behavior? That's what this case study will explore.

Figure 7.11 : How safe is our privacy and free will

Examples
Let's have a look at some - rather worrying – examples:

- ***Example 1: Citizen score***
Imagine that a country has a citizen score. The system is considered a form of mass surveillance, which uses big data analysis technology. The citizen score is a number measuring the "value" or "usefulness" of a citizen from the point of view of whoever rules the country. In the future, it will decide which products and services each citizen is entitled to. According to the system, every citizen gets plus or minus points for everything they do. To determine the citizen score, an algorithm analyses your clicks on the internet, including the videos you watch and the music you listen to, or which links you click. Whether your opinion is consistent with the government position determines whether you get a plus or minus point. The system promotes citizens to police others, as the behavior of your friends and neighbors influences your own citizen score. What you do, what you buy, all of it goes into your citizen score and determines what conditions you get for a loan, whether you can get a certain job or not, travel to certain areas or not. Importantly, in case of resource shortages, the citizen score will decide who will receive what kinds of resources and services, including access to health care and energy, and who won't. For some, this may be very bad news.

  Such a system is pretty invasive: big brother is watching you. Yet, it is already being tested in China, but not only there. A similar "Karma Police" program was revealed in Great Britain as well, run by the secret service. The system is ready to use. During the next disaster or crisis in a certain region , the system could be switched on and turn into a totalitarian system that may persist for decades. Self-determination, democracy and human rights would largely be lost.

- ***Example 2: Ignorance and fake news***
Several people – including the previous president of the European Parliament, Martin Schulz – have demanded that we fight against technological totalitarianism. Former US President Barack Obama also warned us that liberal democracies are under attack by forces that ignore science and facts – forces so powerful that we cannot counter them. Public opinion is now increasingly controlled by intelligent computer programs such as social bots. Thus, we may have arrived in a "post-fact society" in which fake news increasingly determines public discourse. Frank-Walter Steinmeier, former Minister of Foreign Affairs in Germany, called the "post-fact society" created by modern manipulation techniques a "lethal danger" for democracy. And Elon Musk called artificial intelligence perhaps the greatest threat to humanity, possibly more dangerous than nuclear bombs.

- **Example 3: Predictable policing**
Already, **some cities** are using predictive policing. Basically, this means that people can be arrested and even put in jail before they have committed a crime. The amount of time they have to spend in jail depends on algorithms predicting what they might do in the future. Another example is a genetic

test, which yielded different outcomes depending on which company was doing this test. One algorithm might give a reliable prediction about future behavior, but another one might not. This is certainly a use of technology that we should be concerned about.

More concerns
Here are two more concerns:

- #### Cybersecurity
  In terms of security we need to worry about the fact that cybercrime is growing exponentially. **Cybersecurity Ventures** predicts that 'cybercrime damages will cost the world $6 trillion annually by 2021, up from $3 trillion in 2015'. Of course, these figures are not very accurate. Still, we have seen that the White House, the Pentagon and the US military have all been hacked, and the same goes for many big companies. Clearly, we don't have secure systems. Powerful ICT, such as super intelligence systems, could most likely still be hacked by organized criminals, terrorists or extremists, or be abused by dictatorial political powers. In fact, worries about security are increasing even more with the growth of the internet of things.

- #### What to do with all the data?
  Even now, computer power is exponentially increasing. The amount of data produced is increasing even faster. In just one year, we produce as much data as in all the years before, in the entire history of humanity. It is hard to imagine, and even more difficult to understand what the implications are. But it means that there is an increasing amount of data that we will never be able to process or look at. That means we need science to decide which part of all these data actually deserves and needs our attention. Now, there is another important element to this: the degree to which the world is interconnected is increasing the interdependency and complexity of our world exponentially. That means that even the enormous amount of data cannot catch up with the pace at which complexity is increasing. Paradoxically, even though we have more data and more processing power than ever, we run the risk of losing the ability to control these systems top-down.

Conclusion: Let's think twice!
This brings our society to a critical point. We are at crossroads and we need to make up our minds on where we want to go as a society in the future. What we need is to reflect on a new control paradigm, which consists of distributed control, with checks and interventions at various places. So, it's time to say: stop, let's think twice. We need to ensure democratic control of these technologies. We need to create scientific and interdisciplinary points of views on these technologies. We need to ensure their ethical use, we must safeguard transparency and accountability and we should compensate potential victims.

To summarize: we need to learn to use these technologies in the right way.

8. Risk management and safety engineering

8.1 Introduction

In the previous chapter, we learnt about risks and how to anticipate possible risks before or during the design of new technologies. But it is also necessary to manage risks for presently deployed technologies, on an ongoing basis. Moreover, we need to be able to choose which new technology to develop, given multiple alternatives which each preserve and compromise different values to various degrees.

Let's take the value of safety. Abstractly, this a value we hold very dear and do not want to compromise on. The benefits of safe technology are clear: we effectively minimize the risk of failure and damage to people and property. However, safety would necessitate some costs as well - both the opportunity cost of implementing safety features, as well as spill over effects, from deprived opportunity costs to the costs of implement features that uphold other values as well. Similarly, any benefits from implementing safety features would be hypothetical savings in the scenario of an accident occurring. So, research on responsible innovation should include a proper economic evaluation of the safety aspects related to new technologies, quantifying the net benefits and costs derived from pursuing one design over another, while also quantifying the risks that come with each option.

When considering large projects, risk and safety very quickly take centre stage. Risk and safety should be quite familiar words by now, but especially for this context, we need to define them precisely and consider them closely to make them useful to us.

Let us thus start with some definitions, perhaps revisiting some terms we have already seen in the last few sections. Our guiding questions for the moment are: what is risk? What is safety? And very briefly: what is security? We will not go into the latter concept in much detail, as it closely resembles safety for most intents and purposes.

8.2 Definitions

Definition Risks

Risk can be defined as 'the probability of something happening multiplied by the resulting cost or benefit if it does'. Note that this definition is neutral as to our judgement of the actual outcome, whereas risk is normally used only in relation to negative outcomes, and outcomes we want to avoid.

We can put this definition into a simple formula, the risk triplet. The risk triplet is the set of three questions used to define risk:

a) what can go wrong?
b) How likely is it, (probability)?
c) What are the consequences

There are, of course, different types of risk and therefore we can use two dimensions of risk - probability and consequence - to make some distinctions. We can describe risks as having a small probability and small consequences - like bee-stings or being struck by lightning - or we can have risks with a large probability but also small consequences, such as traffic accidents, falls from various heights etc. (We have left out the third dimension of risk, the scenario, for now.) Various consequences can be caused in many different ways. It is therefore important that when we talk about a consequence, we also specify how this consequence could happen.

Definition Safety

Safety is defined as a state: the state of being safe, free from hurt or injury. This is not an objective state, as people also need to feel safe.

Because it is a state, the sense of feeling safe can change drastically from one moment to the next. As we will also see, safety might conflict with other needs or interests, like economic considerations.

Safety only has meaning in the presence of threats. We will call these threats hazards, and they are defined as 'a situation that poses a level of threat to life, health, property, or environment'. Previously, we defined risk using the risk triplet of scenario (what can go wrong), probability (how likely is it) and consequence (what is the outcome). Hazards can also be seen as part of this triplet, the scenario and the consequence. Examples of hazards are: driving a car, running a chemical plant or a nuclear reactor, and flying an airplane. The latter is both a hazard for the people in the airplane as well as for the people and property on the ground.

Definition Security

So, how about security? The difference between safety and security lies in the intention behind the act. In case of safety, the focus is on plausible scenarios and a set of control measures. With security, however, the focus is on intentional actions aimed at creating large consequences.

How can we analyse Risks, Safety and Security? In this chapter we will discuss various ways to do We will start with the more general Cost-benefit analyses and then move to some concrete and powerful instruments and tools like, for example Bow-tie and Fault-tree

8.3 Cost-benefit analysis

Introduction

A cost-benefit analysis (CBA) is an *economic* evaluation in which all costs and consequences of a certain decision are expressed in the same units, usually money. A cost-benefit analysis cannot demonstrate whether one safety investment is intrinsically better than another. Nevertheless, a cost-benefit analysis allows decision makers to improve their decisions, by adding appropriate information on the costs and benefits of various prevention or mitigation investment options. Given the fixed or limited resources that are available to achieve multiple goals (and values), cost-benefit analyses can be very useful to determine which of the different options for investment represent the most efficient use of resources.

Anticipating various types of incidents and events

Decisions may be straightforward in some cases, but this is not always the case. For example, there may be very many types of unwanted events.

- Type I unwanted events can be regarded as 'occupational accidents' - for example, accidents resulting in the inability to work for several days or accidents requiring first aid.
- Type II events can be categorized as 'major accidents', involving multiple fatalities or huge economic losses. Type II events are thus surrounded with more uncertainty.
- Type III events can be regarded as so-called 'black swans'. For type III events, there is no information available whatsoever, and so an economic analysis cannot be carried out for such an event.

The economic considerations differ between Type I and Type II events. Specifically for the latter, a disproportion factor can be used, as we will see.

Net Present Value

One important concept in CBA is Net Present Value (NPV). A safety-related investment project represents an allocation of means and resources - such as money or time - in the present, which will result in a particular stream of hypothetical benefits in the future. The main purpose of a safety CBA is to obtain relevant information about the level and distribution of benefits and the costs of safety. With this information as a guide, a safety-related investment decision can be made in a more objective way.

The role of the analysis is therefore to provide the possibility of a more objective evaluation, but not to advocate either in favour or against any one safety investment, as there are many other aspects that should also be taken into account when deciding - such as social acceptability, ethical issues and regulatory affairs. If a decision maker decides to use a cost-benefit analysis, the recommendation whether to accept or to reject an investment project is based on the following process:

- Identification of costs and benefits
- Calculation of the present values of all costs and benefits
- Comparison of the present values of total costs and total benefits, thus determining the NPV

In order to compare the total costs and total benefits, composed in turn of the costs and benefits that may be incurred at different points in time, we need to use a discount rate in the calculations in order to represent the real present values. Essentially, we are converting all cash flows, including both costs and benefits that may occur in the future, to values in the present. The discount rate thus represents the rate at which we are willing to give up consumption in the

present, in exchange for additional consumption in the future. The higher the discount rate, the lower the present values of future cash flows.

The formula usually mentioned to calculate the NPV is as follows:

$$NPV = \sum_{t=0}^{T} \frac{X_t}{(1+r)^t}$$

Where X_t represents the cash flow in year t, T is the time period considered (usually expressed in years), and r is the discount rate.

NPV calculations are useful because people value (abstract) future experiences to a much lesser degree than (tangible) present ones, since they are more certain about present events than about future events. An investment project can be recommended when the total NPV of all cash flows is positive.

Applied to safety, the NPV of a safety investment expresses the difference between the total discounted present value of the benefits and the total discounted present value of the costs. A positive NPV for a certain safety investment indicates that the project benefits are larger than its costs, at least under the current set of assumptions.

Costs and benefits of safety measures

One can distinguish a great variety of costs associated with safety investments. We may conveniently classify them into a few clear categories, such as initial costs, installation costs, operating costs, maintenance costs, inspection costs, etc. These costs are evidently represented by negative cash flows. Some costs (e.g. initial costs and installation costs) occur in the present and therefore do not have to be discounted, while other costs (e.g., operating, maintenance and inspection costs) occur throughout the whole remaining lifetime of the facility and therefore will have to be discounted to the present.

Similarly, there are different categories of benefits linked to safety investments. But how can we interpret the benefits? We might say that the purpose of safety investments is to reduce the risk of present and future accidents. This means that the benefits are hypothetical, since the accidents - or rather, the accident scenarios – in most cases will never actually occur. They are defined therefore by the difference in consequences with and without a particular safety investment, and, if applicable, taking into account the difference in the likelihood as well.

Since there are usually a large number of accident scenarios that may be avoided, the hypothetical benefits will be much larger than the costs when calculating for any one accident scenario. One way to look at this is that only the most probable accident scenario will happen in reality, but many more would have been avoided.

The benefits as such represent positive cash flows, which all occur in the future and therefore will all have to be discounted to the present. As with the costs, we may also conveniently classify the benefits into a few clear categories. Hypothetical benefits, or avoided accident costs, can be as diverse as supply chain benefits, damage benefits, legal benefits, insurance benefits, human and environmental benefits, intervention benefits and reputation benefits, among many others.

Disproportion factor

Finally, let us look at the disproportion factor. The cash flows, prevention costs and certainly the hypothetical benefits may all be quite uncertain. Various approaches can be used to deal with this fact. For instance, cash flows can be expressed as expected values, taking the uncertainties

into consideration in the form of probabilities; alternatively, we may increase the discount rate to outweigh the possibilities of unfavourable outcomes. This is possible for uncertain and severe Type I risks.

Type II risks - or major accident risks - however, occur at extremely low frequencies and have a high level of uncertainty. To take this into account, the cost-benefit analysis preferably involves a disproportion factor in order to reflect an intentional bias in favour of safety over costs. In case of Type II risks, we can use scenario analyses, essentially estimating cash flows for different scenario cases. For example, we can consider the worst case and/or most credible case scenarios, and use the disproportion factor accordingly. If this equation yields a negative NPV value, then we can say that the safety investment under consideration is not reasonably practicable, as the costs of the safety measure are disproportionate to its hypothetical benefits.

In order to give an idea about the ideal size of the disproportion factor, guidelines state that disproportion factors are rarely greater than 10, and that the higher the risk, the higher should be the factor, so as to stress the magnitude of those risks in the CBA. This means that in cases where the risk is very high, it might be acceptable to use a disproportion factor greater than 10.

8.4 Quantifying and comparing risks

One of the problems with quantifying risks is that they do not have a common denominator. This is important, because if we are not careful, we would be comparing apples with bananas. We should therefore be cautious when we see risk quantifications in different guises. To compare the risks of different activities, we have to always make sure that we are using the same measurement units.

Let us see why this does not work. In the following example, we have some numbers, but we are not sure what they stand for. We can assume they express the overall probability that a person dies of this particular cause during his or her life. For instance, the probability of dying of smoking is 5×10^{-3}. This means that five out of a thousand smokers die because of smoking. Dying in traffic is also possible with a certain probability, as is dying because of a stroke of lightning. Dying because of a chemical accident is the least probable. Let us also add the probabilities of winning some kind of lottery. Actually, winning a lottery is not very probable; in fact, it is actually more probable to die from a bee-sting than to win the biggest prize in the national lottery.

The question is often asked why we use the number of people killed as the unit of measure. The blunt reason is because people agree on when somebody is dead, but they have difficulty agreeing on different degrees of injury. Over the years, the number of dead has proven to be a good proxy for total damage. However, it is not a good proxy for disasters where the number of wounded and the extent of material damage are much more important parameters.

Performing risk analysis
We now turn to the topic of risk analysis. Of course, the main questions for risk analysis are: what can go wrong, and how? (Each answer, and there could be more than one, is a risk scenario.) We can also ask: what is the likelihood of that scenario happening, the probability? And finally, what would be the consequences?

With regard to risk analysis, there are two approaches, the deterministic and the probabilistic approach. The deterministic approach is often limited to preventing the maximum credible accident, like the exposure of the core of a nuclear reactor, or a truck containing a hazardous chemical running into a building. These are things we just don't want to happen, so we go to great lengths to prevent them. In a probabilistic approach, on the other hand, we consider the probabilities of particular accidents.

A risk analysis is part of what is called a risk-based decision analysis. It is shown in the diagram below.
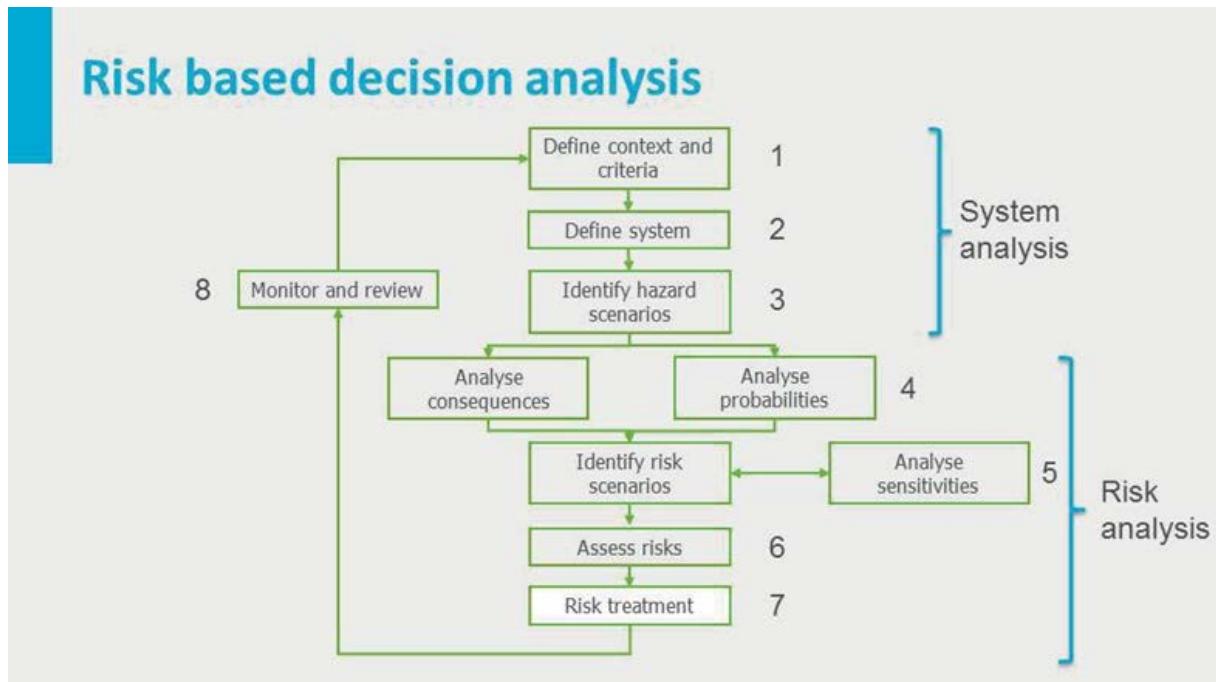
Figure 8-1: Risk-based decision analysis

Risk-based decision analysis consists of eight steps. In the first few steps, we define the context and the system we want to control for its hazards; in the later steps, we actually carry out the risk analysis.

First, we define the context and the criteria for the risk analysis. Why do it in the first place? Maybe we are considering a new technology, and we want to know whether it is acceptable. Or we have decided to introduce a new technology, but we want to know where the risks are. We may also want to know who we need to involve, perhaps in order to control the risks, or whom to convince about the choice of this new technology. Whatever we do, we need to establish the criteria on which we will base our decisions. This looks straightforward, but in practice it is quite complicated.

In the Netherlands, and perhaps in many other countries too, we make a distinction between internal and external safety. Internal safety basically means occupational safety - the safety of the people at work in the plant or in the field. External safety means everything around the plant, reactor or activity. For now, we will only focus on external safety. This is an arbitrary decision, as occupational safety is equally important, given the many occupational accidents that occur each year. However, these are more likely to be specific to each industry, and as such, a comprehensive discussion of them is beyond the scope of this book.

In external safety various types of risks may be defined. The first is individual risk, later redefined as localized risk. This is the probability that one person is killed in a specific year at a particular place because of some hazardous activity. The second is group or societal risk, which is about a particular number of people killed per year with a certain probability. Finally, there is the expectation value, that is the average number of people killed per year.

Risk contours

The probability that one or more people in one particular year will be killed can be placed on what we call risk contours. These contours, seen on a map (see Figure 8-2) surround a potential hazardous place - say, a chemical plant or a nuclear power plant.
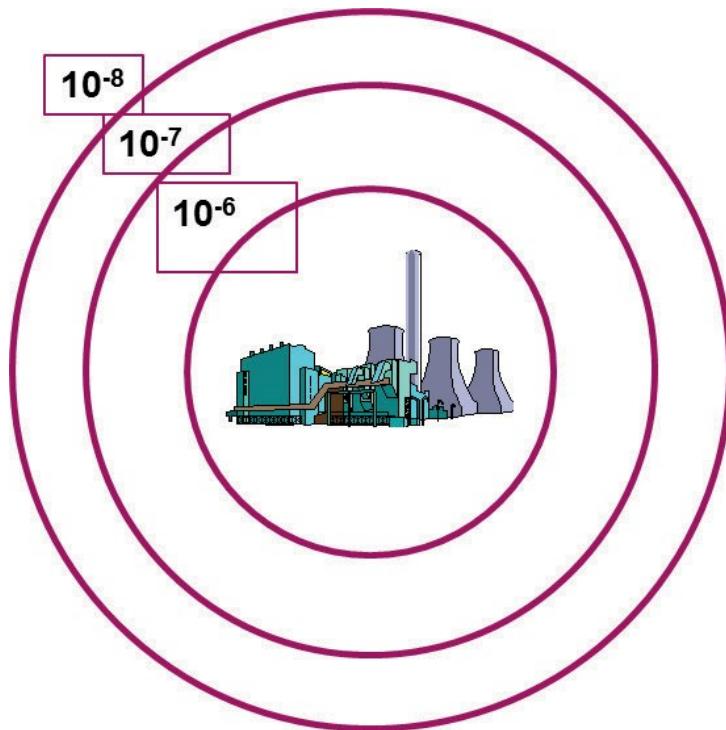
Figure 8-2: Risk contours

All points on the same contour line have the same likelihood. It is customary to draw these lines according to (negative) exponentials of 10, like 5×10^{-5}, 5×10^{-6}, and so on.

Figure 8-3: FN-curve

Group risk on the other hand is usually represented by a graph, an FN-3 curve, in which the frequency (in years) is plotted against the number of fatalities. In the graph below, we see a particular activity, for instance the activity of one particular chemical installation, and the frequency with which it will demand a certain number of casualties.
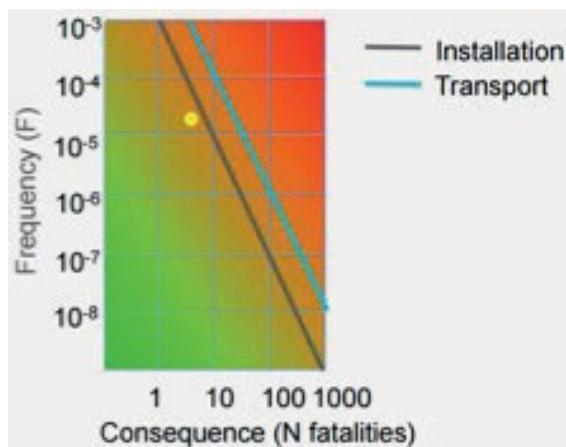
For this particular installation, for example, the accident frequency is about once every 80.000 years, and we can expect between 6 or 7 casualties per accident. In the graph, two lines represent an agreement between the parties involved, defining how many casualties are 'reasonably permitted' and how often these casualties could occur. We can see that the installation lies below this norm.

An interesting phenomenon can occur, which makes clear that there is a certain tension between the concepts of localized risk and group risk. Say the situation is as shown in the picture below.
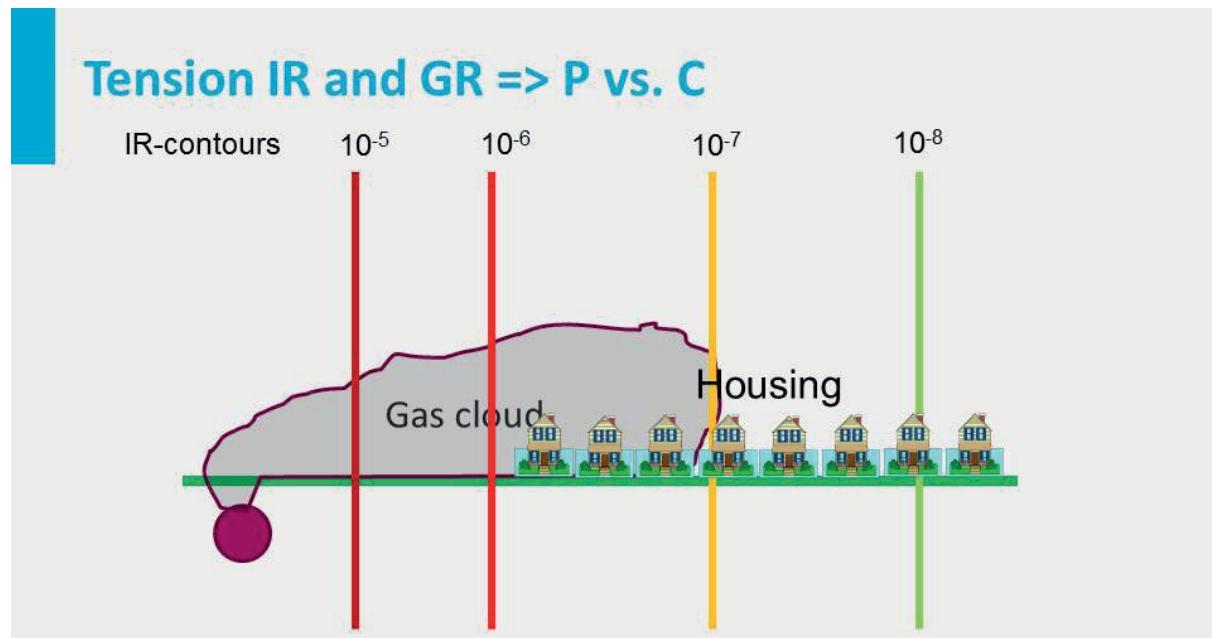
Figure 8-4: Tension between localized risk and group risk

Houses are built beyond the risk contour of 5×10^{-6}. So, once every 10 million years, a certain number of people might die because of a gas cloud escaping the chemical plant. As you can see, the houses are built beyond the risk contour of one million years, so the parties have agreed that this is an acceptable risk.

Now, the plant decides to take measures to make it even safer. Thanks to the new measures, and the agreements made based on the risk contours, houses can now be built much closer to the chemical plant. However, if and when a gas cloud would escape, we can expect far more casualties. So, localized risk and group risk can be at odds, and we have to agree which one we will focus on or prioritize.

This is also shown in the following image, where the yellow dot for the installation moves beyond the agreed lines. This obviously represents an unacceptable risk, based on the previous group risk agreements.

On

Figure 8.5 : Groups Risk

Figure 8-6: Trajectory of the plane that caused the Bijlmer disaster
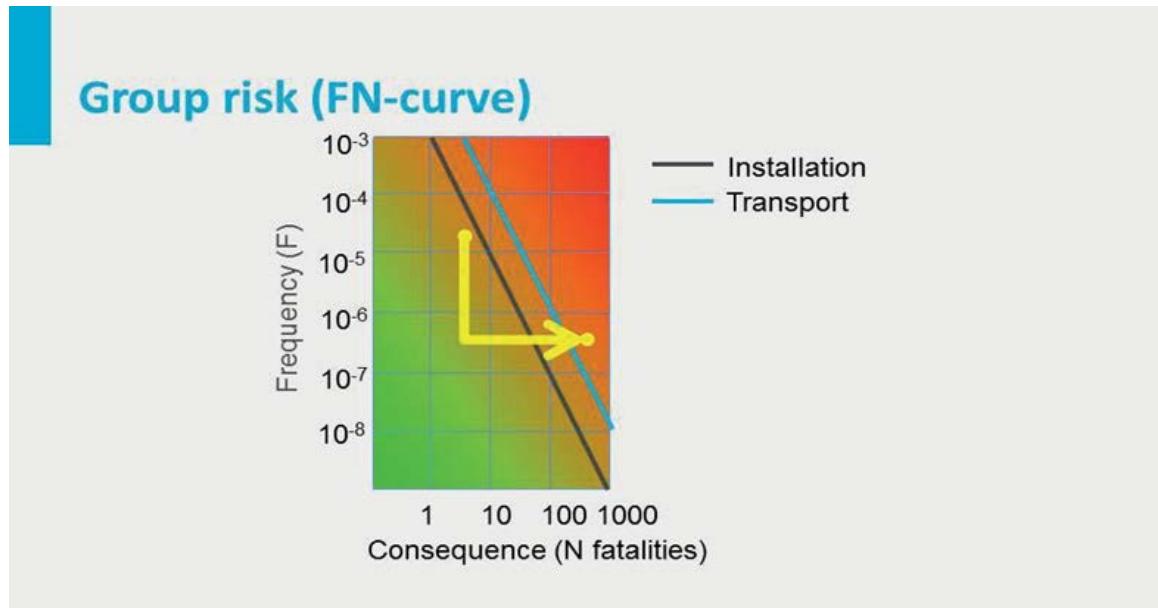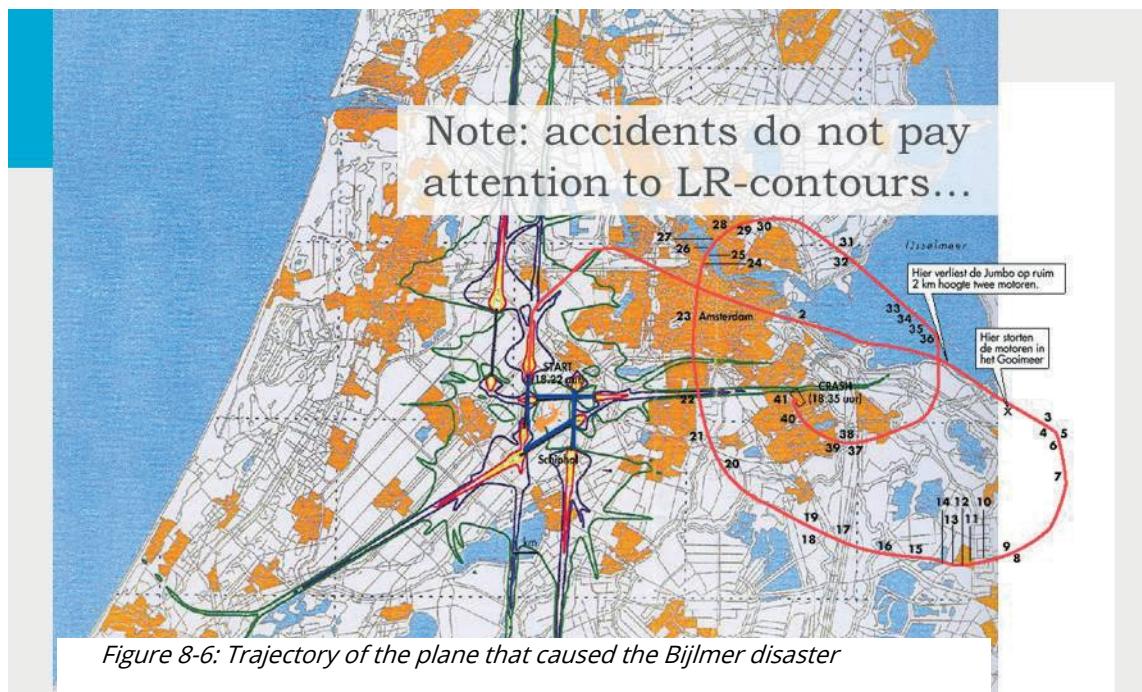
October 4, 1992, a Boeing aircraft crashed into an apartment building in the Bijlmer neighbourhood of Amsterdam.

You can see its trajectory in red in the image below. The green contours are the low, localized risk contours - the ones where a fatality is expected very rarely. As you can see, the Boeing crashed exactly in such a green area.

Defining the system and boundaries

After we have defined our context and criteria, we have to define what the system consists of and which hazards we want to control. We need to think hard about the boundaries of the system we are looking at. Are we looking at a particular installation at a particular plant? Are we looking at the plant alone, or are we looking at an entire work site with many different plants? Are we looking at internal safety - the risks the workers are exposed to - or external safety, which covers everything outside of the system? And what is the level of detail we want to consider? Are we looking at each pipe, vessel and shutter, or are we only looking out for particular hazards and specific activities?

Defining the system and its boundaries is of practical importance. In a way, we can understand events that occur within the system boundaries as outcomes we can prevent or control, whereas events that occur outside system boundaries should be seen as outcomes we can only anticipate and manage but cannot directly control - either because they are beyond our sphere of control, or because we have only limited resources at our disposal. Furthermore, we have to be explicit about what we consider within the scope of the analysis and what is not considered, so that all stakeholders know exactly what the analysis covers.

Hazard analysis

When we have defined the context and the system, we can think about the hazards. We have three questions to guide us: what can go wrong; how might this it happen; and what measures or controls do we have to contain the hazard? We thus gain a more detailed understanding of the system we are looking at. Only after we have exhaustively worked on this step do we consider the identified hazards. Needless to say, this step is crucial for the rest of the analysis and for the validity of the whole exercise.
The following methods may be used for the identification of hazards:

- Standard list or checklist Preliminary Hazard Analysis (PHA)
- Hazard Identification study (HAZID)
- Hazard and Operability study (HAZOP)
- Failure Mode and Effect Analysis (FMEA)
- Failure Mode Effect and Criticality Analysis (FMECA)
- Fault Tree Analysis (FTA)
- Past experience (incident, accident reports or databases).

Some of these acronyms occur frequently in hazard studies. Each method has its own benefits and drawbacks. *Here, we will focus specifically on Fault Tree Analysis and the Bow Tie Model.*
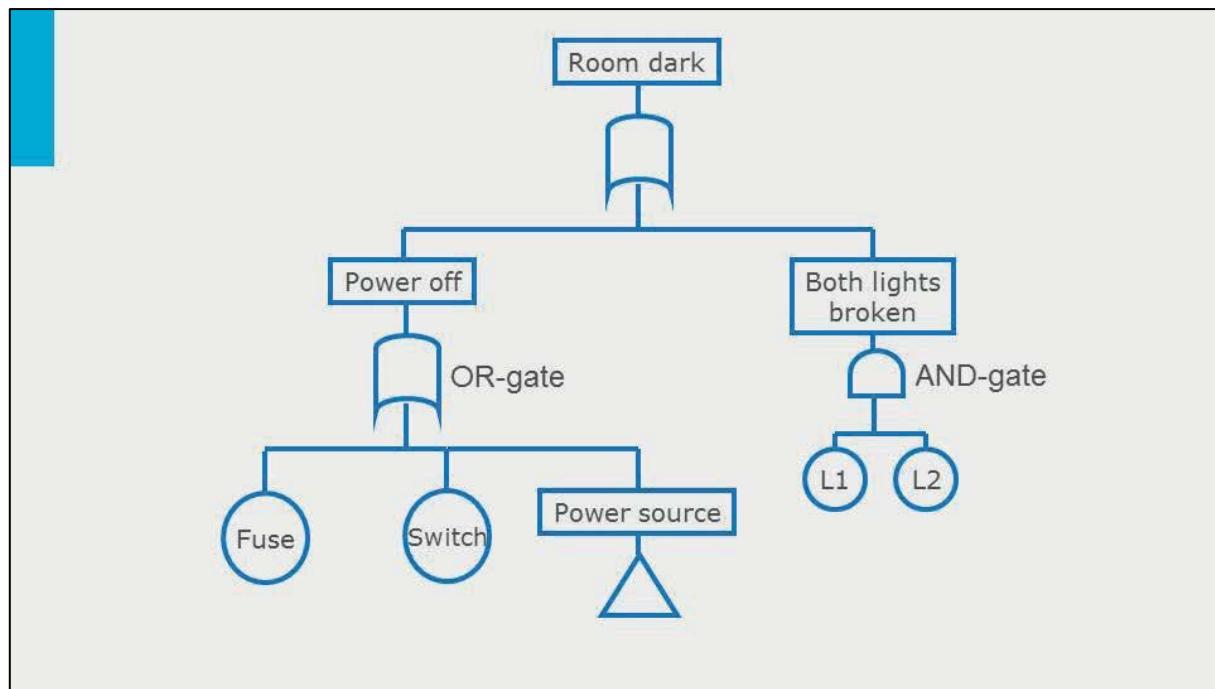
Fault Tree Analysis

A Fault Tree Analysis is a logical structuring of events leading to the top event, the outcome that is to be avoided as much as possible. Because of its logical structure, we can use fault trees to quantify risks.

Although it has one particular event as its top event, we can also use the Fault Tree for events that have not happened yet - that is to say, in a prospective way.

In a Fault Tree Analysis, we start at the top event and work our way down the tree systematically until the point where we decide to stop. Theoretically, however, Fault Trees can go on without end.

Let us take as an example a room that has two light fixtures in it. We press both light switches,

Figure 8-7: Fault Tree Analysis

but the lights don't switch on. We define our top event as a dark room, and we can think of two reasons why this is the case: either there is a power failure or both lights are faulty.

Notice the little symbol that connects the two sub-events to the top event. This is called a gate, and in this particular instance it is an OR-gate. An OR-gate implies that only one of the sub-events has to occur to trigger the top event to happen. There can also be an AND-gate, depicted with a flat base. An AND-gate would mean that both sub-events have to occur for the top event to trigger.

Let's now take a closer look at the system itself, as discussed previously. We see a simple circuit, with a power source, a fuse, a light switch and two light fixtures. We already have tried the switch, but the lights did not go on. We can now finish our little fault tree. Because both lights did not go on, they must be both broken, hence we add the AND-gate. However, it could also be a problem with the power supply. In this system, we have three potential origins of failure, namely the fuse, the switch and the source, which may all be faulty. On the other hand, a fault in a single element is also sufficient to trigger the outcome of a dark room. Therefore, we add an OR-gate.

This small example makes a few points very clear. We first need to define the context and the system clearly. We could have included the power supply of the entire street or neighborhood. We could also include many other ways in which the power supply can fail. Essentially, we have to choose system boundaries depending on what we can effectively prevent or control, and only adapt to or manage events that fall outside the system boundaries. These decisions are essential for our analysis.

Finally, we can add probabilities to the tree, if we know them. What is the probability that a fuse burns out or that a lightbulb fails? There are generic industry tables for these figures. By using specific calculation methods, we can calculate the probability of entering a dark room, if it has

two light fixtures. The Fault Tree also explicitly visualizes the various scenarios, which can be understood as individual paths through the tree.

Bow-Tie Diagram
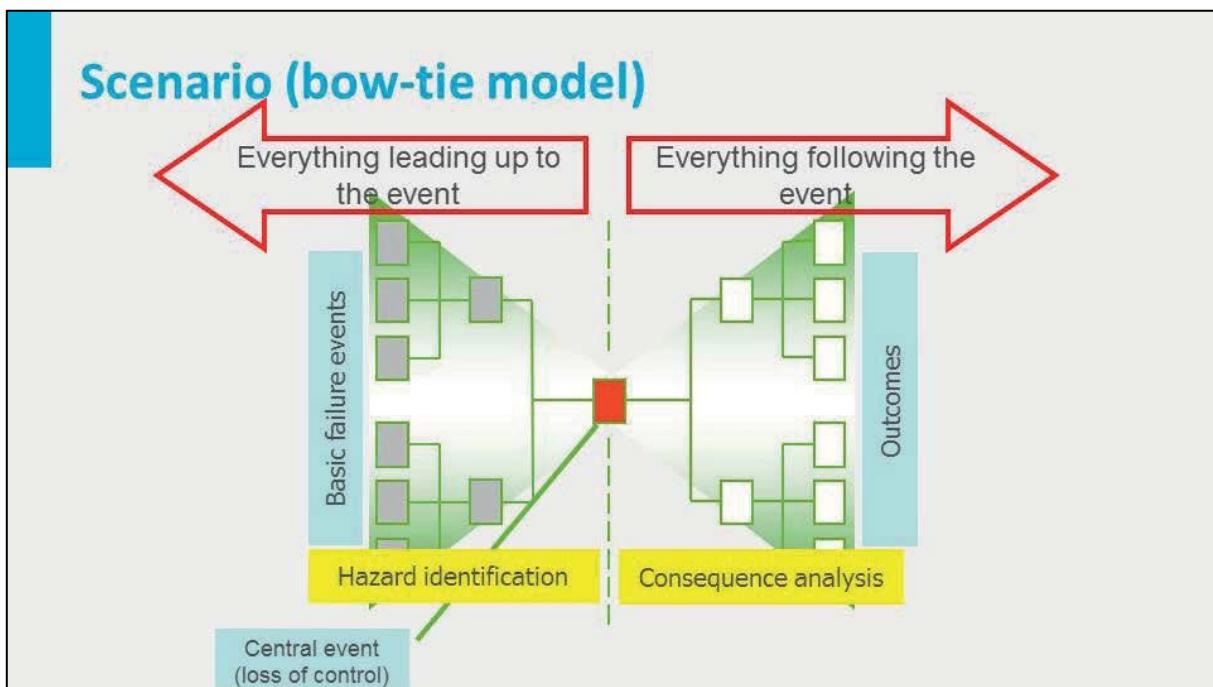
Figure 8-8: Bow-Tie diagram

The Bow-Tie method is a risk evaluation method that can be used to analyze and demonstrate causal relationships in high-risk scenarios.

The method takes its name from the shape of the diagram that is created, which looks like a bowtie. As is clear from the diagram below (figure 36), the fault tree covers the left side of the bowtie in a different shape.

Notably, a Bow-Tie diagram only contains OR-gates, and is mostly used in a qualitative way, with no probabilities added.

The central event is often defined as the point of loss of control. On the left-hand side, events are ordered in such a way that they represent the failures leading up to the central event. On the right-hand side, events are ordered depicting the outcomes of the top event and related cascade events. Put simply, the left-hand side presents the causes of failure and the right-hand side depicts the consequences.

Consequence analysis
In the next step of our risk analysis, we look at the consequences of a scenario. We often define consequences in terms of fatalities, injuries or money.
We first have to agree on a common denominator, otherwise we would be comparing different things. Moreover, what is our time frame? Large accidents bring about consequences that extend

far into the future. Fatalities may often have a large impact on families and companies.

How do we account for this? To answer these questions, one needs expertise pertaining to the domain under study.

	Minor	Critical	Severe	Catastrophic
Plant damage and lost production	Short term loss of production	Damage to machines, reparable in a short term	Major repair costs. Serious loss of production	Substantial damage to plant
Environment damage	Temporary excursion in emission levels	Significant release, Clean up required	Ecological damage for up to 1 year	Ecological damage for more than 1 year
Harm to Personnel	Non-disabling injuries	Disabling or severe injury	Critical injuries	One or more fatalities

Table 8-1: Example of different loss categories

If we want to make the analysis a bit easier, we can classify consequences into a limited number of categories. This is what companies often do. In the particular example above, four loss categories are defined - minor, critical, severe and catastrophic - for three different target groups: the plant, the environment and plant personnel. For each category, a short description is given. Many such matrices can be used to frame discussions about losses and target groups. Note also that there may be many scenarios associated with these consequences.

After we have listed the possible consequences - which must be exhaustive, at least within the scope of the system that is being considered - we have to assign probabilities to these outcomes. This is not a necessary step and we only do this when we can or should quantify certain outcomes.

Let us look at three examples of how to express probabilities. As has been said before, it should be clear and agreed upon which denominator will be used in order to compare the consequences. It is similarly important to be clear about the level of detail of the analysis, and the type of consequences we want to quantify.

Anticipating risk scenarios
In the next step of risk analysis, we identify risk scenarios. This step is predominantly about asking a large variety of different questions, in order to identify as many weak spots and potential outcomes as possible.

Specifically, we quantify and rank the risk scenarios in terms of probability. We can then decide which scenarios we want to approach in a deterministic manner: prevent them all together or in a probabilistic manner.

We may also consider doing a sensitivity analysis, to see if our analysis makes sense and identifies weak links. We can take a closer look at our data sources. For instance, we can ask if we

have taken into account the variability present in our system. Did we capture all factors, interactions and relations? How about the quality of the data we have used?

We can also look more closely at our analysis. What happens if our input variables change? Or if the relationships are re-modelled? In each model, we implicitly make assumptions, and we need to know what happens if our assumptions are wrong. How does the risk increase in that case?

Risk assessment

The sixth step in performing a risk analysis is the risk assessment. Again, we can make this simpler if we work with classes and categories. This is not a requirement, of course, and all stakeholders should agree on these classifications. We can combine the frequencies and outcomes into a matrix.

Frequency of occurrence	Hazard categories			
	I Catastrophic	II Critical	III Marginal	IV Negligible
(A) Frequent	1A	2A	3A	4A
(B) Probable	1B	2B	3B	4B
(C) Occasional	1C	2C	3C	4C
(D) Remote	1D	2D	3D	4D
(E) Improbable	1E	2E	3E	4E

Table 8-2: Risk assessment matrix

In this particular matrix, we distinguish five frequency categories and four outcome classes. We can now label each cell with a particular outcome and associate it with a certain frequency. There are also five classes of severity. The matrix often indicate what will be done about the risk. Some risks in this matrix are accepted, whereas others are prevented or insured.

We have four risk colours ranging from red, meaning that changes have to be made in the design before development continues, to dark green, which indicates risks considered negligible or easily handled with present measures. The most interesting ones are the colours in between, the light green and yellow colours. The risks calculated to be yellow or light green cells can decrease, depending on new insights and technology. Conversely, unforeseen events or new knowledge may make these risks more severe.

In deciding what colour each risk class will get, we often use the ALARA principle, which stands for "as low as reasonably achievable". It demarcates the border between what is acceptable or tolerable and what is unacceptable or intolerable. Red-coloured risks - that is to say, unacceptable risks - are approached deterministically, so they should be eliminated, prevented or well insured. What is tolerable is determined by considerations of costs and practicality. This can change of course, and will be subject to some negotiations.

Safety measures

Finally, we come to the treatment of the identified risks. Here, safety comes into full view. Basically, we have four possibilities: risk avoidance, risk reduction, risk transfer and risk acceptance. These measures are illustrated in the picture below, in which a person is threatened by a crocodile.

We can avoid the risk completely and kill the crocodile, we can keep a safe distance from the crocodile, we can put the crocodile in a cage, or we can put on protective clothing. Such a matrix is also called a **Haddon Matrix**.

environment	man	measure	effect	
		banish danger		Avoid
		separate person & hazard		Reduce probability
		shield danger		Reduce probability
		protection of person		Reduce consequence

Figure 8-9: Haddon-Matrix

William Haddon was a medical doctor who tried to think of all possible strategies and ranked them in order of effectiveness, eventually visualizing them in the form of a matrix. We find that elimination of hazards is the most effective, followed by minimizing exposure to hazard sources (reduce the probability of their occurrence). We can also try to prevent the release of hazards, which equally reduces the probability that they will occur; or we can modify the way in which they are released in order to minimize damage. Haddon defined ten different strategies, with the last few strategies pertaining to coping with consequences in a particular way.

Interestingly, we can also project Haddon's strategies onto the Bow-Tie model discussed earlier. Before control is lost, we have time to pursue prevention strategies; after control is lost, we shift to mitigation strategies in order to cope with the consequences. This means we can try to:

1. Avoid negative outcomes,
2. Reduce their likelihood,
3. Minimize their consequences,
4. Transfer risks - for instance by insuring them, or
5. Accept and live with the risks.

Another nice and simple model to help our thinking about risk measures is called the Hazard-Barrier-Target model. It describes the situation we often find ourselves in: there is a hazard that presents a possible threat to some kind of target, but which is protected by one or more barriers. Barriers can be of different kinds: physical, procedural, or a combination of both. These barriers

prevent an unwanted energy flow from reaching the target and prevent harm. In a bow-tie, barriers can be represented as blocked pathways or scenarios. Unfortunately, there will always be pathways that remain open. They are exposed when an accident happens or when a sabotage attempt succeeds. Therefore, the open pathways need to be monitored closely.

Risk analysis in practice

In the previous sections, we have seen the main steps of a risk analysis so far. Note, however, that this is not a linear process with a fixed endpoint. Risk analysis ideally never ends; it is a continuous process of anticipation, preparation and (pre-emptive) prevention or mitigation.

We need to be on the alert constantly. Systems keep changing, modifications are continuously being made, people and their practices change. We need to take these changes into account. Accidents inevitably happen, and we need to learn from them. Over time and with effort, our knowledge and inventory of means also increase, and thus - by the ALARA principle - our risk analysis and mitigation efforts will continuously change accordingly.

Part VI
Value Sensitive Design

Values

Environmental safety
Reliability

Norms

GMO should not leave the lab
Should be able to use and store

Design requirements

Cell fee system
Reliable detection
method
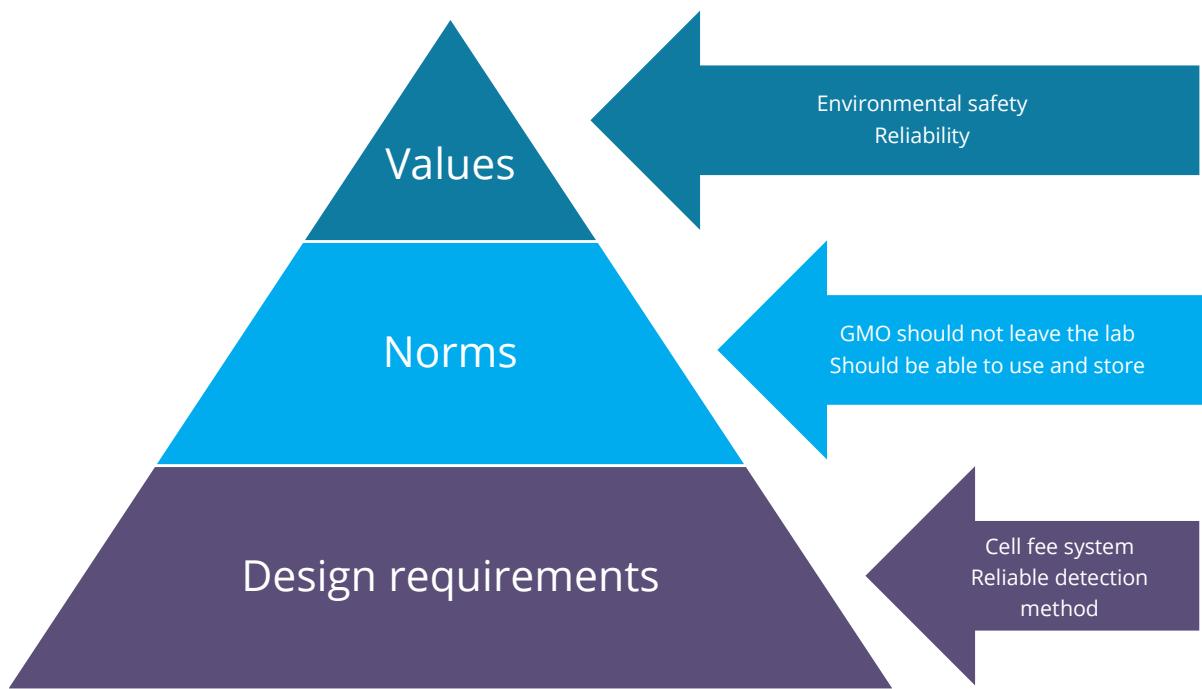
Figure 9.1: Value Hierarchy Matrix based on GMO

9. *Value Sensitive Design*

9.1 Introduction to Value Sensitive Design

Introduction to VSD

We have now come to the final part of our introduction to Responsible Innovation and will now focus on the relevance of Value Sensitive Design (VSD) for RI. VSD aims to provide a more actionable conception of how to take abstract moral values and shape them into tangible technical parameters in our technologies and innovations.

Please note that Design for VSD is also known as Design for Values "

Historically, VSD originated as a discipline within the computer sciences, even though the idea itself has a much wider purchase in technology. When the computer was first introduced around the middle of the 20th century, much of the scholarly attention was focused on the new technology itself. The computer was correctly seen as a general-purpose technology that could empower solutions to a wide range of problems across many disciplines. However, there was little attention for the social context and the users of computing machinery at this early stage.

In the second stage of the development of the computer - in the 70s and 80s - many started to realize that computers were being used in real-world organizations, supporting a multitude of users, each with specific needs and requirements, in different work environments and within a variety of social and institutional settings. Thus, the social and behavioural sciences became increasingly relevant for Information Technology (IT) applications, in the form of Human-Computer Interaction (HCI), Participatory Design and Social Informatics.

This shift of attention to the social context, usage patterns and user behaviours was at this point only motivated by attempts to identify potential barriers to the successful implementation of systems, to prevent failures and avoid failed investments. Still, it eventually led to the study of user-friendliness, usability and user acceptance.

We can discern the start of the third stage of development somewhere around the turn of the 21st century, when the successful applications of IT were increasingly understood to be dependent on their capacity to accommodate a broad range of human values, rather than just user-friendliness. Human beings, whether in their role as employers, consumers, citizens or patients, all have their own moral values, moral preferences and moral ideals. In every society, there are ongoing moral and public debates about values like equality, property, privacy, sustainability, autonomy and accountability, among many others. Even our computer networks and systems should accommodate these values in some way or form, whenever possible and appropriate.

In the last decade, values have emancipated from the status of mere constraints in implementation to constitutive aims and proactive driving factors in the development of IT. In California for instance, a Centre for Information Technology Research in the Interest of Society

(CITRIS) was founded in 2001. We seem to have entered a logical fourth stage in the development of IT, where the needs and values of human users - citizens, patients, consumers, decision-makers and so on - are considered as important in their own right. IT is conceived of as a technology to serve and support human beings, qua moral persons, in individual moral and social endeavours.

We have thus changed perspectives, from considering mere technicalities to framing technology in social contexts. Similarly, we have moved from seeing moral values just as constraints to abide by, to a more humanist vision of technology serving the needs and goals of society. This development is neatly summarized in the term "Value Sensitive Design", which has gained currency over the last decade.

The core idea of Value Sensitive Design is that moral values can be tangibly expressed in engineering terms, and that we can tangibly impart the fruits of ethical reflection - concerning sustainability, safety and privacy among others - to the things we design and make.

Converging lines of thought
A number of converging lines of thought and research have come together in the concept of VSD:

- ### 'Do artifacts have politics?'
First, an important step in this line of thinking was a seminal paper written by Langdon Winner in 1980. It was titled "Do artifacts have politics? (see also chapter 1) ", and it drew attention to the fact that values can be manifested in real-world objects and technologies, profoundly shaping the behavior of populations. His illuminating illustrations of how values and political views and power embedded in technology may shape and constrain the actions of people were very influential in thinking about the ethics of design of technology.

The example that captured everyone's imagination was that of New York's bridges having low-hanging overpasses. The famous architect and urban planner Robert Moses had designed the overpasses on New York parkways to be intentionally low, so that they were accessible for cars, but not buses. The socio-cultural impact of this was that the white middle-class population, who owned cars, could easily access Jones Beach on the other side, but people from poor black neighborhoods, who were more likely to take the bus, could not pass. Indirectly, the overpass functioned as a racist border-mechanism, separating the wealthy from the poor, the white population from the black population.

There has been some controversy about the historical accuracy of this case, but once we are introduced to this example, we immediately grasp the wider implications of how values can be ingrained into the things around us, profoundly yet invisibly shaping our lives.

- ### Science and Technology Studies
Other studies in the 80s looked into the philosophy and sociology of technology as well. This line of research was referred to as Science and Technology Studies (STS). This too revealed numerous examples and provided detailed case studies proving that socio-political biases (especially those concerning race, gender and income) could be inscribed in(to) technical artefacts, systems and infrastructures. Researchers like Geoff Bowker, Susan Leigh Starr and Lucy Suchman have contributed much to this body of work.

- ### Concerns by engineers
Some specialized areas of design and engineering also started to use this basic concept of Design for Values or VSD at around the same time. We can use the following two examples to illustrate   this.

First, let us look briefly at Privacy-Enhancing Technology. In the 80s, a number of privacy scholars started to work on ways to design IT systems and applications in such a way as to increase the likelihood that users would comply with privacy norms. Instead of relying only on the goodwill of users to comply with privacy regulations, the artefacts themselves would be designed in such a way that user compliance would naturally be within the desirable pathways.

The other example concerns Architecture and Built Environment Studies in the 80s. In architectural design and urban planning, steps were taken to pre-emptively design for security and against crime. Factors like lighting, variety in architecture, the spacing between buildings and lines of sight, among others, were all found to be influential in reducing crime rates. As such, these factors were carefully embedded into the design parameters of buildings and neighbourhoods.

9.2 Defining the method of Value Sensitive Design

The most clear and precise formulation of the VSD concept originated in a movement at Stanford in the 1970s-80s in the field of Computer Science, advocated strongly by *Terry Winograd*. It has now been adopted by many research groups and is often referred to as Value Sensitive Design (VSD).

VSD is an approach to systems development and software engineering which was developed in the last decade of the 20th century, by Batya Friedman et al. They built on insights from the human-computer interaction (HCI) community to draw attention to the social and moral dimensions of design. In VSD, the focus is on incorporating a wide range of human and moral values into the design of (information) technology.
For references see also this article

Even though VSD does not commit to a specific normative framework, according to Friedman, the practice is primarily concerned with values that center on human well-being, human dignity, justice, welfare, and human rights. VSD connects the people who design systems and interfaces with the people who think about and understand the values of the stakeholders who are affected by the systems.

To **quote** Friedman:: "*Ultimately, Value Sensitive Design requires that we broaden the goals and criteria for judging the quality of technological systems to include those that advance human values.*"

At TU Delft, we frame VSD as a way of applying ethics with the aim of making moral values part of the process within technological design, research and development.

The main methodological structure used by VSD initiatives is an integrative and iterative tri-partite methodology, consisting of conceptual, empirical and technical investigations (Each of the conceptual, empirical and technical investigations and analyses is carried out iteratively, mutually informing and being informed by the other investigations.

See **this** article: Value Sensitive Design and Information Systems by Batya Friedman, Atya Friedman, Peter H. Han and Alan Boring

Value Sensitive Design has a number of features that are aligned with responsible innovation. The values and moral concerns of all stakeholders need to be articulated at a point in time when

they can still make a difference to the design; they need to be formulated in such a way that they can inform the design; and the designs and artefacts need to be evaluated in terms of the values upheld and the moral concerns raised.

It should be clear that although VSD originated in the fields of IT and computer science, it has a much wider purchase and is relevant to all types of innovation and design of new technologies, as well as the diffusion and deployment of technological artefacts.

9.3 Applying VSD in practice

We have seen why the concept of VSD is important. In modern complex socio-technical systems, we are confronted with serious challenges. On the one hand, we all have values we hold dear as individuals and as a society. These are values such as safety, sustainability, justice, privacy, human well-being and so on. In the past, such values were mainly achieved and upheld through human behaviour and institutions like the law and government policies.

Increasingly, however, we live in a technological world, in which technologies shape how we live. We have only to think of how different present-day lifestyles to the lifestyles of generations that came before us, with ubiquitous technologies like the internet, computers and smartphones. The challenge we are confronted with is how to see to it that these technologies reflect and embody the values we hold dear.

We thus need to make a translation from the world of values and ideas to the world of technology and materiality - a translation that is hard to make, as these worlds have been very much separated in the past. Therefore, this is an opportune time to ask: *how can we embody values in design*?

Does technology embody values?
Let us start with the first question: does technology embody values, and if so, how? We can take three positions to answer this question: Instrumentalism, Substantivism and Interactionism.

- **Instrumentalism**
  Instrumentalism states that technology is value-free, because it is merely an instrument in the hands of human beings. Whether a technology serves or obstructs a certain value depends on how it is used. A bread knife can be used to cut bread, but also to kill someone. Instrumentalism is for example expressed in the slogan of the American National Rifle Association: "Guns don't kill people, people kill people."
  However, it is much easier to kill someone with a gun than without a gun; and when a burglar breaks into your house, you will probably behave differently with a gun at your disposal than without.

- **Substantivism**
  Substantivism takes the position that technology itself is value-laden and that humans have no influence on this. For example, it has been argued that technology embodies values like efficiency, or that technology inherently leads to environmental degradation, to a lack of authenticity, or even reduces human interactions to a minimum.
  A problem with this position that it overlooks the influence that people can have, both by using and designing technology.

- **Interactionism**
  The position we will defend here is therefore an interactionist position. It holds that value

is created and embedded in the interaction between human and technologies, both in how technologies are used and designed. In this book, we will focus especially on the design aspect.

What values should be included in technology design?

A first thing to note is that a wide range of values may be important in engineering design, and that we may derive these from a number of sources like the design brief (that states the motivation of project), the designers (and their professional communities), users and stakeholders, laws and government policies, technical codes and standards, codes of ethics and other moral concerns. Listing all these values, however, would not tell us which values to include, because this is a normative question - a question about what we should do.

Answering this question is complicated further by what we call value pluralism. There can be a plurality of values and people can reasonably disagree about which values are the most important. Obviously, value pluralism makes it harder to decide which values to include in a design. Still, it does not make it impossible, for a number of reasons:

- Firstly, despite value pluralism, there will often be agreement on at least some values that need to be integrated in the design of a technology.
- Second, value pluralism often means that people disagree about what values are the most important, but they may still agree on the broad spectrum of values which are the most relevant to take into account. For example, one may disagree whether safety or sustainability is most important in the design of a technology, but most people would agree that both safety and sustainability should somehow be incorporated in the design of say, a new car.
- Third, it may sometimes be possible to design technologies in such a way that they respect the different values of various groups and stakeholders.

9.4 How can we translate moral values into design specifications?
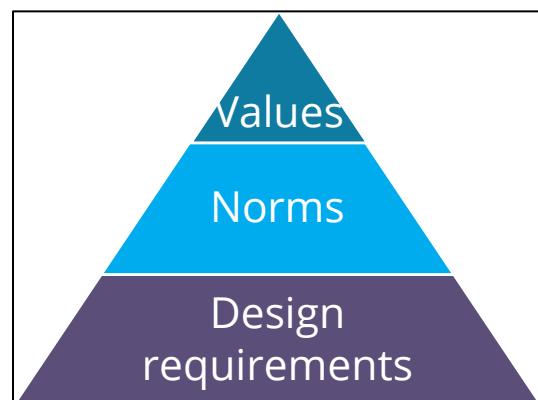
Instrumental and intrinsic values

When it comes to the question which values are most important, philosophers often make a distinction between instrumental values and intrinsic values.

Instrumental values are values that are important for the sake of something else. Money is, for example, often seen as instrumentally valuable, because it helps us to attain other important goals and values in life. Intrinsic values, on the other hand, are values that are important for their own sake, and thus are not used to attain something else. Typical intrinsic values are well-being, justice, beauty, honesty and truth.

Now we should ask how we can translate abstract moral values into tangible and effective design requirements. To answer this question, we will make use of a values hierarchy.

A values hierarchy consist of three layers: values, norms and design requirements. See the figure.


Figure 9.2 Values Hierarchy

Two Examples of a VSD
Below we will give 2 examples of a VSD-matrix: Cages for battery hens and biofuels.

Example 1: Animal welfare
The figure below is based on a European directive for the design and production of cages for battery hens. The directive was meant to guarantee the value of animal welfare in the design of battery cages. We can see how this value may be translated into several norms. For example, it is mandated that chickens should have enough living space. These norms are then translated into more specific design requirements, like that there should be at least 450 cm^2 of floor area per hen.
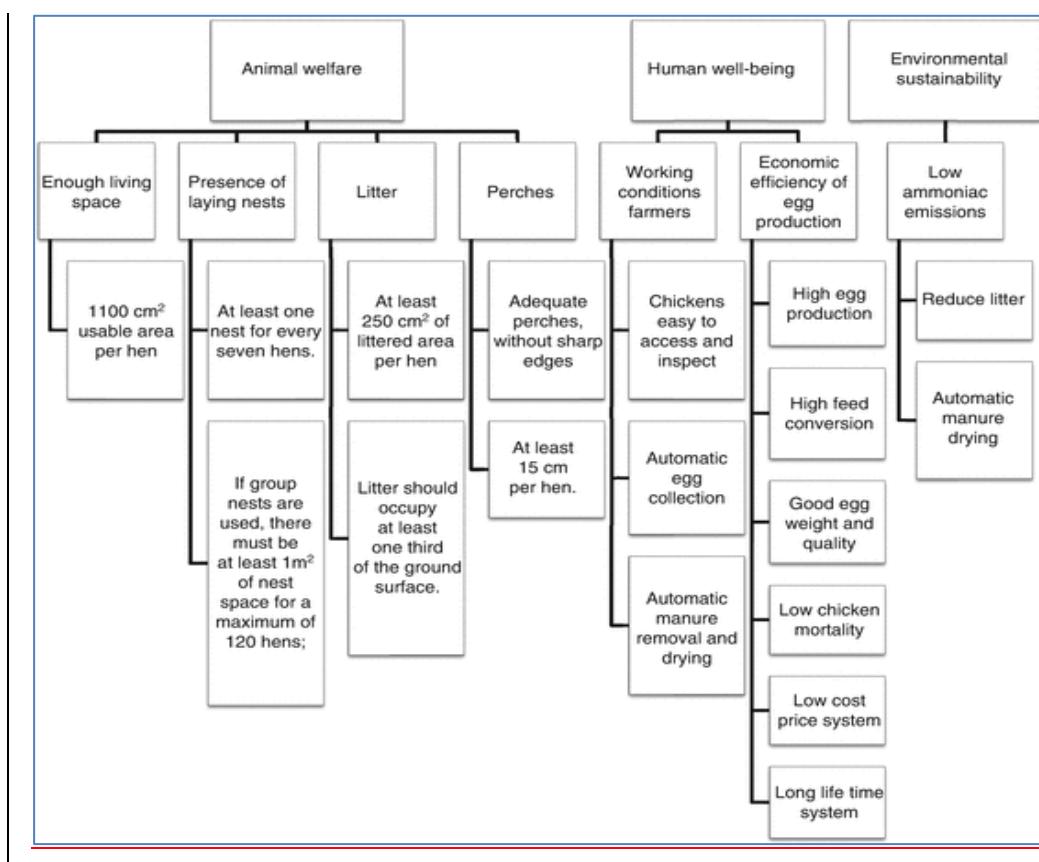

Figure 9.3 VSD-matrix for hens

Note: in this case, the values hierarchy has been reconstructed on basis of a European law, but we can also make a values hierarchy ourselves.

Example 2; Biofuels
Below is another example: an attempt to make a values hierarchy for biofuels.

Biofuels are based on relatively recent lifeless or living biological material. They have been introduced in order to deal with the expected shortage of fossil fuels, and to reduce emissions of greenhouse gases. They have, however, been met with fierce criticism for their environmental effects and for their effects on food production and food prices. Organizations like the **Nuffield**

Council on Bioethics have in response formulated ethical principles that biofuels should meet in order to be ethically acceptable. The figure below is an attempt to organize all such concerns into a values hierarchy.
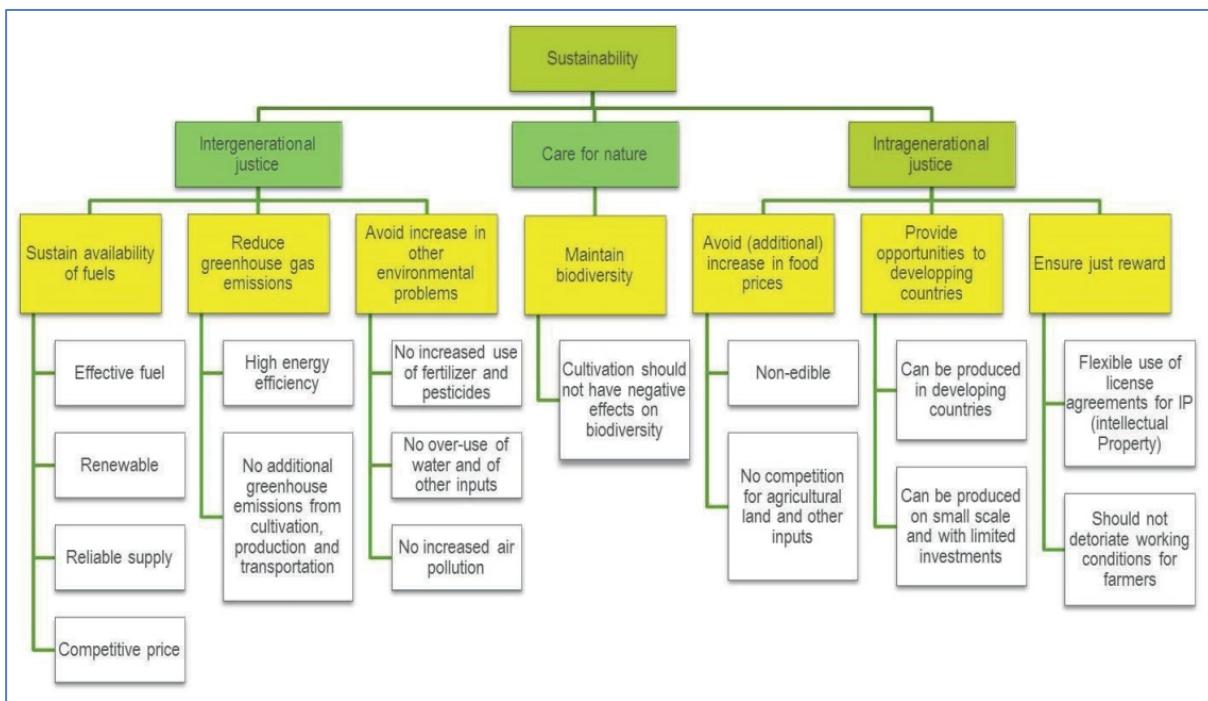

Figure 9.4: VSD-matrix for Biofuels

At the top is the value of sustainability, which is supposed to be the main value behind the development of biofuels. This value is broken down into three more specific values that are important in the light of sustainability: intergenerational justice, care for nature and Intragenerational justice.

A number of norms is associated with each of these values. Let us look at the example of intergenerational justice. Three norms are associated with this value, namely the need to sustain the availability of fuels, to reduce greenhouse gas emissions and to avoid an increase in other environmental problems.

Each norm is in turn translated into a number of more specific design requirements. For example, the norm that fuels should be available means that such fuels should be effective, renewable and reliable, and should have a competitive price. Another example is the norm that biofuels should avoid an increase in food prices, which means that they should be non-edible and not compete for agricultural land and other inputs.

There are currently no biofuels that meet all these requirements. Most current biofuels are first- or second-generation, which means that they are edible or compete with food-crops for land. However, third-generation biofuels are now being developed that allegedly solve these issues.

Returning to the values hierarchy, this can be constructed top-down, starting with a certain value like animal welfare or sustainability. We can then specify this value going down in the hierarchy. They can also be constructed bottom-up, starting with given design requirements. The key question to be asked in this case is: what ultimate goal do these requirements achieve?

An important question is whether the specification of a value in a values hierarchy is adequate.

We can question if meeting a specific lower-level design specification contributes towards meeting a higher-level norm or value (thus going bottom-up: from design requirement to values)? Let us look again at the example of animal welfare and battery cages. The question is whether meeting these design requirements is enough to attain the value of animal welfare. Many would doubt that this is the case. Indeed, the European Union has since changed its laws and formulated stricter design requirements that effectively forbid the battery cage altogether.

In this way, the values hierarchy helps us to structure and translate abstract moral values into tangible design requirements.

9.5 Complicated process

The example highlights a number of aspects that are more generally illustrative for the translation of values into design requirements.

- First, the translation of values into design requirements, especially of new values, may be a lengthy and cumbersome process.
- Second, translation may require specific expertise, sometimes from outside engineering. In the case discussed here, ethology provided such expertise. In cases of environmental values, environmental science or ecology may be relevant. For values such as privacy and trust, philosophical analysis may help to better understand these values and translate them into more concrete norms. Even values like safety and usability, which are more familiar to engineering, may require specialized expertise, as witnessed by the emergence of such disciplines as safety science, safety engineering and ergonomics.
- Third, translation will often partly take place outside specific design processes. The chicken husbandry example is extreme in this respect; often the final translation from more general norms into specific design requirements will take place within the design process. Nevertheless, in these cases as well engineers will often rely on specifications that are more generally available. Apart from legislation, a main source of such specifications are technical codes and standards, which are usually drawn up by engineers on standardization committees and which lay down requirements or guidelines for dealing with general values and considerations such as safety and compatibility.
- Fourth, the translation of values into design requirements is value laden. It can be done in different ways. Sometimes different (sub)disciplines offer different ways of specifying a value. Sometimes specification is made dependent on what is feasible with current technology or on trade-offs with other relevant values. The reason why Directive 88/116/EEC only addressed one of the four more general ethological norms was that it was deemed economically undesirable to formulate requirements that would de facto forbid the commonly used battery cage. From a philosophical point of view, a main question is when certain specifications are adequate or at least tenable.
- Fifth, the translation of values into design requirements is context-dependent. Although animal welfare is a general value, its specification is different in the context of the design of chicken husbandry systems than, for example, in the context of toxicity tests or medical experiments. EU Council Directive 1999/74/EC contained as many as three different specifications of requirements for chicken husbandry systems applying to three different types (layouts) for such systems.

- Sixth, the example illustrates that values and design requirements have a hierarchical structure. In this case, the general value of animal welfare was first translated by ethologists into a range of norms for holding chickens, and then governments translated these norms into very specific requirements.

Values are part of designing any technology. Choices about which values to include or exclude, and to what degree, are not always clear and obvious. Rather than ignore or overlook value conflicts, we need strategies for addressing them in a deliberate and thoughtful way. First, we can design innovations – both technical and institutional – for the purpose of solving or avoiding value conflicts.

Second, we can strike a balance between opposing values by establishing minimal thresholds that must be satisfied and then optimizing the design for the best balance. Here, we can follow the satisficing strategy – a combination of satisfy and sufficient. This means that both values need to first be at a sufficient level by meeting required thresholds. When designing a car, engineers must comply with a number of regulatory restrictions and standards about safety and sustainability. In addition to these thresholds, we still need to investigate the extent to which each value can be satisfied above the minimum level. In other words, satisficing the conflict between safety and sustainability means that we need to first establish the minimum acceptable (or, sufficient) level in each value, and then optimize the balance of each value above the threshold.

Case study #10: Autonomous weapons

What is an autonomous weapon? This is in fact quite a difficult question, partly because the concept of machine autonomy is very complicated, but also because it is contested. The easiest way to define an autonomous weapon is to say that an autonomous weapon can carry out certain tasks without a human operator. Once it has been pre-programmed, it does not need any further input or guidance from a human operator.

Figure9.5 Drones for military purposes

Now, this is obviously not very helpful, because we already have automated weapons. Automated weapons of course can also do the same things without a human operator. If we look at missile defence systems, for example, they have been automated because speed is a crucial issue in intercepting a hostile missile. Of course, they can intercept a missile once they have been programmed, without the operator having to do anything.

The really a big question in the debate and in the academic literature on new weapons systems, it seems, is whether autonomy should be seen as separate from automation. Some people agree with this, but there is no complete agreement.

One argument that we often hear in the debate is that an autonomous weapon can make a decision about targeting (by) itself. It can itself generate a targeting decision. Unfortunately, when people say that it makes a decision by itself, it is actually not clear what decision making means in this particular circumstance. This is something

we try to tackle through policy and research, just to shed light on this particular issue.

If people were really serious about machines making decisions, then we would be faced with machines that essentially would be able to apply intelligent criteria themselves. We could deploy them, we could put them into a particular situation, and then they could, by themselves, apply the criteria that regulate the use of armed forces. Realistically, this is still a long way off and, for now, it is more a science-fiction scenario.

We prefer to see autonomy as a more sophisticated form of automation. This means that the kind of machines we have in mind cannot make decisions in the sense that they apply intelligent criteria themselves. Still, they differ from automated systems. A cruise missile for example, is an automated system. We can program it with a GPS coordinates and it will find and hit the target by itself.

With an autonomous system, however, we would be looking at something more complex. For example, an autonomous system could be deployed in a very complex and challenging environment, in which it could navigate its own way to a target, without a human operator providing the GPS coordinates.

Autonomous weapons in a way exist on a continuum with automated weapons. They are the next step up, as it were, from common levels of automation that are already common in the military. As policy-makers, but of course also as citizens, we wonder about the risks and the advantages of these systems. On the one hand, we would have very narrow military advantages. With an autonomous system, we could, for example, fight at much greater speed over much greater distances.

Consider a stealth drone. Drones are currently remote-controlled by an operator. For that to be possible, there needs to be a link between the operator and the drone, so the drone can be controlled. The problem is that if the drone flies into enemy territory, the link between the operator and the drone could potentially be tracked by the enemy, and the enemy would know that something suspicious is happening.

With a very sophisticated autonomous drone, for example a highly automated stealth airplane, we could pre-program it. It could fly into enemy territory by itself. It could track certain targets and attack those targets without the operator having to do anything, apart from initially pre-programming it. Such stealth mission scenarios could be very useful to the military.

These are very narrowly (and) best described as military advantages. On the other hand, we could, argue that there could be ethical and legal advantages to using these types of systems. In particular, there are some roboticists who argue that by increasing machine autonomy, we can prevent war crimes and thus wrongdoing. If this is true, these systems certainly seem desirable. Who would not want fewer war crimes? Still, this is a very big if.

It therefore seems that the burden of proof falls upon those who think these kinds of

systems can really make such a difference on the battlefield. The risks are increasingly highlighted in the literature. There is a very large risk, for example, that these machines might not be able to adequately identify their assigned targets. We are talking about machines that could operate in very complex battlefields and under very complex circumstances. And there is a very genuine worry that it will be hard for them to find the kind of targets they've been programmed to look for and to attack those targets. This is obviously a very big risk. There are other technological risks as well, for example the possibility of our system being hacked and re-programmed by the enemy, and then re-deployed to commit war crimes or to attack our own troops.

We may conclude that there are some advantages, especially with regard to the military. There might be some ethical advantages, but there are also significant risks resulting from this type of weapon.

What is it that makes people most uncomfortable about the use of robotic weapons? It seems that a lot of the campaigning surrounding robotic weapons explicitly stresses the risks. There are two ways to answer this question.

The first answer is that people very often are scared of new technologies. In a way, robotic weapons or autonomous weapons are not entirely new. There are precedents which are very widely accepted, for example automation within missile defence. This is very widely accepted and people are not really worried about this at all.

The question therefore is: does this have something to do with the negative perception of particular technologies, which seem to raise these big issues? Or is it something entirely new, which is actually building on what's already there? Clearly, perception is a big issue.

Aside from the perception issue, we should also take the worries that people have about these weapons seriously. One worry could be reliability for example, and the way in which such weapons could be deployed in a safe manner. Could these weapons really find targets in a very complex battlefield? In a way, this is the crucial question in this whole debate. How reliable are these systems? And to what extent would their deployment be within an acceptable level of risk? Or will they impose excessive risks on others? Clearly, there is sufficient reason for worry.

There are obviously guidelines when it comes to the development of weapons and new weapon systems. These apply to autonomous weapons as well as any other type of weapon. There are also, of course, guidelines for their deployment, which need to be kept in mind. This means we are not operating in a vacuum - legally or morally - when we consider the development of these kinds of weapon systems.

The question always is, would people comply with these sorts of guidelines? There is a worry that these guidelines might be undercut by states that are very keen on developing these kinds of technologies. We're already seeing the use of drones during counterterrorism operations by the United States, which is legally and morally

ambiguous. It is not really clear how the law applies to this kind of operations and these specific situations. Compliance is clearly a further reason for worry.

We therefore need a strong response from civilian society. We also need a strong response from international institutions like the UN, the Red Cross and so on, in order to make absolutely clear to all parties that this is the process of the development and deployment of new weapon systems needs to take place in accordance with international law. We hope that individual militaries will hold the law in high regard. Many militaries do try, but it seems there is also a case to be made for very strict international supervision.

We may conclude that a key theme in this debate is transparency. Countries and armies should be transparent about what they are developing, within certain limits of course. They should be transparent about the use of such systems and proactive in ensuring that these systems are used in accordance with the law.
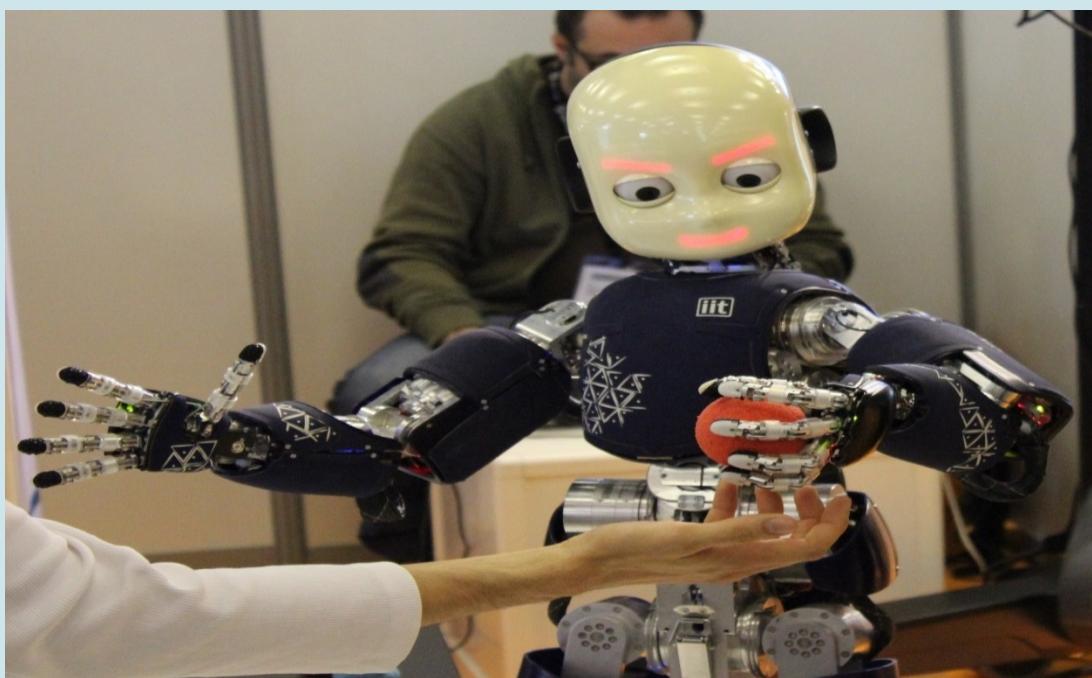
Case study #11: Care robots

For an extensive description of this case study see **this** *article by Filippo Santoni de Sio and Aimee van Wynsberghe.*

A revolution in healthcare is before us, one that involves the introduction of robots, in the surgical theatre and throughout the rest of the care sector. There is no universal agreement on the definition of a care robot. We propose that a care robot is a technological device integrated into care practices to assist healthcare personnel in their role as caregivers.

But when should we use care robots? After all, it is up to us to decide which roles we give them in our society!
In order to decide whether the introduction of robots in care activities is desirable or not, in addition to considerations such as cost-effectiveness and health benefits, we must understand what the values involved in those activities are. What about human dignity, privacy, attentiveness, responsibility, competence and reciprocity? One way to identify these values is to understand what the aim(s) of care activities is/are. With this information we will then be able to assess if these aims may be furthered or endangered by the introduction of care robots. We may run the risk that, given the high economic interests at stake, once assistive robots are introduced in larger numbers in care institutions and practices, they will be used no matter the preferences, values and needs of patients and users.


Figure 9.6: Robots are getting more and more important in healthcare

Let's have a look at two concrete examples.

Example 1: lifting of patients

If one were to consider the activity of lifting exclusively in terms of its immediate external goals ('goal-directed'), the activity could be described as moving a patient from a bed to a wheelchair in order to move him/her elsewhere, e.g. to the toilet, to an appointment etc. From this perspective, the activity of lifting simply consists of safely raising the patient out of bed, at a certain angle, with a certain speed and force, and safely placing them in their wheelchair.

However, seen through the practice-oriented lens, the same activity appears much more complex. During lifting the patient is vulnerable and responsive; he/she must learn to trust the caregiver and the caregiver must establish themselves as an agent who can be trusted (among other things). Lifting is an act through which the caregiver and care-receiver form a therapeutic relationship with each other. This relationship has a value in itself, but is also necessary for the future care of the patient, because it motivates him/her to be honest about their symptoms, to take their medication and to comply with their care plan. Lifting is also an act through which the caregiver is able to assess the neurological and physiological status of the patient, to make eye contact with the patient and to socially interact with the patient. Thus, under this practice-oriented description, the caregiver efficiently and safely moves the patient from one place to another, but also assesses the status of the patient and meets important social and medical needs of the patient.

So, should the operation of lifting be delegated to robots? Once we realize that it can be legitimately described appropriately in multiple ways, we can also understand the presence of a wider range of different, potentially contrasting values embedded in the activity:
- Seen simply as a process of transport, lifting is an activity that requires the safest and most efficient means to be fulfilled.
- Seen as a moment of socialization, trust-building and care-taking, lifting is a "practice" (in the sense of care ethics) that requires human responsiveness and human attentiveness to be fulfilled.

Example 2: Urine collection

As a second example of a healthcare activity, let us consider urine sample collection in pediatric oncology. Urine samples are routinely collected for testing for the presence of chemotherapy toxins in pediatric patients undergoing chemotherapy. Following the care ethics tradition, we may describe the activity of urine collection as practice-oriented in the sense, for instance, that the activity realizes the care skills of the nurse: the nurse remains attentive to, responsible for and competent as a care provider throughout the activity. Moreover, even urine sample collection can be seen as a moment in which patients have the chance to get in touch and briefly interact with a human caregiver. However, it is also true that the collection of the sample has a clear external goal, namely testing for chemotherapy toxins in patients. Sample collection may not only be embarrassing for the patient, but also dangerous for the nurse's health. Nurses often do not have time to put on protective clothing that shields them from chemo toxins which are able to cross the skin barrier. As a result, nurses put themselves at risk in order to satisfy the goal

of sample collection, while ensuring the wellbeing of the patient.

Now, in contrast to the activity of lifting, in this case the different aspects of the activity cannot only be conceptually distinguished, but also materially separated. We should not simply ask whether it would be permissible to remove the nurse from the complete activity of sample collection and to replace him/her with a robot. Rather, we should ask whether it would be permissible to remove the nurse from a portion of the activity of sample collection, namely the part which is goal-directed and harmful to the nurse, provided that a connection between the nurse and the activity of sample collection is maintained.

This Care Centered Value Sensitive Design approach provides a framework for designing future care robots in a way that systematically accounts for the recognition of care values throughout the design process of the robot. Understanding a care activity allows a robot designer to create a robot whose functioning is compatible with, or ideally can promote, the realization of care values. What's more, understanding the distribution of roles and responsibilities entailed in care activities allows robot designers to wisely choose the roles and responsibilities (or lack thereof) delegated to the robot.

To do this, the design suggestions are as follows: the robot is designed so that it travels autonomously to reach the patient's room. Once there, it requires information from the nurse to indicate whether he or she is present. This design consideration enforces that the nurse be present for urine testing. The robot then enters the patient's bathroom autonomously to collect the sample; it collects the sample from the patient's toilet or waste bin (as opposed to attaching itself to the patient's organs), and then exits the bathroom. It travels to the nurse waiting outside the bathroom or outside the patient's room and again confirms the presence of the nurse. The robot transmits the information that it has obtained the sample, and perhaps has already done the testing. With this information, the nurse can choose to send the robot, carrying the sample, to the oncology lab or to have the robot complete the analysis and send the results to the oncologist. Whatever the nurse decides, he/she is aware that the sample collection has taken place and that he/she is responsible for passing on the results of the analysis to the oncologist. This design suggestion is intended to ensure that a human agent is responsible for the successful completion of the sample collection.

The above means that it is possible to safely collect the urine sample, by removing the nurse from harm, while at the same time:
- Allowing the nurse to remain connected to both the patient and the patient's care by being present for the activity of sample collection;
- Preserving the nurse's accountability for the process.

Summary

We have come to the end of this book on responsible innovation. We hope you have enjoyed it, and have gained insights into the ethics behind the technologies we build and use on a daily basis. This chapter offers a short summary of the course material. See if you can refresh your memory as you read along.

In *Chapter 1*, we elaborated on the present context of complex socio-technical systems, and we introduced the notion of responsible innovation as an important and necessary aspect of developing new innovations and technologies.

In *Chapter 2*, we introduced various thought experiments, in order to explore how different dilemmas arise from the lack or confusion of values and responsibilities (the Trolley problem, "Many Hands", etc.). We saw that when there are multiple values to uphold, each of them important and desirable in its own way, moral overload can occur due to the inability to satisfy all these goals at the same time, given the constraints of time and resources.

Moreover, emotions may run high due to potential conflict of values, in which case, counter-intuitively, emotional responses could be seen as an opportunity to explore those values, rather than a liability preventing the emergence of a solution. Moreover, one can be optimistic about the use of innovation to satisfy multiple (conflicting or constrained) values; after all, isn't that what innovation is about?

In *Chapter 3*, we discussed the institutional context of modern innovation. We discussed how institutions - that is, embedded or explicit social conventions and rules that structure social interactions between individuals and groups - can profoundly influence favorable values and how they are manifested, as well as contribute to their preservation.

In *Chapter 4,* we focused on how companies think about innovation, in the context of competition and opportunities. We learnt how incremental and radical innovations come about, the factors that influence them, and how to manage these innovations in a conducive way.

In *Chapter 5*, we highlighted frugal innovations, a type of innovation that is specifically targeted at Bottom-of-Pyramid consumers. Frugal doesn't (just) mean the use of cheaper technology, but rather, these innovations are tailored for the lifestyle and living conditions of the communities they will be deployed in. That said, frugal innovations are not automatically "responsible", and the issue of social standards must be justified before this question may be answered.

In C*hapter 6* we described a new approach (roadmap) that companies can use to develop a strategy for RI and to include this in their Corporate Social Responsibilities policies. This starts with ethical leadership and is based on the pillars for RI: Anticipation, Reflection, Inclusiveness and Responsiveness.

In *Chapters 7 and 8*, we looked at one of the most important values for any technology, namely safety and security. To ensure the potential safety of a technology, we discussed how to assess a new technology for potential risks. One of the reasons for this is best illustrated by the Collingridge Dilemma: when a technology is new, it is easier to shape its development in a way that is desirable, but we may not always know all the risks. On the other hand, once the technology becomes embedded in society, the dangers might become apparent, but it becomes much harder to change it.

So, not all risks can be foreseen and there will always be the possibility of 'unknown unknowns'. In this case, we proposed the Precautionary Principle as a good maxim, since it allows us to develop new technologies with pre-emptive safeguards in order to mitigate the known risks as much as possible.

In addition to understanding and identifying risks, it is also possible to quantify them and engineer for safety. As such, risk analysis and safety engineering were introduced. First, we looked at one of the most commonly deployed methods for risk analysis: Cost-Benefit Analysis. Of course, there are some ethical concerns with this method, namely: how can we put a value on something which is essentially impossible to value?

We also introduced comprehensive risk analysis frameworks, with tools like the Fault Tree Analysis, Bow-Tie and Hazard-Barrier-Target model. These allow for both a quantitative and logical understanding of risks and their consequences.

Finally, in Chapter 9, we introduced Value Sensitive Design (VSD) as a framework for operationalizing the values we want to preserve in our technologies. VSD can be formally represented in a Values Hierarchy matrix, and can be approached both top-down and bottom-up.

The visual and explicit representation of values allows stakeholders to debate and negotiate these values in a constructive manner. Moreover, one can critically deconstruct and question the operational criteria: are the values that we hold dear incorporated in the design, or conversely, do the criteria achieve the desired values?

Appendices

Appendix 1: Overview of EU funded Projects in the field of RI

EU projects on RRI in Industry/SME's

- RESPONSIBLE-INDUSTRY
  The project explores how private corporations can conduct their research and innovation activities responsibly.
- COMPASS
  The project aims to support Small and Medium-sized Enterprises (SMEs) in three emerging technology industries to manage their research, development and innovation activities in a responsible and inclusive manner.
- KARIM (http://www.karimnetwork.com/)
  KARIM is the Knowledge Acceleration and Responsible Innovation Meta-network, a European project that aimed to develop transnational connections between universities, innovation-support agencies and SMEs, with a focus on responsible innovation

EU funded RRI process oriented projects: communication/research/governance/ awareness

- RRI-ICT Forum
  The project aims at analysing, supporting and promoting the contribution of Social Sciences and Humanities to, and the Responsible Research and Innovation (RRI) approach in ICT research and innovation under H2020.
- RESPONSIBILITY (http://responsibility-rri.eu/)
  The goal of the Responsibility project is to develop a virtual observatory for enhancing the interaction among research outcomes and policy making, incorporating the full potential of scientific achievements in the policy development and implementation.
- RRI-tools http://www.rri-tools.eu/
  This project develops a toolkit for RRI.
- ENRRICH (http://www.livingknowledge.org/projects/enrrich/)
  The Enhancing Responsible Research and Innovation through Curricula in Higher Education (EnRRICH) project will improve the capacity of students and staff in higher education to develop knowledge, skills and attitudes to support the embedding of Responsible Research and Innovation (RRI) in curricula by responding to the research needs of society as expressed by civil society organisations (CSOs).
- HEIRRI
  The aim of the HEIRRI project (Higher Education Institutions and Responsible Research and Innovation) is to start the integration of RRI within the formal and informal education of future scientists, engineers and other professionals involved in the R+D+i process.
- IRRESISTIBLE
  Design activities that foster the involvement of students and the public in the process of RRI. The consortium aims to raise awareness on RRI by increasing pupils' content knowledge about research.
- ENGAGE2020 (http://engage2020.eu/)
- MoRRI
  The project's objective is to provide scientific evidence, data, analysis and policy intelligence to support directly the Directorate General for Research and Innovation's (DG-RTD) research funding activities and policy-making activities in relation with RRI.

- **Ark of Inquiry**
  The project aims to foster RRI by teaching pupils core inquiry skills needed to evaluate the credibility and consequences of scientific research and by offering opportunities for pupils to engage with different societal actors involved in the research and innovation process.
- **FoTRRIS**
  The project aims to foster a transition of the existing Research & Innovation system to a Responsible Research and Innovation (RRI) system.
- **NUCLEUS**
  The project aims to develop a new understanding of communication, learning and engagement in universities and scientific institutions.
- **PROSO**
  The project has the objective to provide guidance on how to encourage engagement of citizens and third sector organizations, such as non-governmental organizations (NGOs) and civil society organizations (CSOs), in Europe's research and innovation processes.
- **TRUST**
  The project aims to foster adherence to high ethical standards in research globally and to counteract the practice of "Ethics dumping" or the application of double standards in research.

Other related domain-specific projects

- **https://canvas-project.eu/canvas/**
  The CANVAS Consortium – Constructing an Alliance for Value-driven Cybersecurity – aims to unify technology developers with legal and ethical scholar and social scientists to approach the challenge how cybersecurity can be aligned with European values and fundamental rights.
- **NanoDiode|**
  The project establishes a coordinated programme for outreach and dialogue throughout Europe with the aim to support the effective governance of
- **NERRI**
  Project which aims to contribute to the introduction of RRI in neuro-enhancement (NE) in the European Area. The project will involve different stakeholders and will promote a broad societal dialogue about neuro-enhancement.
- **RESAGORA(http://res-agora.eu/news/)**
  The ResAGorA project aims at doing extensive research about existing RRI governance across different scientific and technological areas, continuous monitoring of RRI trends and developments in selected countries, and constructive negotiations and deliberation between key stakeholders.
- **SMART-map**– The project will define and implement concrete roadmaps for the responsible development of technologies and services in three key time-changing fields: precision medicine, synthetic biology and 3D printing in biomedicine. SMART-map draws explicitly on the work of Res-AGorA.
- **SYNENERGENE**
  Mobilization and mutual learning action plan (MMLAP) for RRI in synthetic biology. The goal is to establish an open dialogue between stakeholders concerning synbio's potential benefits and risks, and to explore possibilities for its collaborative shaping on the basis of public participation.

- [PIER](#)
  Pier (Public involvement with an Exhibition on Responsible research and innovation) is a RRI project dedicated to European research on the Sea.
- [SPARKS](#)
  Te project has the objective to familiarize and engage European citizens with the concept and practice of Responsible Research and Innovation (RRI) through the topic of technology shifts in health and medicine.

Appendix 2: Questions for consideration

Below, you will find some question for you to consider. These can help you think about the concepts and theories presented in this book, and help you form your own opinion.

Questions for par. 2.2

- Can you think of some real-world problems or concerns that are typically presented as dilemmas? Would it be possible to introduce an innovation into the mix, in such a way that the dilemma effectively disappears?
- Do you think it's useful to try and resolve the various thought experiments mentioned here - e.g. the "Trolley Problem" or the "Fat Man" problem? Why?

Questions for par. 2.3

- Can you think of other kinds of co-operative schemes that could address other scenarios of the "tragedy of the commons"?
- Why might moral motivation work, and why could it fail?

Questions for par. 2.5

- Think of a controversial or risky technology. What are some of the values and emotions underlying the controversy?
- What do you think about the technology, regardless of the technical specifications?

Questions for par. 2.6

- Can you think of a moral dilemma raised by car crash testing?
- Can you think of other moral dilemmas in your area of expertise?

Questions for par. 4.3

Do you know any examples of innovations that never appeared in the market, or appeared but did not succeed?

Questions for par. 5.4

Do you know any examples of frugal innovations that also meet the condition for inclusiveness and social standards

Questions for par. 6.4

- Do you know of companies that apply RRI? What are the barriers? What are main KPIs for RRI for your profession?

Questions for par. 7.3

- What is your view on the intergenerational justice issue concerning nuclear waste?
- How does the Precautionary Principle influence your opinion on nuclear power production?

Questions for par. 8.1

- Despite the systematic effort undertaken during a CBA to capture every advantage and disadvantage of a given problem, there are still some ethical concerns about the practice. We might ask how you can put a price on what is essentially impossible to value. What do you think?
- What are other concerns? What are the underlying values behind a CBA methodology? Are these values the ones we want to emphasize?

Questions for par. 8.3

- As stated before, in complex socio-technical systems, we inevitably come across the problem of "many hands", not to mention uncertainties, innovations and interdependencies. On top of this, organizations need to adapt to competitors, changes in the market, changes in regulations and so on. Any one failure at some point may lead to a cascade of events, with high potential for negative impact. What does such complexity and constant change mean for risk assessment?

Questions for par. 9.3

- How would you formulate a VSD-hierarchy for autonomous weapons?
- How would you deal with conflicting vales (which there always will be)

Appendix 3: Teachers and link to weblectures

A large number of teachers provided input for the MOOC on RI. Below is an overview of the teachers (with a link to further information) including a link to their weblectures on YouTube and reference to the related paragraph in this book.

Teacher (click on name for more details)	Weblectures with link to YouTube	Paragraph in this book
Professor dr. Jeroen van den Hoven (also course director)	• **Introduction to the course** • **Introduction to RI** • **Trolley problem** • **Moral overload** • **Value Sensitive design, part 1** • **Value Sensitive Design, part 2**	• 0.1 • 0.1 • 2.1 • 2,2 • 7.1 • 7.1
Prof. dr. Ir. Ibo van de Poel	• **What is innovation?** • **Technology Assessment** • **Case study Coolants**	• 4.1 • 4.2 • Case study Coolants
Prof. Ir. Bert van Wee	• **Case study self-driving vehicles**	• Case study self-driving vehicles
Prof. dr. Neelke Doorn	• **Problem of the many hands**	• 2.4
Prof. dr. Cees van Beers	• **Economic determinants of innovation** • **Frugal innovation**	• 4.2 • 5.1
Prof.dr. Peter Knorringa	• **Frugal innovation**	• 5.2
Prof. dr. Sabine Roeser	• **Emotions and Values**	• 2.5
Prof. dr. Rolf Künneke	• **Four Layer model Williamson** • **Wind Energy**	• 3.2 • Case study wind Energy
Prof. Genserik Reniers	• **Cost Benefit Analyses**	• 7.1
Dr. Eefje Cuppen	• **Institutions and values**	• 3.1
Dr. Ir. Behnam Taebi	• **Case study nuclear energy and values, part 1** • **Part 2**	• Case study Nuclear Energy
Dr. Roland J. Ortt	• **Management of Innovation**	• 4.3
Dr. André Leliveld	• **Frugal innovation**	• 5.3
Dr. Philip Robichaud	• **Individual responsibility** • **Collective responsibility**	• 2.2 • 2.3
Dr. Rafaela Hillberbrand	• **Precautionary principle**	• 6.1
Dr. ir. Frank Guldenmund	• **Risk and Safety, part 1** • **Risk and Safety, part 2**	• 7.2 • 7.2
Prof Dr. ir Nick van de Giesen	• **Case study TAHMO**	• Case Study TAHMO
Dr Filippo Santoni de Sio	• **Care Robots and Values**	• Case study care Robots
Prof. Dirk Helbing	• **Digital Revolution**	• Case study big data meets big brother"

Appendix 4: Credit figures

Figure #	Link
0-1	https://en.wikipedia.org/wiki/Asbestos
0-2	https://www.interregeurope.eu/marie/
1-1	https://www.flickr.com/photos/stevensnodgrass/4385045639
1-2	https://commons.wikimedia.org/wiki/File:Goma,_Nord_Kivu,_RD_Congo_-
1-3	https://en.wikipedia.org/wiki/Campaign_to_Stop_Killer_Robots
1-4	https://en.wikipedia.org/wiki/Sustainable_Development_Goals
1-5	http://teachersinprogress.blogspot.com/2006/01/do-artifacts-have-politics.html
1-6	https://en.wikipedia.org/wiki/Closed-circuit_television
1-7	http://responsibility-rri.eu/
2-1	TU Delft
2-2	TU Delft
2-3	TU Delft
2-4	TU Delft
2-5	https://en.m.wikipedia.org/wiki/File:Trawlers_overfishing_cod.jpg
2-6	TU Delft
2-7	https://en.wikipedia.org/wiki/Deepwater_Horizon_explosion
2-8	https://en.wikipedia.org/wiki/Control_room
2-9	https://en.wikipedia.org/wiki/Anti-nuclear_protests
2-10	https://en.wikipedia.org/wiki/Ruth_Barcan_Marcus
2-11	TU Delft Slide
2-12	https://www.flickr.com/photos/us-mission/25005875442
3-1	TU Delft
3-2	http://www.sienna-project.eu/digitalAssets/721/c_721882-l_1-k_sienna-infographic-ai-robotics-pdf.pdf
3-3	TU Delft
3-4	https://www.researchgate.net/publication/322226825_Current_marine
3-5	https://www.reboostende.be/offshore-wind-farms
3-6	https://link.springer.com/chapter/10.1007/978-3-662-48847-8_3
3-7	https://en.wikipedia.org/wiki/Ligier
4-1	TU Delft
4-2	TU Delft slide
4-3	https://commons.wikimedia.org/wiki/File:Firmen_im_Silicon_Valley.jpg
4-4	TU Delft Slide
4-5	TU Delft Slide
4-6	http://conf.montreal-protocol.org/meeting/oewg/oewg-40/events-publications/Observer%20Publications/GIZ_Side%20Event_Greenfreeze_Brochure.pdf
5-1	http://www.globalhealthmgh.org/camtech/portfolio-item/2014-camtech-india-jugaad-a-thon/
5-2	TU Delft Slide
5-3	TU Delft
5-4	TU Delft
5-5	https://public.wmo.int/en
5-6	TU Delft
5-7	TU Delft

Figure #	Link
6-1	https://www.rri-prisma.eu/road-map-rri-for-companies/
7-1	http://www.indiaenvironmentportal.org.in/media/iep/infographics/Bhopal/Gas Disaster/index.htm
7-2	https://visual.ly/community/infographic/environment/climate-change-asia-and-pacific
7-3	https://www.nemokennislink.nl/publicaties/wetenschappers-pleiten-voor-co2-opslag-pilots/
7-4	http://meganck.com/mix/mx_04.html
7-5	https://link.springer.com/article/10.1007/s10551-015-2769-z
7-6	TU Delft
7-7	https://en.wikipedia.org/wiki/Yucca_Mountain_nuclear_waste_repository
7-8	https://en.wikipedia.org/wiki/Advanced_boiling_water_reactor
7-9	https://www.researchgate.net/publication/270842410
7-10	https://www.researchgate.net/publication/270842410_
7-11	https://en.wikipedia.org/wiki/Pebble-bed_reactor
7-12	https://en.wikipedia.org/wiki/Molten_salt_reactor
8-1	TU Delft
8-2	TU Delft
8-3	TU Delft
8-4	TU Delft
8-5	https://en.wikipedia.org/wiki/El_Al_Flight_1862
8-6	TU Delft Slide
8-7	TU Delft Slide
8-8	TU Delft Slide
8-9	TU Delft Slide
9-1	TU Delft
9-2	TU Delft
9-3	TU Delft
9-4	TU Delft
9-5	http://www.freestockphotos.biz/stockphoto/14725
9-6	https://en.wikipedia.org/wiki/Robot#/media/File:ICub_Innorobo_Lyon_2014.JPG

Responsible Innovation: Ethics, Safety and Technology
2nd Edition

Joost Groot Kormelink (Editor)

This textbook is based on the MOOC Responsible Innovation offered by the TU Delft. It provides a framework to reflect on the ethics and risks of new technologies. How can we make sure that innovations do justice to social and ethical values? How can we minimize (unknown)risks?

The book explains:

•The concept and importance of responsible innovation for society

•Key ethical concepts and considerations to analyse the risks of new technologies

•Different types of innovation (e.g. radical, niche, incremental, frugal)

•Roadmap for Responsible Innovation by Industry

•The concept of Value Sensitive Design (VSD)

It includes a link to all the web lectures as well as case studies ranging from care robots and nuclear energy to Artificial Intelligence and self-driving vehicles.

Joost Groot Kormelink
TU Delft | Technology, Policy and Management

Joost Groot Kormelink is manager Open and Online Education at the Faculty of Technology, Policy and Management and responsible for offering MOOCs in the field of Responsible Innovation for various target groups. Joost is also involved in international projects focusing on the risks and societal impact of new technologies and secretary of the Human Research Ethics Committee of the TU Delft.