Responsible innovation
From MOOC to book

TUDelft

March 2016

Table of Content

Foreword

From MOOC to book

This E-book is based on the Massive Open Online Course Responsible Innovation, which was offered by the TU Delft in November 2014 - January 2015 on the edX-platform.
This is the link to the re-run in 2015/16: https://www.edx.org/course/responsible-innovation-ethics-safety-delftx-ri101x.

This E-book contains all the content covered by the web lectures and will cover various aspects of RI, in an effort to provide an in-depth insight into applied ethics of technology and engineering, innovation management, stakeholder dialogues, risk perception, controversy and emotions, and science communication. The reader will also highlight interesting examples and case studies at relevant sections, casting light on the real-world applicability of these concepts across a host of different domains. The primary objective then is to demonstrate how to think about, and translate moral values as technical requirements for new technologies. Moreover, we will cover how to articulate value conflicts and think of creative (and moral) solutions for these quandaries. On the next page you will the course objectives.

About Responsible innovation and this book

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 have had serious negative consequences, or that just failed to address significant problems and make meaningful contributions to society; recall for example carcinogenic asbestos or the ecological devastation of DDT.

At the same time, we do need new technologies to find solutions for grand societal challenges such as energy scarcity, ageing demographics, water management and/or food security. So we are looking for responsible innovation in multiple upcoming fields that demand urgent attention in this regard: nanotechnology, biotechnology, artificial intelligence, policy-making based on big data analytics, and so on.
Our goal then is to give you an in-depth knowledge of what responsible innovation entails - an ethical perspective to help shape socio-technical solutions for global and regional problems.
This reader is intended to give a comprehensive but by no means exhaustive primer to responsible innovation. Responsible Innovation (abbreviated as RI henceforth) is a broad term that refers to the acts of analysis, reflection and public debate concerning the ethical principles and moral acceptability of new and emerging technologies.

To do this, we must be able to answer key questions such as these:
1. Do our efforts in applied science, technology and engineering contribute to the solution of the big problems of our age?
2. How do we find solutions for global problems in a responsible way?
3. Can technical solutions accommodate the plurality of moral values and the needs of all parties affected?

The term "Responsible Innovation" itself was first introduced in 2006 in the context of the Dutch Research Council Program entitled *Socially Responsible Innovations*.: it is now incorporated into the larger Research and Development agenda of the European Union (EU). As recently as November 2014, the policy was endorsed and extended in the Rome Declaration on Responsible Research and Innovation.

While thinking about RI has its roots in Europe, it is a concept with a true global reach. Consider that we live in a hyper-connected world: science provides knowledge of the fundamental building blocks and processes in Nature, our technologies scarcely leave any resource on the planet untouched. So, it is of the utmost importance, our duty even, to define an adequate and shared conception of responsibility for our innovations and technologies.

Can our innovations save lives? Will they produce more jobs? Can they save the planet, or only contribute more

waste and pollution? Are they safe for users and secure from abusers? Do they respect 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?

Learning Objectives

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

- Understand the need for responsible innovation
- Understand the complexity of responsible innovation and the operational challenges
- Be able to explain the difference between individual and collective responsibility
- Understand the role for communication and dialogue in responsible innovation
- Understand the relationship between values, institutions and responsible innovation
- Understand various types of innovation
- Understand the implications of radical innovation
- Understand the economic aspects of innovation
- Understand the concept of frugal or 'bottom of the pyramid' innovation
- Understand the concept of Value Sensitive Design and the VSD-framework
- Understand the concept of Constructive Technology Assessments
- Understand the relationship between various types of risk and responsible innovation
- Be able to critically discuss and assess various real-life cases

Structure of this book

The content is loosely divided into 4 parts. Part I will cover the concept of RI, delegation of responsibilities and the definition of values. Part II is concerned with the different forms of innovation (incremental, radical, frugal and so on), the modern innovation process, and socio-technical considerations of innovations. Part III covers the concept of risk, risk perception and safety engineering, and their implications for RI. And finally, Part IV will discuss Value Sensitive Design (VSD), a collaborative and visual framework to translate moral concerns into technical requirements in the design of technologies.

We hope you enjoy the course content. Good luck!

Part I

1. Introduction to responsible innovation

1.1 The real-world context of Responsible Innovation

Before getting into the definition of Responsible Innovation (abbreviated to RI going forward), it is good to put the discussion into the right context.
Try answering the following questions which have been designed to get you thinking about RI with real-world examples.

A. Dealing with hazards

New technologies can bring dangers and the possibility of not being able to control or contain those outcomes. We expect a certain level of risk that coincides 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 think about whether the harm is controllable. For instance, if we find out something is hazardous, would we be able to contain its effects by removing the specific technology from society/stopping its effects (or even reverse the effects)?

What are your opinions on the following statements?
To what extend 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 new innovation?

B. 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 of likelihood to knowing the probability of failure or having being certain knowledge of the dangers. If we are not certain of the outcomes, who is responsible for finding out, monitoring and taking precautions?
When assessing a new technology, how much knowledge about the hazards and risks is enough before deciding to introduce that technology in society? Should we assume that important risks and hazards will turn up every now and then, and it is not possible to anticipate and assess them beforehand? Or should we beforehand be certain of possible hazards and risks, and thus, should we have the capability to prevent or contain some extent of negative outcomes?

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

C. Distribution of risks and benefits

How should the risks and benefits of new technologies be distributed? What constitutes a 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.
How should risks and benefits be distributed? Should they be distributed equally across groups and generations? Or, as it is often the case in real life, , benefits and harms cannot always be equally distributed?

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 technology development? Do they have the power to block the development of potentially harmful technologies if need be?

How much influence should citizens have on the development of new technologies? Should citizens be able to block or discontinue the development of new technologies? Or do only producers and experts have enough knowledge and capability to make critical decisions?

As you may have observed, answering questions of this nature - which could potentially affect large populations - is not always easy. And yet, if we are to continue to invent and innovate new solutions to complex problems, we cannot avoid these questions. Instead we, have to confront them systematically in order to make the right decisions and implement appropriate measures.

1.2 Why discuss Responsible Innovation?

Innovation often brings wonderful and unimagined new functional abilities that are in demand and may lead to new business, new jobs and thus, economic prosperity; 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 in the course of history; much of it is clearly desirable.

But surely innovation is not a good in itself. If we agree that something is really innovative and brings interesting new functionality, 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 seen and sold as wonderful new technological inventions, but are now associated with a greatly increased risk of illness and even death.

The UN Millennium 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.

Innovation in our times is no longer about building bigger SUVs, but instead, it is 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. 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 those disabled,, cognitive neuro-enhancement may make some of us smarter someday. Whether these are acceptable innovations will depend on their precise features and on how we shape our technology.

So we have to take responsibility for our innovations and realize that technology is never neutral, but always value-laden; many 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 famously wrote his essay "Do artefacts have politics?", where he argued that the low-hanging overpasses in New York at the beginning of the 20th century had been designed intentionally low, so as to prevent busses to go from poor black neighbourhoods to the white middle class beaches.

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, others with perhaps less noble intentions may insert their views on sustainability, safety and security, health and well-being, privacy and accountability.

So, not only will our innovations have to be geared towards solving our grand challenges, they will themselves have to be expressions of our shared moral values. Technology is too central, and the science underlying it too fundamental; we should not first wait for outcomes and only reflect after the fact. This is why we need to think and act on responsible innovation, either by making embedded values 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 Responsible Innovation

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. 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 these in greater detail later in the course).

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 have been listed below.

- Fairphone is a start-up that makes smartphones from conflict-free metals, such that human rights, sustainability, fairness, security are accommodated in one design.
- Plans for the water-works in the Netherlands are storm surge barriers against flooding, but they are also ways to manage the ecosystems, 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 of fatal road accidents in Sweden has triggered a lot of innovation in the automotive industry. Volvo is now a leader in safety.

Innovation can thus be construed as a moral concept in the sense that it helps to change the world so that the set of obligations we can satisfy is amplified. There is no guarantee of course, that there always will be wonderful solutions to our pressing moral problems, and, in some cases, we may need to go for more drastic and fundamental approaches. However, we do have an obligation to see whether there are such possibilities. This, one could say, is the outcome or substantive aspect of RI.

There is also a process aspect to RI; in order to appreciate how responsibility is assigned in a complex system, we have to take a look at the criteria for being held responsible at all: 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 did not know it". "I did not mean it", "I was forced". "It wasn't me'. 'I didn't understand". In everything we do, we can at the same time go about it in such a way so as to enhance the conditions for responsibility. Or we may undercut or weaken them in order to make it more difficult for others to hold us responsible or accountable.

There are for example 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 here. Our company did not have the resources at that time to critically consider this".

This brings us to the defining clauses of Responsible Innovation in the EU report Research and Innovation for Strengthening Options. If some innovative organization or process would be praised by virtue of its being "responsible", this would imply among other things that those who initiated it and were involved in it must have been acknowledged as moral and responsible agents, i.e. they must have been enabled:

A. to obtain – as much was possible – the relevant knowledge on (i) the consequences of the outcomes of their actions and on (ii) the range of options open to them and
B. to evaluate both 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 and the possibility of resolving problems by design, another aspect of Responsible Innovation is the capability of relevant moral agents
C. to use these considerations (under A and B) as requirements for design and development of new technology, products and services leading to moral improvement.

On the basis of this characterization of innovation and the implications of (A), (B) and (C), we can define Responsible Innovation in summary as follows:

> **Responsible Innovation is an activity or process, which may give rise to previously unknown design and functionality either pertaining to the physical world (e.g. designs of buildings and infrastructure), the conceptual world (e.g. conceptual frameworks, mathematics, logic, theory, software), the institutional world (social and legal institutions, procedures and organization) or combinations of these, and which - when implemented - expand the set of relevant feasible options regarding solving a set of moral problems.**

We thus suggest a core conception of responsible innovation which refers to, among other things, a transition to a new situation, and which has as its defining characteristic the amplification of possibilities to meet more obligations and honour more duties to fellow human beings, the environment, the planet and future generations than before.

2. Applied ethics for Responsible Innovation

2.1 Applied ethics with 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. Let's look at it now.

The Trolley Case

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. Further ahead on the track, there are five people 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, you notice that there is one person similarly tied up on the side-track.

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?

1. Flip the switch to maximise the number of lives saved (1 person dies so 5 can live).
2. Flip the switch because you are a compassionate person and it is the right thing to do.
3. Do not flip the switch as it would be a form of killing, and killing is inherently wrong.
4. Do not flip the switch because you feel aiding in a person's death is culturally inappropriate, not to mention illegal.
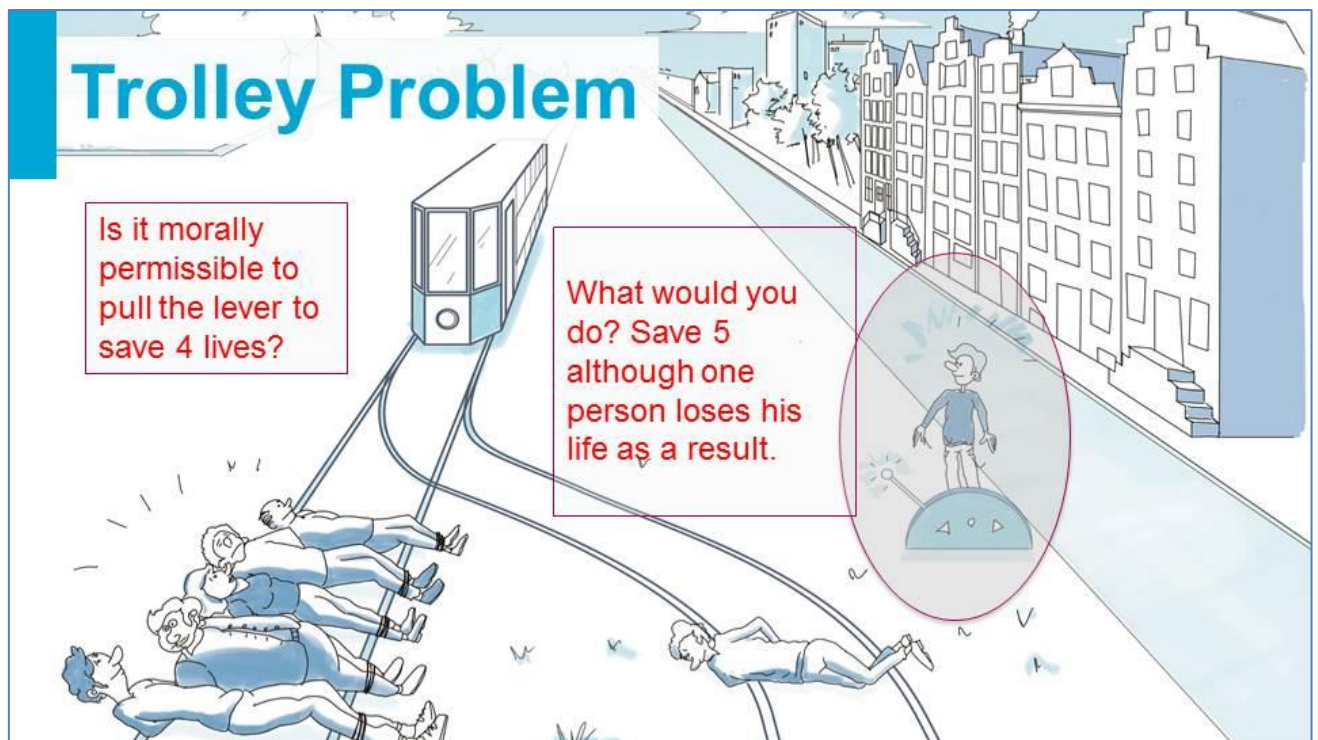
Figure 1: the trolley problem

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 3 decades have been raised on such so-called "trolley cases". If you would like to do a PhD in trolley analysis, that would actually 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 pay attention to this artificial thought experiment is not to introduce you to the extensive body of literature – that could be the subject of a separate MOOC - 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. Instead, there are still five people tied up on the track and the trolley is barrelling towards them, but there is also a fat man standing on a bridge over the track. By pushing the fat man onto the track one can stop the trolley before it hits the five tied up 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.

Empirical research shows however – some argue even that brain imaging studies also point in that direction- that we react in a different way to both cases, although the numbers and the calculations are the same. In the lever case, we primarily relied 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 such a preposterous idea to use someone as an obstacle and by doing so, to end up killing him.

How engineers answer the "Trolley Problem"

So we've generally seen the philosophical questions that such dilemmas expertly bring to the surface. What you will not find in the trolley literature however, is the following: engineers and designers of technical artefacts often reply to trolley cases by saying that it is a stupid piece of railway infrastructure and a very bad design - and they would be right! Engineers especially would immediately start to think of better system designs and innovations to 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. Again, 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 most interesting move in another context, namely the one that pertains to preventing deaths in rail transport, and safety of rail infrastructure.

What this response clearly brings to the surface, is that in 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 design and so, their conception of responsibility therefore extends to even the design histories which lead to 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 that 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 so, it is as important to learn to prevent dilemmas from coming into existence 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, designer foods, untested 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.

Does it mean that the study of trolley cases is useless? No, not at all! The types of moral considerations, concerns, lines of reasoning, moral categories that are invoked in such discussions can be fruitfully used in the design of high-tech innovations, systems or infrastructures. Moreover, these discussions also reveal different sets of values that people have. So, it is important to involve the stakeholders and address their values as well.

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 have 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 or any other innovation deployed in society in some way or form.

2.2 Individual Moral Responsibility

Now that we have been introduced to a general 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 different 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 technologies and to innovate in a way that promotes responsibility. In order to know how to promote responsibility, however, one needs to have a clear understanding about what responsibility is. For now, we will focus on backward-looking responsibility - i.e. judgement of past actions - rather than considering their future obligations.

The "Wrong Switch" Case

Let's consider a chemical accident case. We'll call this case: "Wrong Switch". Imagine that an operator of a chemical plant notes that there is leakage coming from a tank, and in an attempt to contain that spill, the operator accidentally turns the wrong switch...

Imagine further 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 some outcome is to ask whether the same outcome would have been obtained if the person did not act as she did.

This way of understanding causal responsibility seems uncontroversial and seems to apply in the "Wrong Switch" case. But notice here that it's quite another 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. To see this, look again at the "Wrong Switch" case. The operator's moral responsibility seems to depend on the explanation for why she 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 slightly explains 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 he or she believes to be the right switch, but instead of stopping the leak, there is an explosion killing a poor worker again. 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 he or she 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" morally responsible for the worker's death?

One way of coming to an answer on this question 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 he or she couldn't prevent the explosion from occurring despite trying, it seems to be a mistake to think that he or she is blameworthy for the worker's death. 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 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. In this version, the operator intentionally and knowingly turns the wrong switch in order to kill the worker!

This case differs from "Mixed Wires" in two important respects. Firstly, the operator has ill will towards the

worker who dies in the explosion, whereas in "Mixed Wires" the operator had no such ill will, and was actually motivated to try to stop the explosion. The second difference is that in "Hateful Operator", the deadly explosion is avoidable. The operator knew that she 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 blameworthy 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 had ill will toward the worker entail that the operator is morally at fault. Importantly, this kind of moral blameworthiness 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 blameworthy for it. Let's call this case "Extra Effort". This case is similar to "Mixed Wires" in that, the operator doesn't know and couldn't have known he or she was turning the wrong switch. Imagine however, that when the operator realizes that he or she 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, and you might find yourself with a strong intuition that the operator is morally praiseworthy for doing so. She had to think very quickly and had to carry out a very difficult action in order to save the worker's life. She was motivated to go the extra mile in order to save the victim, and that seems to be good grounds for thinking that the operator is morally praiseworthy. In this case as well, it is important to notice 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 "Everyday Thing" Case

The final version of the chemical spill scenario highlights yet another important aspect of moral responsibility. Let's call this case "Everyday Thing". 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 her actions were perfectly ordinary and didn't require a tremendous amount of effort or achievement.

It also seems obvious the operator is not blameworthy given that she does nothing wrong. Still, we may want to reserve judgment whether the operator is morally responsible for avoiding the death of the worker. The operator did the right thing freely and intentionally, and she knew what she 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. First and most minimally, we have the notion 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 "Everyday Thing" case, it seemed like the operator was morally responsible for preventing the worker's death, even though she doesn't seem to merit either praise or blame. The third notion of responsibility involve cases where the agent's actions seem to merit praise or blame. We saw that the agent was praiseworthy when he or she went above and beyond the call of duty to do the right thing, and we saw the operator was blameworthy when he or she knowingly and intentionally killed the worker. These agents are indeed morally responsible, but we are inclined to add that they are also praiseworthy or blameworthy.

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, if she is to be morally responsible for it. Without causal responsibility, we cannot have moral responsibility. The second connection is that both praiseworthiness and blameworthiness presuppose moral

responsibility. For example, if the operator could not have avoided causing the worker's death, then he or she is not morally responsible, and therefore not blameworthy either. Thus, judgments of praiseworthiness and blameworthiness 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 about the various factors that affect the attribution of responsibility should help to design processes 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 their processes, and the system itself should be designed with multiple fail-safes that are easy to access.

2.3 Collective moral responsibility

In some cases, individual moral responsibility alone is not enough to address some 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". They can arise in the context of shared resources such as rivers, the atmosphere, and national parks. We shall focus with a typical example of a tragedy of the commons scenario, as is found in overfishing.

Tragedy of the commons

Imagine that some seaside villagers rely on fishing for their economic livelihood. As it happens, each fishing boat in the village must compete with the other fishing boats to bring in a catch. Because of this competition and the constant demand for fish, there is overfishing. 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? Well, the tragedy has to do with the way that overfishing seems to be inevitable, namely due to the individual fishermen acting in their own rational self-interest.

It is important to notice that it is in each individual fisherman's rational self-interest - let's call him fisherman A to catch as many fish as he can. This is because if fisherman A catches less fish than his maximum capacity, he will make less money, and meanwhile his competitors, namely fishermen B, C, or 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 well. As each fisherman only acts in his individual rational self-interest, the common stock of fish is soon depleted!

Although the individual fishermen took apparently rational actions, this behaviour does not add to the best interests of everyone collectively in the long-term. The community's interests are set back because they risk losing an important source of nutrition, 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 could practice sustainable fishing. Sustainable fishing means taking only that 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. Such a scheme would limit the size of the catch for each boat but 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 free-riding

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 for 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 yet again 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? Well, 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 that 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 end 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 can't just 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 can rely on enforcement. For the quota system to work, some significant degree of compliance must be achieved, and there are ways of 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. This enforcement shifts the individual rational self-interest through fines or penalties such as revoking the license to fish, thus aligning individual interests with collective rationality for the most part.

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 themselves, if they are acting in their 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. So some kind of enforcement becomes necessary, although it is not sufficient either. What if both the fishermen and the maritime police were both 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.

2.4 Responsibility in complex systems

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 then to pinpoint who is causally or morally responsible, or who is to blame.

What we see is that the actions of the four people together lead to some dramatic outcome but none of the

individual persons can be held responsible. This phenomenon is called the problem of "many hands". Because there are different people involved, it is impossible to identify one single person that is responsible. This problem is very urgent in engineering because there are often many people involved in the development of technology, even in the development of risky technologies – 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 those 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:

- The first condition is the freedom condition: the person should be free to act and not be under external pressure. So if I put a gun to a person's head and ask this person to do something illegal or immoral, this person cannot be held responsible. He was not free to do otherwise.

- The second condition is the knowledge condition: a person should have the knowledge that his action would lead to some negative outcome. If the person does not know this, he will generally not be held responsible. If for example, someone painted the door of his house without putting a notification that the door was wet. If 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 or could not know that the door had just been painted and that therefore, you should not have touched it.

- The third condition is 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. So if someone else also touches the previously mentioned painted door, it is not fair to hold you and you alone responsible. Note that sometimes, doing nothing is the wrong act. So if one has the possibility to save another from harm, not doing anything is the wrong act.

- The last condition is about the transgression of a norm: if what you did was somehow faulty, then we can say you transgressed a norm. This norm can be a legal norm, but also an ethical or social norm.
  This is a difficult condition but especially when we talk about institutions later in the course, this is an important condition. Institutions should be designed in such a way that they represent the right norms.

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 with 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 he is responsible for cleaning the firemen's outfit.

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 none of the individuals fulfilled all conditions. The person working in the laboratory may have known that this material could have some chemical reactions 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.

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

Think of the oil spill of the BP platform in the Mexican gulf. The impact of this disaster was huge and it immediately prompted the question: who is responsible for this disaster?

The problem of many hands is often discussed in a backward-looking sense, that is, after some 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 the person has the ability to fulfil his responsibility: does this person have the freedom to act? Does he have the necessary information? Are the right norms in place? And so on.

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 a technology taking over responsibility from a person. But equally, can technologies be developed such that they enable people to assume their responsibility? We think that technology can indeed have this role, but we should pay attention to specific aspects of responsibility when technology is being developed.

The first example we consider is the V-chip. The V-chip is a technological device designed to prevent children from watching violent television content. TV stations broadcast a rating as part of a program. Parents just program the V-chip by setting a threshold level rating and all programs above that rating are automatically blocked by the V-chip when it is turned on, so children who are watching TV cannot view the blocked programs.

Some people argue that by using the Vchip, parents transfer responsibility to the TV stations because they are letting the TV stations decide on the exact rating of each program and thus 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 violent content and as such, gives them more freedom to check what their children are actually 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.

Another example would be a control-room. A control-room is a central space where a large facility or service can be monitored and controlled. These rooms are often equipped with multiple monitors and screens. The people working in a 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 indeed able to make the correct decisions. We could argue that a badly designed control room may hinder people from assuming their responsibility. And vice versa, a well-designed control room may enhance a person's ability to carry out his responsibility.

So, technology can enable, but also hinder people when carrying out their responsibilities. One aspect of responsible innovation is then to develop technology in such a way that it may indeed facilitate or strengthen people in their ability to carry out their responsibility.

2.5 Emotions and values

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

As such, 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 grasp morally salient features of risky technologies, such as fairness, justice, equity and/or 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, emotional public debates. While experts emphasize scientific studies that point out supposedly low risks, the public is often concerned about the impact of such technologies on society. The experts emphasize that the worries of the public are due to a lack of understanding.

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 me call these two extremes the technocratic pitfall and the populist pitfall respectively. In both pitfalls, there is no genuine debate about emotions, public concerns and moral values; 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 with cost-benefit analysis, as conventional, technocratic approaches take it to be. In other words, the experience of risk is something quite different from the equation 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 a risk assessment. Interestingly, as we know from the influential work of the psychologist Paul Slovic, these considerations do come up in the risk perceptions of laypeople. Apparently, the pre-theoretical connotations that people have with risk include ethical considerations that normally get 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 than experts.

It has become more and more clear that laypeople's risk perceptions are largely influenced by their emotions. Social scientists struggle how 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 philosophy and psychology these days is a socalled cognitive theory of emotions, according to which, emotions are a form or source of cognition and knowledge. These ideas can shed 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 then 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 worried about the technologies they develop. Fear can point to concerns about

unforeseen negative consequences of a technology. Fear and anxiety can indicate that a technology is a threat to our wellbeing. Disgust for example, can point to the ambiguous moral status of clones and human-animal hybrids. Meanwhile, indignation may be an indication of violations of autonomy, in cases of risks we are exposed to against our will.

It is often thought that emotions are by definition against technology and therefore one-sided, but this is not necessarily the case. Enthusiasm for a technology, for example, may suggest 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 risk emotions in public debates, by voiding both the technocratic pitfall and the populist pitfall

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 for important insights into moral constraints and desirable parameters of responsible innovation. For example, in debates, experts should not only focus on 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 needed 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, since emotions are a predominant source of motivation.

When developing risky technologies, we argue that emotions and moral concerns have to be seriously taken into account 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 again in debates about risky technologies; this way the public will feel that their concerns are taken seriously, contributing to participative and responsible innovation.

2.6 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, more sustainable, to create new jobs while simultaneously protecting privacy and fighting terrorism, to give autonomy and freedom to future users, to improve 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 and sustainability. We value our security but also our privacy. We demand safety but are not willing to sacrifice some freedoms. We require accountability but insist on the right to confidentiality as well.

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.

> 1. The agent ought to do A.
> 2. The agent ought to do B.
> 3. The agent cannot do both A and B.

We think it is important for a better understanding of Responsible Innovation that you 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.

Two views on 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 security?
- Nuclear power plants: do we want energy security or less exposure to risk?
- Drones: do we want our soldiers to be safe or accountable?

Dilemmas such as these are typical of a moral dilemma. Anyone who confronts a situation with a dilemma typically has two obligations: however, he can only fulfil one of them, but not both of them at the same time. There are two views on such dilemmas:

The first view is that, based on a balance of good reasons, the alternative that you end up choosing is your true moral obligation. The other obligation, might have seemed like a serious moral contender, but it turns out after some deliberation not to be among the things you are obliged to do. And so the dilemma was resolved. One could say that the dilemma does not really exist, it only appears to exist. What seems to be a situation of conflicting obligations is just an apparent dilemma.

The second view of moral dilemmas holds that both options are genuine obligations and exert a moral pull on us - and they keep exerting this moral pull - regardless of what our final choice is. According to this second view, both options are genuine obligations, but we must unfortunately choose only one of the two. The problems with this second view are twofold.

Firstly, it seems to be inconsistent with a general moral principle, which is 'ought to implies can'. This principle tells us that it only makes sense to speak of a moral obligation for someone to do A if that person can actually do A. So, if we have an obligation to do A and an obligation to do B – assuming we can conclude from this that we have an obligation to do both A and B at the same time – then we must conclude, applying the principle 'ought to implies can', that this only makes sense if we can do both A and B. The nature of the dilemma however is precisely that we cannot do both A and B at the same time; if we could, there would be no dilemma in the first place. There is a contradiction, in other words.

The other problem also applies to the first view on moral dilemmas. It concerns the question why we often experience remorse or regret for not being able to do the other thing that we were obliged to do, or appeared to have an obligation to do.

For the first view this poses a problem . We should expect that there is no regret or remorse over the road not travelled because that option is not our obligation. The psychological facts are usually different however, and we often do feel bad about the lives we were unable to save, the animals we could not rescue, the CO_2 emission targets we could not achieve. Even if we know we did the right thing, we may still experience these feelings.

For the second view on moral dilemmas, the fact that we cannot do both A and B at the same time should help us to put our feelings of having failed and thus being blameworthy in a sense in the right perspective. But this also does not work; we still feel bad!

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. Barcan Marcus suggests that if we have an obligation to do both A and B, we have a second order obligation to see to it that we can indeed do both A and B. We have termed this second order obligation a "meta task responsibility".

According to philosopher *Jeroen van den Hoven,* a "meta task responsibility", is an obligation to assess prior to the task performance whether the environment in which they will have to work, is likely to allow them to do what they ought to do in situ and to ascertain that it at least does not prevent them from doing what they ought to do.

If we accept this idea of a higher order obligation, we can see how the fact that we have not succeeded - or not even tried or even thought about trying - is what gives rise to the feelings of regret and guilt. It leaves us with a moral residue.

Do note that this second order obligation is not subject to the principle of 'ought implies can'. It may turn out not to be possible that both A and B can be done. The meta task responsibility to see that one can do what one will be morally required to do at a later stage, is an obligation that we implicitly endorse and mentally focus upon, if and when we feel that we have failed from a moral point of view.

The moral force - both ex ante[1] and ex post[2] - of the sheer possibility and the opportunity to change the world in such a way that our obligations become co-obeyable, and our responsibilities co-satisfiable, can in itself act as a driver of innovation, invention, and creativity in engineering and design. So, instead of changing our values, ranking our obligations or reducing them to one metric or KPI (key performance indicator) , we are encouraged to bring about a new state of the world which allows us to have our cake and eat it too. Essentially, this moral force prompts us to solve the problem of moral overload by innovation!

Moral obligations 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 also been raising concerns about their impact on privacy. We have to design meters that satisfy all the functional requirements - thus serving the goals of sustainability - and 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:
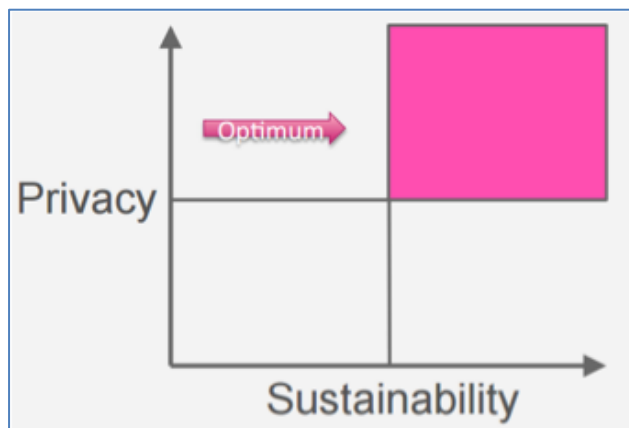

Figure 2: meeting different requirements

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 v2.0 may give us one, but not both. The smart meter 3.0, which is what we are ideally looking for of course, is designed to accommodate both of the functional requirements in order to make management of the grid more efficient, while 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 as we frame them - in one smart design, some of which were actually in conflict before.

Similarly, if we would like to benefit from RFID technology in retail, but fear situations in which we can be tracked and traced 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 be active and transmit data.

There are more examples like these which illuminate how we can take moral obligations seriously - towards customers, future users, future generations, 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. Dealing with moral dilemmas can help stimulate creativity and innovation, and innovative design may help us to overcome problems of moral overload.

Part II

3. Institutional context of Innovations

3.1 Introduction to institutions

Let us comment briefly on the role of the institutional context in which a technology is being developed and implemented. This is 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. Think of infrastructural projects such as the construction of roads and dikes, or of energy projects such as wind farms or even natural gas production. Many of 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. This alludes to values such as transparency, fairness, and procedural justice. 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 enable 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 ensure and safeguard the value of environmental health and safety. It is interesting to note that in controversies, often institutional rules such as environmental impact assessment 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 practically prioritized in the decision-making process at

the expense of others. This shows how values embedded in institutions are intertwined with the acceptability of technologies.

The 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 in this case. 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. Of course this is something that may hamper the involvement of public in decision-making. The 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 the 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, it changes over time and place. This means that the acceptability - 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 being reformulated as a reaction to both changes in the perceived threat and in the degree of acceptance of high dikes as the primary means of protection. 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 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 literally 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.

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 value sensitive design comes about, it is also important to consider the values and beliefs of technology developers and/or relevant policymakers as well. This is because we know that the way the public respond 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 that of NIMBY, which is 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 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. So, if a project developer thinks that someone

is ill-informed, he will probably focus in his communication on providing technical facts and explain 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, leading to the paradox that efforts to prevent opposition by providing "the hard facts" may actually provoke even more public opposition.

Designing for values thus means that we 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 diversity of values and problem definitions at stake in case of technology.

### Accounting for institutional values in innovation

So, 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/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. It should be specified to specific contexts, not based on a cookie-cutter blueprint
4. It should include a humanistic reflexivity of the designer/technology developer/policymaker

3.2 The Four Layer model of institutions

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

Figure 3: Dutch North Sea and its various uses
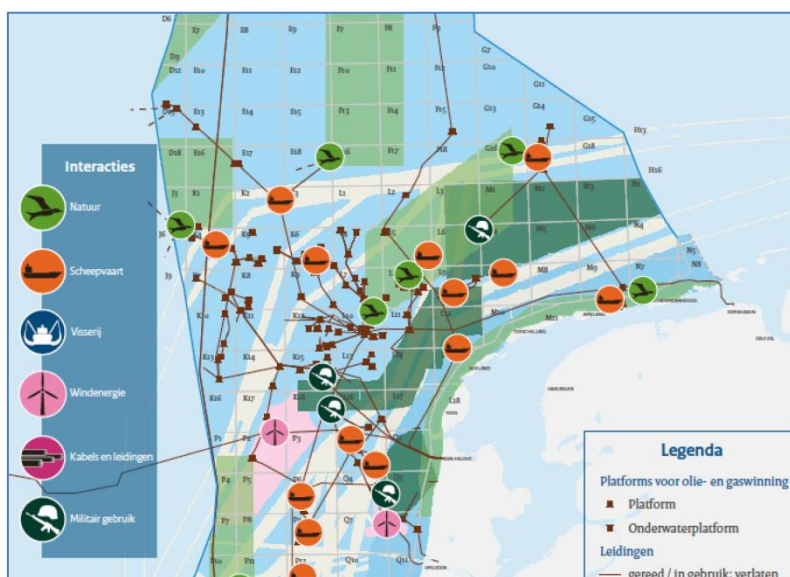
As illustrated by the different lines and dots, this North Sea is not just an empty abandoned space, but is used for very different purposes like natural recreation, naval transport, fishery, cables or even military uses; now, wind energy production is also being introduced in that area. So the area has many different users that make use of the North Sea and naturally, these users need to adhere to certain social rules. This is simply a matter of practicality in most cases: it would not be useful to have 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 in an unspoken way.

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 certain areas precisely where the military may practice their exercises. So we need formal rules to dedicate these areas for specific purposes.

But there might also be embedded 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, instead they are embedded - that is to say based on certain traditions that people did for long periods of time.

Building on this understanding of social rules, let's go back to the possibility of offshore wind turbines in the North Sea: we might have to alter some 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 them if possible. For example, if we dedicate certain areas for wind power, fishermen might not get good catches in these areas, naval transports need to find other routes and military exercises of course would be prohibited close to this area. So we are not talking about single rules, but we are looking at a system of different social rules, which in turn both prescribe and describe social behaviour.

Types of institutions

Going a step further, there are different categories of institutions, identified by Oliver Williamson to be four different layers which are illustrated here in the figure below.
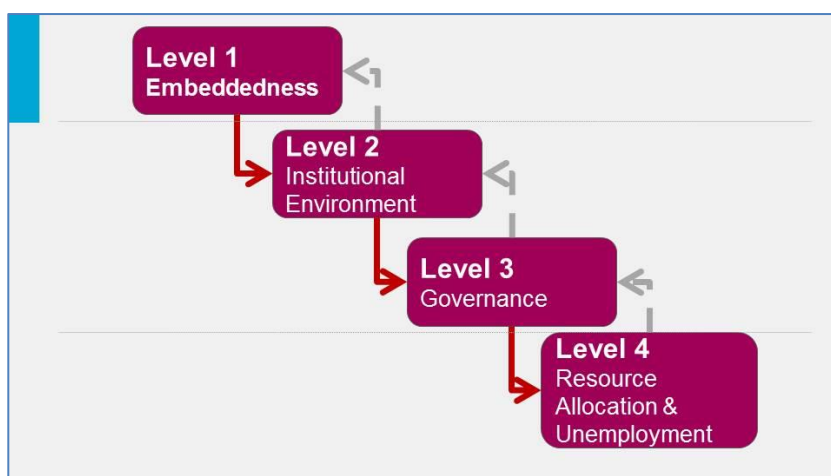

Figure 4: Four Layer Model by Oliver Williamson

- *Embeddedness*

Let us start with the upper layer L1, which is called Embeddedness, which refers typically to informal institutions, customs, traditions, norms and religions. As mentioned earlier, fishermen who traditionally fish in certain areas for generations, or citizens who are used to camping in certain areas over the years, have done 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. So Williamson uses an indication of once in hundred or thousand years; this is really deeply embedded in the behaviour of people, or actors. The purpose, which is a third category that Williamson identifies in this scheme, is often non-calculative. These informal rules just evolve very spontaneously, and they are very difficult to plan. Equally, trying to influence these informal rules, these embedded rules, can become quite a difficult task. But this is a layer that is very important if we are looking to pursue responsible innovation.

- *Institutional environment*

Layer 2 is the Institutional Environment. These are the formal "rules of the game". Examples include the constitution 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 is about once in ten or 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 getting the institutional environment right, and Williamson refers to this as first-order economizing. Answering which institutions would serve best the purpose of stimulating the development of offshore wind energy, might be one of the conditions under which we design specific formal rules for the establishment of the same.

- *Governance*

Layer 3 is Governance, or the play of the game. Given the formal rules and the embeddedness, 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, which organizations serve best certain individual objectives, firms or 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.

- *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, and these constant interactions individually and collectively shape how the institution works in practice. This is referred to as third-order economizing.

Applying the Four Layer model of institutions

So 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 above scheme by the arrows, top-down and bottom-up arrows. The arrows suggest that the different layers are not to be analysed in isolation.

For example, if there are certain norms or customs, or certain informal institutions in a country or in a region, it is important that these rules and informal institutions are taken care of by the institutional environment. So, the formal rules of the game are to a certain degree based on the informal institutions. And if this were not the case, we would have a serious problem because these 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 in this layered scheme, we can also argue that the governance and the resource allocation also 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 by each other; 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. It might be that that when resource allocation or governance changes, this would have an influence on the formal rules and the informal institutions.

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

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

So 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, towards change of the institutional environment and embedded institutions, values and customs. In this respect, the institutional context is crucial to understanding and shaping the success or failure of new innovations.
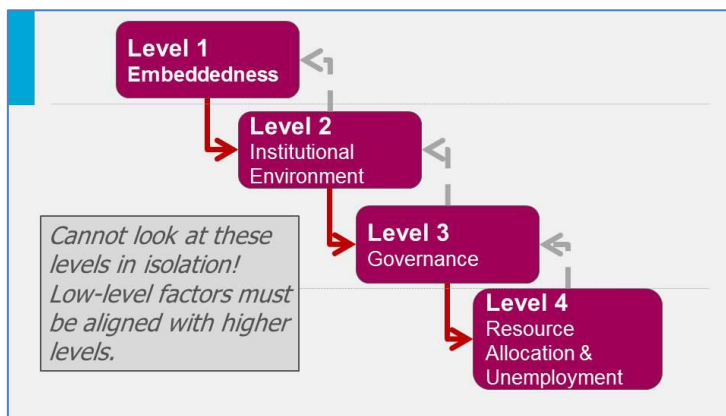

Figure 5: Four Layer Model by Oliver Williamson: not be seen in isolation!

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 theproduction process itself. So, 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; offer new functionalities; reach out to new user groups; or they may serve new types of values.

Here, we will rely on taxonomy for innovation that was 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 lead to the quadrants shown below.
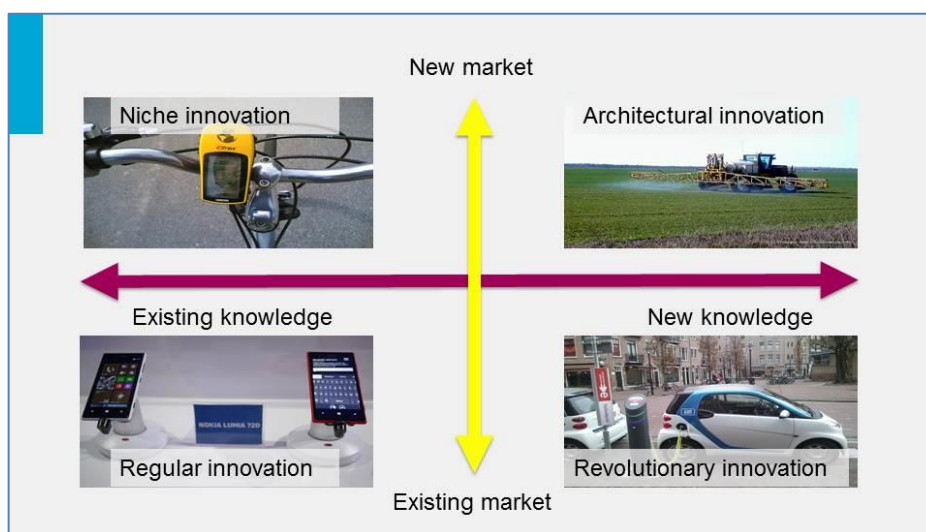

Figure 6: taxonomy for innovation by Abernathy and Clark

Let us briefly discuss the different kinds of innovation:

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

The link between radical innovation and responsible innovation

Architectural innovations have a few identifiable characteristics. First, they only occur once in a while. Secondly, they lay the base 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. Think for example of Apple entering the mobile phone market with their 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 into account the values 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 privacy of smart meters into account, we need knowledge about what privacy is. We would 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. So, responsible innovation may change the relation with customers, and 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, incumbent companies would not always be best suited to introduce responsible innovations, and such initiatives might typically come from outsiders or newcomers.

Ethical considerations of radical innovations

We can also ask if 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 pre-natal 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 disease, possibly one without a cure. This information raises completely new ethical questions for parents. What should they do? Should they consider an abortion?

For existing technologies, there are often moral habits and rules. In cases of incremental innovation, these rules and habits are usually still adequate. But in case of radical innovation, the same rules and habits are often insufficient for the new conditions.

A good example would be the time during 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 of 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.

Case study: Coolants

Coolants are used in everything from refrigerators, air conditioners, fire extinguishers, aerosol sprays, medical devices and even semiconductors. From the early 1930s until the early 1990s, virtually all domestic refrigerators used ozone-depleting chemicals - chlorofluorocarbons, or CFCs - as refrigerants. CFC-12
for example had a number of very attractive properties. It was inflammable, nontoxic, had very good thermodynamic properties and it was chemically stable. However, it contributed to ozone layer depletion. When it was discovered that there was a growing hole in the earth's ozone layer, these chemicals were eventually banned.

Unfortunately, the alternatives that were adopted - hydrofluorocarbons, or HFCs - while less damaging to the ozone layer, turned out to be harmful in a different way. They are actually powerful greenhouse gases that contribute significantly to climate change.

Thankfully, Greenpeace successfully initiated in the nineties the development - convinced that there was a way to avoid HFCs through innovation. - of 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 refrigerator industry world-wide. Today we have over 700 million refrigerators using this ´Made in Greenpeace´ technology
In this section, we will briefly explore the interesting story of how green freeze became the modern coolant of choice.

In the search for alternatives, three values played key roles: safety which was specified as inflammability; health which was specified as non-toxicity; energy efficiency operationalised as having good thermodynamic properties; environmental sustainability specified in terms of ozone depletion potential; and most practically, cost, availability and compatibility. As you can imagine, all these values are important, but it is not easy to satisfy all these values at the same time. We will focus here on the conflict between safety, health and sustainability.
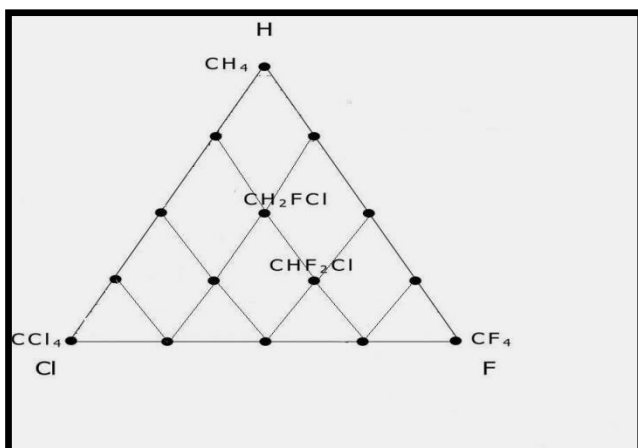
Figure 7: CFC

This figure above is a graphic representation of CFCs based on a particular hydrocarbon chain. At the top, there is methane or ethane, or another hydrocarbon. If one moves to the bottom, the hydrogen atoms are replaced by chlorine atoms if one goes to the left. And by fluorine atoms if one goes to the right. Now, properties related to safety, health and the environment can be mapped along these spectra.

As we move towards the upper-right corner, substances become increasingly toxic. In terms of health, we should better move to the bottom-left. As we move towards the top, substances become more flammable. So, in terms of safety, substances near the bottom are preferable. However, as we move to the bottom, the atmospheric lifetime increases which in turn means that both the ozone depletion potential and the global warming potential increase. From a sustainability perspective therefore, we should move towards the top. But, we cannot move in all three directions at once. We face a value conflict. Initially, this dilemma was s solved by choosing a coolant that met all three values to some degree.
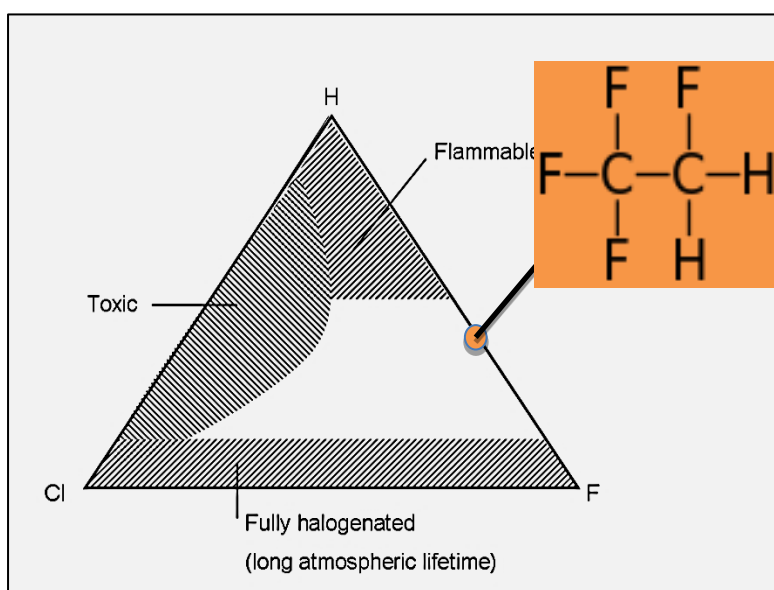

Figure 8: CFC coolant

This figure is from a publication by two engineers. They argued that all coolants in the blank area are

acceptable. The industry indeed chose a coolant from this region, namely HFC 134a. There were several reasons for this choice. First, it was attractive to the chemical industry because this coolant could be patented.

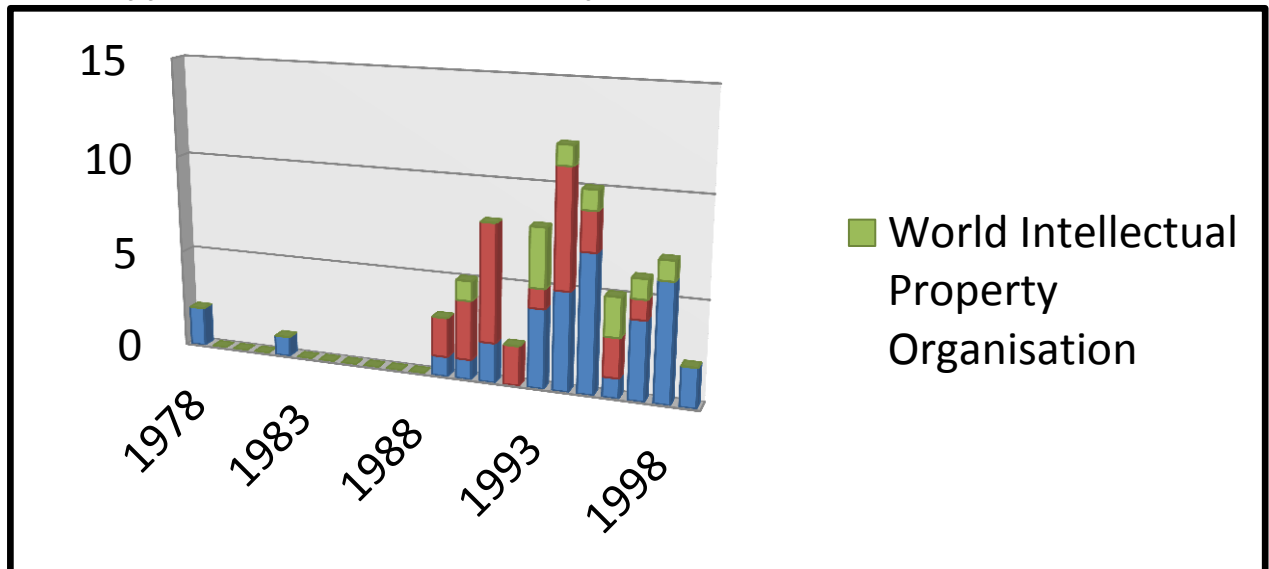As the following graph shows, several patents have been granted for it to chemical firms.

Figure 9: patents

For the fridge industry, the price of a coolant is only a very small part of the price of the entire fridge. Therefore it was more important that a new coolant that served refrigeration purposes would be available. So, they followed the choice recommended by the chemical industry. Indeed, by around 1990, there was a general conviction that HFC 134a was going to be the new universal coolant.

The editor of the International Journal of Refrigeration expressed it as follows: "In my nearly 25 years of working. I have never seen this industry put so much time and effort into one problem as they have into the CFC problem. It is doubtful that any coolant has been tested more in as short a period of time as 134a." *He could not have known then that the future would be quite different!*

To understand the success of the GreenFreeze, let us look at the properties of three coolants: the original CFC 12, the industry-recommended alternative HFC 134a and isobutane used in GreenFreezes. Today, GreenFreeze is used in three hundred million fridges worldwide.

If we compare the latter two, isobutane contributes far less to the greenhouse effect but, it is flammable. This was a main reason why the fridge industry opposed it initially. In fact, existing technical codes banned the use of flammable coolants. Still, it came to be accepted. Why? To understand, we have to look at the interpretation of safety as a value.

Before, safety was understood as the inflammability of coolants. 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 fridges contain only small amounts of coolants. When the standards for inflammable coolants were formulated, fridges still contained much more of the coolant because they had a much lower efficiency. Moreover, it became possible to further diminish the ignition risk by utilizing a clever refrigerator design.

In summary, the GreenFreeze innovation was a responsible innovation because it met the values of safety, health and sustainability. It was also radical because it was the first time in sixty years that a flammable coolant was used. And, third, it was initiated by outsiders, in this case Greenpeace and the East German company Foron.

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 also, what their motivations are. Then we have to understand how we can scale up innovations and 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 these 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 then 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, such 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 that find and provide technological solutions for daily problems.

The determinants of innovation

Let us now draw attention to 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. The uncertain benefits originate from the fact that novel ideas should first prove themselves useful before customers 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 per cent. This means that only 1 in more than 1000 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 discussed. We can start with the external factors: these are factors outside the firm such as the technical, economic and legal environment. We end with internal factors, which play a big role inside the boundaries of 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 say, more mature 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 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 lead to more innovations. 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 after possible innovation failures.

Empirical studies are rather inconclusive. It seems that the technical environment is an important factor. For example, 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 it 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 get 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 it is
in the experienced 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 do meet each other, share a few drinks and exchange views on their tacit knowledge regularly. Other examples also exist, such as 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 i.e.benefits of R&D of an innovating firm spills over its boundaries and increases the benefits of another innovating firm. A second advantage is that the local labour force is technologically welleducated. Yet another advantage is that suppliers and distributors, as well as other supporting firms such as accountants, lawyers etc., will be present at hand, which increases the chance of commercial success of 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 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 design as well as how they relate to each other in a technical system. Then they produce and sell their own version of the innovative product or system, but without incurring the high development costs 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 is not allowed to 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, driving industries and the economy forward.

Finally, the fifth determinant 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 is caused by 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 with procedures. However, procedures and creativity are two opposing forces. Creativity is intrinsically unpredictable whereas procedures are developed to make the innovative outcomes less unpredictable.

That means that a centralized hierarchical organisation of a firm 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 a workable idea.

Hence, ambidextrous organisations are often considered as 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.

4.3 Management of Innovation

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 while trying to be successful and profitable.

In order to manage innovation, acting like an entrepreneur when necessary, it is necessary to understand both the innovation creation and subsequent diffusion process in companies. The way innovation has to be managed, and the way entrepreneurship enters the process, strongly depends on our perspective on these processes.

Management of innovation in companies

We have seen earlier the different types of innovation are distinguished. For now, we will particularly focus on product innovations -i.e. new-technology based innovations. Examples of technology-based product innovations from history are: communication appliances such as telephony 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 deal 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 and that 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 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.
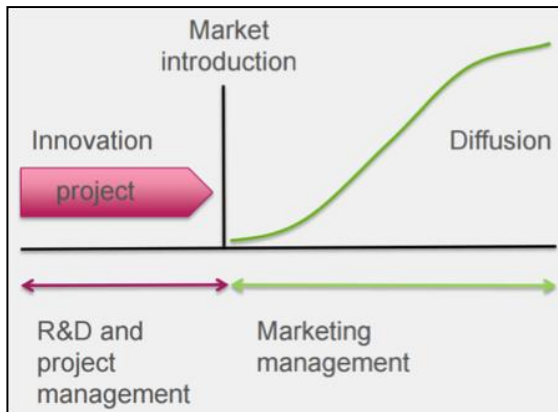
Figure 10: Marketing strategies in market introduction plan

From the 1980s on, mainstream innovation management handbooks presented innovation as a project. From this perspective, the management efforts require close interactions between marketing and R&D teams. It depends on the type of company and the particular industry of course which of the two competencies are in the lead, and that leading department usually provides the project manager going forward.

The success of these joint efforts is reflected in a large increase of sales or in a steep diffusion curve after introduction. In innovation management literature, we find a considerable body of knowledge on the so-called Marketing/R&D interface. This line of thinking continued until the turn of the 21st century. 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.

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 just a product development project.

Firstly, technology development and product development usually proceed in parallel. Usually the first products are unreliable and therefore, 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 just be introduced out of the blue. The initial market usually lacks all types of complementary products and services (e.g. an infrastructure), or sometimes cost-effective production facilities are not yet available. Materials such as nylon, and strong fibers such as Dyneema were developed long before their largescale production was even possible. Sometimes consumers do not really understand the product. As a result of all the elements that might be lacking, the market introduction becomes a bit of a trial-and-error process in which development proceeds even as the product is already diffusing in the market. So, market development and product/ technology development proceed in parallel.

And 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 mastered the principles of flight.

Case study: 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 that implies for the types of management and entrepreneurship that are required to complete these processes successfully.

The invention of the principle 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 (on average a decade) to turn an invention into the first product. This is the first so-called innovation phase in the larger process.

When the he television was first introduced, it was not the appliance that we know today. At first, televisions in Germany and the UK were introduced 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 1950's on (more 20 years after the original invention) did the large-scale diffusion of televisions begin 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 case in terms of the time between invention and large-scale diffusion.
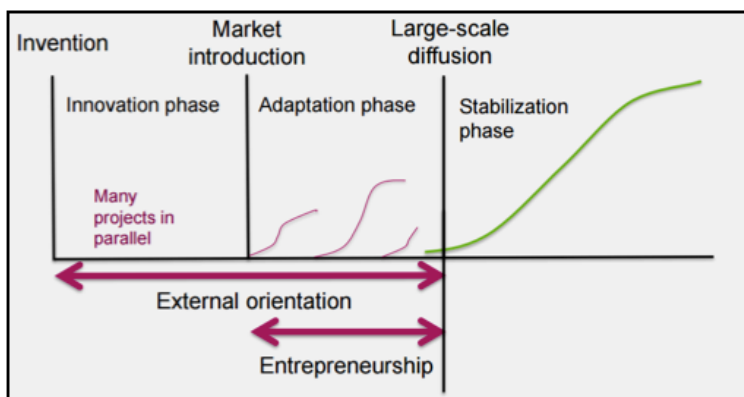

Figure 11: The resulting process

The modern innovation process

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

Firstly, companies need to be oriented more externally 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 by rival companies and to track the latest market developments.

Secondly, entrepreneurial competencies are required to develop a market. The second phase - adaptation phase - is usually when many companies go bankrupt or leave the market before they crash and burn. Most importantly, entrepreneurs are required to bring a vision for a certain product - combining the technological functionality with a (latent) market need and designing a product that fulfills 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.
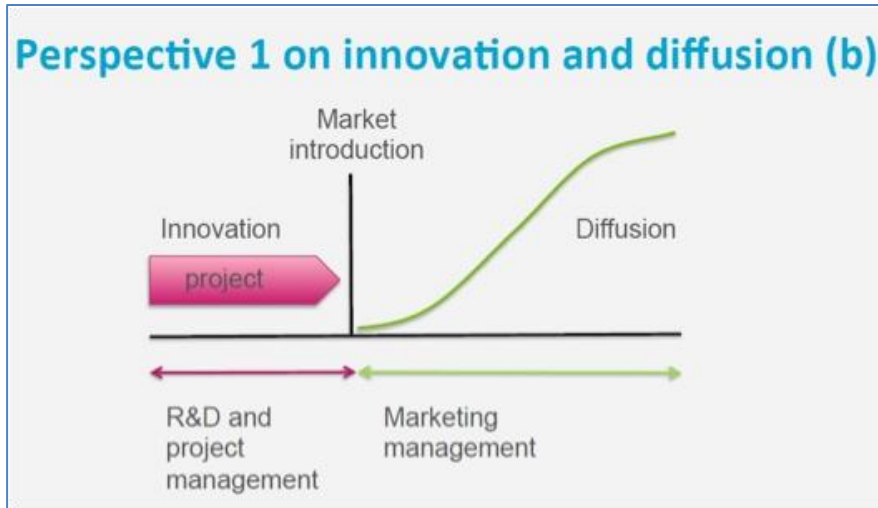
Figure 12: management contribution in innovation and diffusion

Here, there are some implications for responsible innovation. In the adaptation phase (he 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 also 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.

5. Frugal Innovation

5.1 Introduction to 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, and 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". Take note! 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.

What is frugal innovation?

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" - i.e. almost four billion people 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 anywhere between 50% to 97% (Rao, B. C., "How disruptive is frugal?", published 2013).

From an economic perspective, frugal products and services seek to minimize the use of material and financial resources in the complete value chain with the objective of substantially reducing not just the price point, but the complete cost of ownership/usage of a product; and 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. So, the design - and the mind-set of the designers - has to take this into consideration. It has to serve users who face extreme affordability constraints, in a scalable and sustainable manner.

Generally we can distinguish two versions of frugal innovations. The first type is an existing technical product, service or system that is stripped from its luxury attributes but its basic technical functionalities remain intact in order to guarantee the quality of its workings. 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 cell phone Nokia 1100. This cell phone was targeted at low-income users in developing countries who do not yet require advanced features beyond making calls and SMS text messages.

The second version of frugal innovation is creating new technical products or systems that originate from demand by the potential customers, such as a frugal thermometer. This is a thermometer developed in 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 such that illiterate people are able to use it in a responsible way. For example, the body temperature can be measured by holding a scan to the forehead. The temperature can be read in colours. Red means: go to the physician, green means: everything is fine.

The case for frugal innovations

There are some misconceptions with regard to 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'' (BoP) 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 largely 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 customers are in the Bottom of Pyramid (BoP). Again, these are people that have to live off less than US$2 a day. This is more than half of the world's population 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 Bottom-of-Pyramid customers, which would create opportunities leading to 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 the unorganized and fragmented local markets into an organized private sector market where, in the future, products can be supplied at much lower costs than today.

One example of multinational firms providing frugal innovations to the BoP markets is General Electric Healthcare, with 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. It costs about US$1000, which is just a tenth of the price of electrocardiograms developed for the US market.

The link between frugal innovation and responsible innovation

There are two elements of responsibility when speaking of frugal innovation. First up, 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 will 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 low-middle classes, and particularly those in the BoP, have to be thrifty and can only afford frugal products at low prices. Here we see a tension between the cost increasing social and sustainable regulations and the more prosaic immediate needs of the world's poor. This requires design processes of products different than we are used to in high-income countries so far.

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 whole new business eco-system should be designed. Such an ecosystem means that the innovating firm has to collaborate with external partners, such as governments, NGOs, but also local entrepreneurs.

Local entrepreneurs especially can be very important for two reasons. Firstly, they can be a clear distribution channel of 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 done at the innovating firm. Note that BoP customers may live in extreme resource-constrained environments. Solving around these constraints is 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 knowledge of this to the innovating firm.

5.2 Case Study: TAHMO Weather Stations

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

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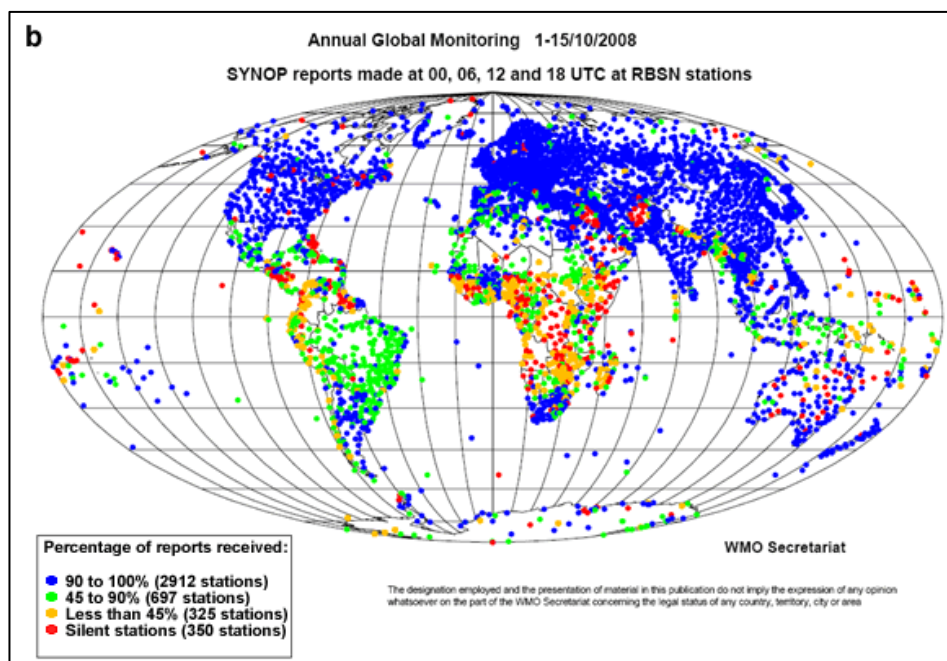

Figure 13: operational weather stations around the globe

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

Maximizing functionality and minimizing costs

Researchers from Delft University of Technology, together with other researchers at Oregon State University are trying to build a self-sustaining observation network. Each TAHMO station is a simple stripped version of an existing product (weather stations as we know them) but uses 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 $5000 and US $15000. This would be prohibitively expensive for Sub-Saharan Africa. Moreover, such stations demand specialized technicians for their continued maintenance. This is why we aim for low-cost and robust weather stations that hardly need maintenance.

There are many considerations to take into account the harsh environment as well. 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.

Figure 14: insects in and around moving parts, Ghana

Figure 15: web of caterpillars around sensors, Ghana

Similarly, a standard weather station usually has a nice and well ventilated screened housing for temperature and relative humidity sensors. When we opened such a housing in Ghana, we noticed a web of caterpillars around the sensors. We also want to reduce the costs of our station. This is a typical research grade radiation sensor that costs about US$6000. By using mass-produced sensors, we hope to significantly reduce the costs. 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 nice 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 how much rain falls 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. So our researchers had to come up with a robust design without moving parts.

After trying several materials, we found a simple piezoelectric element which can be found in any smoke alarm and costs about US$1. Such an element produces an electrical signal when it is mechanically excited. In other words, when a drop falls on it, a signal is produced and this 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.
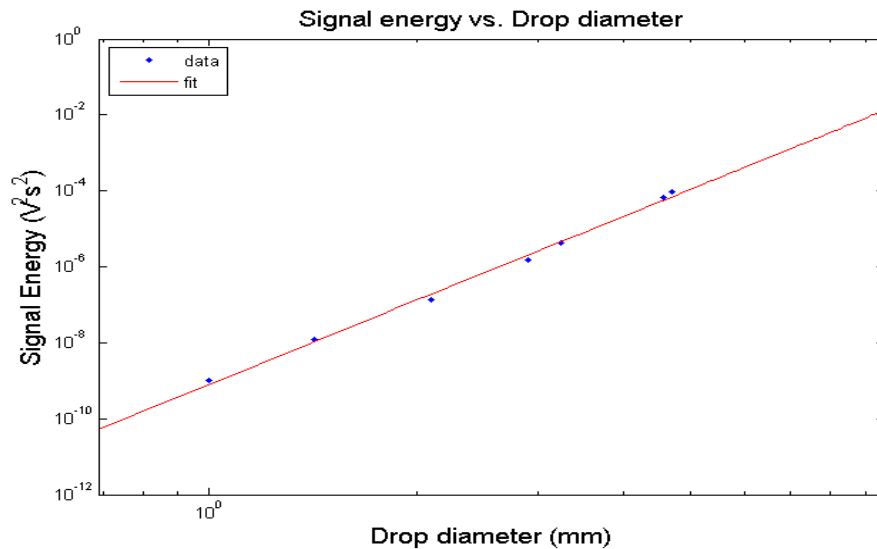
Figure 16: correlation between drop size and signal strength

Still, there is a lot of work to do, which is why we are linking up with companies like Decagon and IBM to speed up the development of the TAHMO network.

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.

We have done some early pilots in Ghana to see what needs to be done to include weather and water stations in the curriculum. We are also piloting a school-to-school program in which richer schools pay 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 is then followed by a series of lectures on climate, water, weather and the exchange of information between the sister schools. The first exchange happened in 2014 between schools in Idaho and Kenya[4].

We are also developing cooperation between African universities in support of TAHMO. In 2013, we ran a sensor design competition where teams were tasked with designing new sensors along the TAHMO design criteria. First, we asked interested parties to register, and then send in their first designs. There were 43 registrants, resulting in 23 design submissions.

One example of such a design 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 being used for alternative uses as well. 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 was an attempt to determine wind speed and wind direction by measuring the movements of a flag. Strictly speaking, this design does have moving parts but a relatively robust one. (And besides, who doesn't like a waving flag?)

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

Eight teams were invited to the final in Nairobi in August 2013. There at the iHub Innovation Center, the teams built their sensors and integrated them through a Raspberry Pi, and then uploaded all the incoming data to IBM's

online system for data and operations.

Business models for the TAHMO project

At the same time we see that scaling up is an important issue right now, and 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. So, we are trying to develop public-private partnerships and business cases that are financially compelling to build, operate and sustain TAHMO stations.

The initial financial numbers are significant but not staggering. It is probably necessary to start off on the basis of some grants and subsidies. 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 are about US $31 billion per year. In Africa, we would only need to capture a very small percentage of the value produced to maintain the program. 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.

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 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 information for taking business decisions; for instance, decided when to seed crops. Similarly, insurance companies would be interested in these data as they provide them with better estimates on crop failure, which underlies their calculations for the crop insurances they sell. Thus farmers insure their inputs and insurance companies pay them out when rain fails. Forecasting of rainfall could be a service provided by nearby weather stations. We are now partnering with Kilimo Salama in Kenya, an insurer, who leverages the possibilities of mobile phone networks to sell insurance and organize payments.

In conclusion, we think that TAHMO will be able to let 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. If you are interested, please visit our website www.tahmo.org or use our ''Contact us'' form at http://tahmo.org/contact-us/ .

5.3 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 if and when the social standards applied in the production processes are either too low or too high.

What are social standards?

Social standards, sometimes also called labour standards, have two main elements.First is ensuring decent working conditions for labourers such as implementing a reasonable minimum wage, and installing proper health and safety precautions. In a factory context, this includes simple things like making sure the fire extinguishers really work. On farms, this responsibility could mean providing protective clothing to 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 even worse, mistreated by employers.

These 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 just from looking at a T-shirt, or from drinking a cup of coffee, whether the processing of that product involved some child labour. We call this a 'credence' good, which means that we need to put 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 are more effectively monitoring social standards compared to others, all governments face challenges with 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 as possible working conditions. What it means is that one cannot only look at the technological or ethical dimensions of the product as such, but we 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 means also 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 entrenched firms use the argument of unnecessarily high social standards to block the entry of new firms into their markets, and this is not done for ethical reasons.

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

5.3 Innovation and inclusive development

Another dimension that frugal innovations have to satisfy in order to qualify as responsible innovations is to answer 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 it systematically.

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 $1.25 a day live in two regions: Southern Asia and Sub- Saharan Africa. Nearly two thirds of these people, the extreme poor, can be found in five countries: India, China, Nigeria, Bangladesh and the Democratic Republic of Congo. However, since 1990, the GDP growth rates have been quite high in these regions of the world, as seen in the following table.

Region / country	Average GDP growth p.a. (%)	
	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 1: average CDP 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 earn 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 also 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 or other critical goods and services remains elusive for many people who live in developing economies.

These trends in growth, poverty and inequality highlight the character of current development. In many countries, people have been and still are excluded from the fruits of economic growth and development. This exclusion takes two forms. On the one hand are those which have gainful employment or access to land, but who are 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 those who are middle- or upper-class. Taken together, such innovation trajectory 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?

Serving poor consumers

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 between households or, in some cases, a single purchase for an entire village or community organisation (an oxen plough, a water pump, a weather station, and so on).

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. Then more people can afford the product or service, and therefore the frugal innovation will be more inclusive. An example is the OMO sachet for washing in cold water. Providing a small portion allows many 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. This figure below depicts nine different product characteristics which may reflect consumers' incomes.
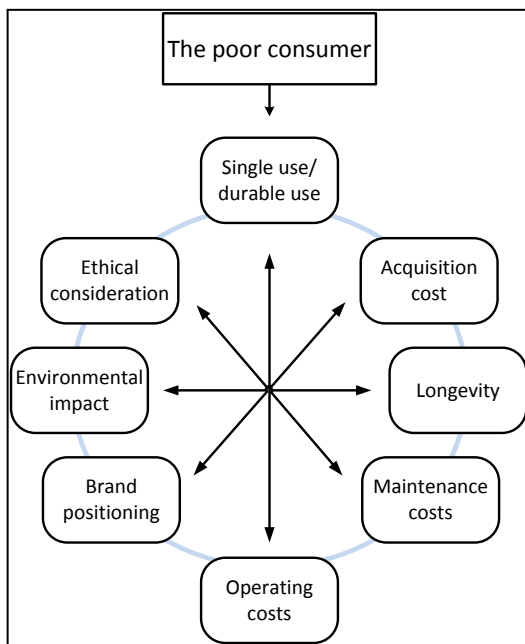
These characteristics reflect whether the product is for single use or repeated use, acquisition cost, longevity, costs
of maintenance, operating costs, brand image, impact on the environment, and the extent to which that product or service has characteristics which reflects local, environmental and ethical considerations. Frugal innovations typically reflect a characteristic that match with 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. So, the inclusiveness of frugal innovations may be at odds with some other dimensions which would make them responsible innovations.

Figure 17: product characteristics

Serving poor producers

We can also ask if frugal innovations are more inclusive towards poor producers. 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 are have to use their own or family labour as very little capital is available to them. Production is generally done small scale, using unskilled or semiskilled labour.

As a general rule then, 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 involve in Exportation 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 auto-mobiles and watches

Table 2: Innovation for poor producers and innovation for poor consumers: some examples (adapted from Chataway et al. 2013)

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

There might still be inclusive effects, for example, in the decentralized marketing and distribution of these

products, which can create employment for poor traders, salesmen and women, as well as through the employment of unskilled or semi-skilled labour for the production stage. But generally the role of local producers will be limited, unless they are able to become part of the value chain of the multinational company. For example, they might do so by becoming part of the marketing and distribution network of the multinational, or by becoming a local source of inputs and information.

This might be different from frugal innovations which originate from the informal sector itself. Local producers are involved in the design, production and marketing of these innovations, there is local ownership as it were. But the spillover 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, it is not by definition that frugal innovations are also inclusive innovations, in the sense that they allow for 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".

Poor producers do have some comparative advantages to multinational companies however, when it comes to the design and production of frugal products and services. For example, they know the demands and preferences of local poor consumers better, and they are less vulnerable to reputational damage which may arise because of 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.

Case Study: revisiting TAHMO weather stations

We can find links to the discussion of frugal TAHMO weather stations we saw earlier. Let us briefly explore the inclusiveness of frugal weather stations.

If you recall, these weather stations have been developed by Delft University of Technology together with Oregon State University, and will be distributed and marketed in West Africa and Kenya. One attribute which makes the frugal weather station probably more inclusive than conventional ones, is the simplicity of its technology, which allows for low maintenance and operating costs, and it is also easy to handle for people with lower educational levels

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 or NGOs, which may need the data of the weather station to serve poor clients better and reach them with new services tailored for poor consumers (like the weather-based insurance discussed earlier).

For the moment, possible gains for local entrepreneurs in West Africa are 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 the 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 always, because it asks for 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 enterprise. Moreover, through new initiatives like weather-based insurance, they could become less vulnerable to income shocks.

Caveats for frugal innovation

We have shown that frugal innovations are not by definition responsible innovations unless we satisfy the dimension of inclusiveness. Like with the discussion on 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, the inclusiveness is not necessarily guaranteed. Bottom-up frugal innovations may allow for better inclusiveness, but poor producers of frugal innovations still face various constraints for upscaling and creating spillover effects in the local and national economy. Still, the example of the weather stations shows that frugal innovations can have huge potential to be or become inclusive and thus serve as responsible innovations.

Part III

6. Understanding Risk

6.1 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 dealt with. 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 uncovered 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 also of probable harm.

The difference between risk and uncertainty

Sometimes we may assign probabilities to indicate the uncertainty about the harm's occurrence. It is the natural, social or engineering sciences that provides the probability that a certain valve in a nuclear power plant will break, for example. However, what constitutes a 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 that is actually so, and also why we need to care about the environment for future generations. So, risk is per se an interdisciplinary concept; 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 with climate models in the narrow sense, but require so-called welfare economic impact models, made by economists. Hence, better political decisions as to how to react to the threats of climate change may not require better climate models, but better climate impact models – an area of research not nearly as active as climate models today.

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

The 2007 report by the Intergovernmental Panel on Climate Change (IPCC) predicted the following threat of global warming when CO_2 concentrations double: "temperature rise is likely to be in the range 2° C to 4.5° C with a best estimate of about 3° C, and is very unlikely to be less than 1.5° C. Values substantially higher than 4.5° C cannot be excluded". The 2007 report still associated terms such as likely and very likely with probability ranges. For example, likely would mean probabilities larger than 66%. The latest IPCC report however does not provide such probability estimates anymore.

Without these probability estimates for the occurrence of harm, we say in technology assessment that reacting to climate change does not actually 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. 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 where not even the probability space is known. This recently became famous as the former US Defense Secretary Donald Rumsfeld termed these the "unknown unknowns", or what Nassim Taleb calls the "black swans".

Mitigating climate change is an example for 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 large parts of engineering and science.

There are also advocates of a more subjective view, in which probabilities are grades of belief instead of relative frequencies. This is sometimes referred to as 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 fragmentary, as we cannot update our beliefs and assigning a priori probability distributions in Bayes formula is difficult.

So, how do we deal with risk and uncertainty?

For risk, we may use what is known as maximizing expected utility analysis, or, formulated negatively, risk minimization or risk analysis. This is nothing else but the utilitarian paradigm of the greatest good for the greatest number, but as 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 for 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 different 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 referred to as a weak formulation, because it advises us to also 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, at least making 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 however - also known as the strong formulation - does advise us on how to act. It tells us even 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.

Precautionary Principle and moral overload

Applying this formulation to the issue of climate change means that we need to mitigate any impacts climate change may have. Here it comes in handy as some economic assessments suggest, that reducing anthropogenic greenhouse gas emissions is not very costly. 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 damages 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 put it into relation with other figures. Consider that US $1.3 billion per year is needed to fulfil one of the UN's Millennium Goals, which is to provide 80% of the rural population of Africa with safe water and sanitation. So this comparison shows that simply applying the Precautionary Principle falls short of adequately accounting for this comparison. If we do invest 1% of global GDP per year to avoid climate change, we must accept that that money is not there for other goals. So... how to 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 be quantified in terms of a probability.

6.2 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, the computer he was thinking of was large mainframes like ENIAC. He could not have known that in just a few decades, there would be PCs, laptops, tablets, smartphones and so forth, 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 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 so, 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 time severity. However, the problem is that we often do not know the probabilities, in which results in uncertainty. Sometimes we do not even know all possible consequences – so this end up in ignorance. As a consequence, we cannot actually determine the risks!

An alternative approach is the Precautionary Principle. There are various formulations of it, but an often used formulation is the following: when an activity poses threats to the environment or to human health, precautionary measures should be taken, even if some causeand-effect relationships are not fully established scientifically. Note that this principle does not require the establishment of probabilities. It can therefore deal with what I have called uncertainty.

Drawbacks of the Precautionary Principle

But, this 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. This proposal led to heated opposition!

Now 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 so, on the grounds of the principle alone we cannot make a decision!

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 below.

Figure 18: unkown unkowns (painting by Robert Meganck/ http://meganck.com/mix/mx_04.html)

The image above suggests that there will always be surprises and unexpected developments when we introduce new technologies into society. 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. And then to ask the question: under what conditions are such experiments morally acceptable?

Case Study: nanoparticles in sunscreens

We can illustrate this by a case study on titanium dioxide nanoparticles in sunscreens. Some fear that these nanoparticles might cause cancer. In 2006, the International Agency of Research on Cancer mentioned titanium dioxide as a possible cause of cancer. However, titanium dioxide comes in different sizes and it is unclear whether the nanoparticle size is also dangerous. Nevertheless, some opponents of nanoparticles already call such particles ''the new asbestos''.

In the same year, the Health Council of the Netherlands offered the following advice on nanoparticles, based on the precautionary principle. They said: before nanoparticles are brought onto the market, their toxicological properties should be properly investigated. The goal of this investigation is to make nanoparticles a 'simple' risk problem instead of an uncertain risk problem. This is sensible advice. However, the second part overestimates the possibility for assessing risks beforehand.

For three reasons, the risks of nanoparticles can only be determined beforehand to a limited extent. First, they may have long-term cumulative and interaction effects that cannot be tested in the lab. Second, laboratory tests are often not representative of real-life circumstances. And thirdly, we have ignorance, or

unknown unknowns. An interesting example of ignorance is the appearance of defects on pre-painted steel roofs. These turned out to be caused by titanium dioxide from sunscreen used by workers during installation.

Responsible innovation as acceptable social experiments

Due to the aforementioned reasons, the current introduction of titanium dioxide particles in sunscreens is a kind of social experiment. But is it also an acceptable experiment?
We have formulated four principles for acceptable experimentation in this case.

1. The absence of alternative ways of acquiring the knowledge required for a complete risk assessment.
2. The controllability of the experiment. This includes the monitoring of possible effects, the feedback of such effects, the containment of possible effects and a conscious upscaling of the experiment.
3. Informed consent: informed consent is one of the main ethical principles to judge the acceptability of experiments with humans. It states that human subjects should be completely informed about risks and expected benefits. They should also freely and knowingly consent to participating in the experiment. In this case, this translates into two more specific conditions, namely that a) consumers should be informed and b) consumers should be able to end their participation in the experiment.
4. Proportionality of risks and benefits: in this case, the risks were still so uncertain that we reformulated this condition. It now reads as follows: there should be a continuous review of the risks. And decisions about continued use should be based on such a review.

If we apply these conditions to the case at hand, it turns out that some conditions are not fulfilled. First, more risk assessments could have been done before introducing nanoparticles to society. Also, we find that the monitoring, labelling and continuous review were absent. So, we made the following recommendations: first, close the existing knowledge gap as far as possible; second, monitor the possible effects of nanoparticles; third, if necessary, take action on basis of such monitoring; fourth, conduct ongoing design for safety. One possibility is to design nanoparticles in such a way that they can be traced, as this would be helpful in monitoring. And the final recommendation would be to change the law and legally require monitoring and labelling.

Applying the Collingridge dilemma

Let us briefly return to the Collingridge dilemma. We have seen that anticipation tries to attack the first option of the dilemma by making technology more predictable. There are however limits to that.

Treating new technology as social experiments deals with the second option of the dilemma. It accepts that some effects only become clear as technologies are introduced in society. However, it tries to avoid that technology will get entrenched in society too fast, without proper monitoring of developing consequences. Hopefully, this chapter has made a serious case for this alternative approach.

6.3 Technology Assessment

We can now ask, how do we achieve responsible innovation? What kind of approaches can we take and how should we use the tools available? This line of questioning falls under technology assessment. Responsible Innovation approaches in general have evolved from the larger practice of 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 (ELSI); the 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. Its history goes back to the late 60s to early 70s. It is 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 is also a form of impact assessment. It can be described as an attempt to objectively predict 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 counterpart organisation.

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. This evolution of TA can be described as follows: from the objective prediction of expected consequences to anticipation of possible consequences; across governments, companies and research organizations; from reactive to proactive approaches: and even influence 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. These are: Constructive Technology Assessment (CTA), Midstream Modulation and Network Approach for Moral Evaluation (NAME).

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: first, learning about social consequences, second is reflexivity - which implies awareness of the other actors - and the third is anticipation of possible technological development and possible social consequences. In striving for these goals, CTA also aims at broadening technology 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.

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 broader social dimensions of their work. The term midstream is used to stress that the method focuses on modulating R&D practices.

The reason for this is that, rather than making the 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. (A nanoscale is 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 Nano technological 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 different issues and giving feedback with different perspectives.

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 the network dynamics in order to discern moral issues. The approach also consists of network norms to judge such networks. The two main networks norms in the NAME approach are: first, learning and reflexivity; and second, 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 - like 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.

We applied the NAME method to an innovation in sewage treatment, the NEREDA. This is an innovative wastewater technology that won several innovation prices. It was developed by researchers from Delft University of Technology and Royal Haskoning DHV. The main innovation is that the living conditions of the bacteria that clean the water have been changed. The effect is that they grow in granules instead of flocs. As a result, the sludge will settle quicker, which results in a smaller foot print for the plant.

In our research we found that it was not clear who in the innovation network was responsible for so-called secondary emissions. These are emissions that are not regulated by law, but may nevertheless be harmful: think of heavy metals. This is an illustration of what we earlier defined as the problem of many hands. The graph below shows how the users attributed the responsibility to address secondary emissions to the research phase, while the researchers attributed it to the use phase.
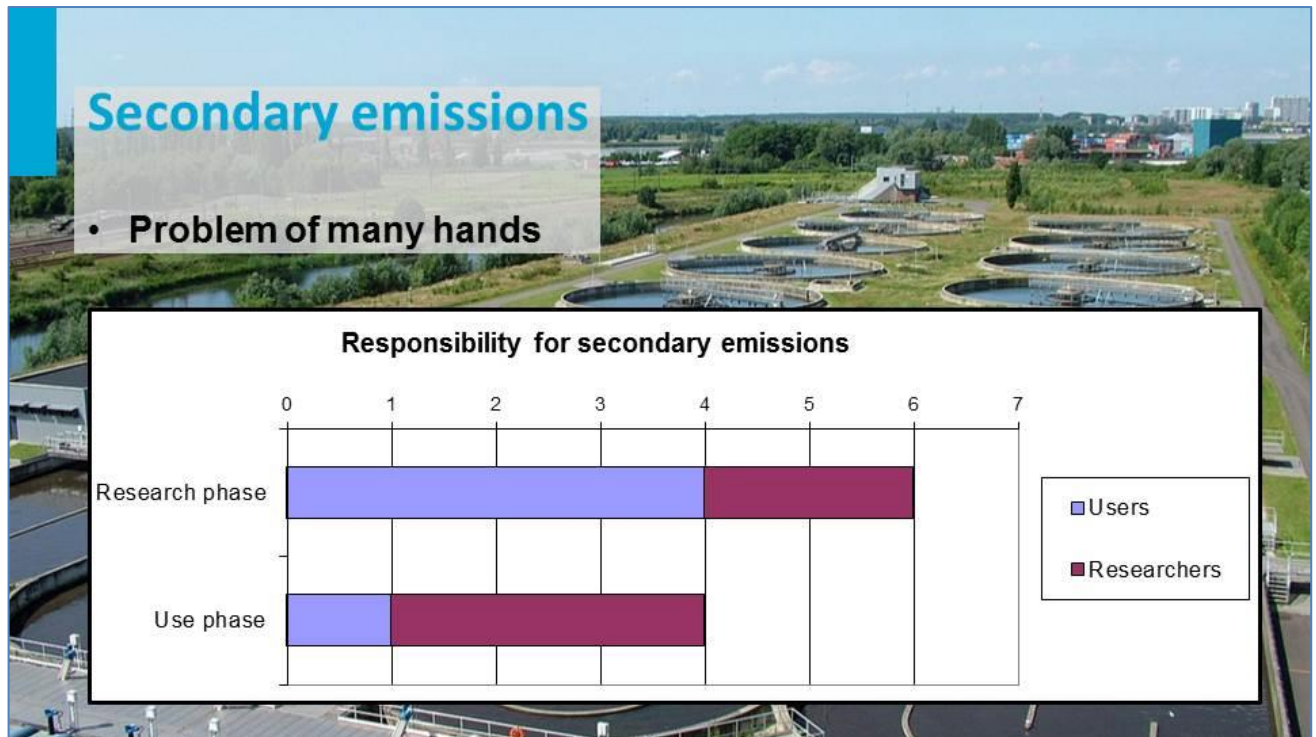
Figure 19: responsibility for secondary emissions

As a result of our findings, the researchers looked better into secondary emissions, which turned out not to pose a problem under the new conditions.

A framework for responsible innovation

Let us revisit responsible innovation. In a recent book, Richard Owen and a number of colleagues describe a framework for responsible innovation which has four main components. They suggest 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: This means that 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 values are inspired by, and hark back to earlier Technology Assessment approaches.

6.4 Case Study: the debate on Nuclear Energy

To put what we have just discussed into real-world perspective, we will present 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 exante analysis, which is an analysis of a technology that does not yet exist. This is actually the more important analysis when we speak of responsible innovation, trying to accommodate the

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 we see there are several values at stake. We will argue here that sustainability is to be considered as a moral value that consists of five other values. Each of these values has a temporal and spatial dimension. Temporal relates to time and spatial dimension relates to space. 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 then 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. So this is how we can frame sustainability.
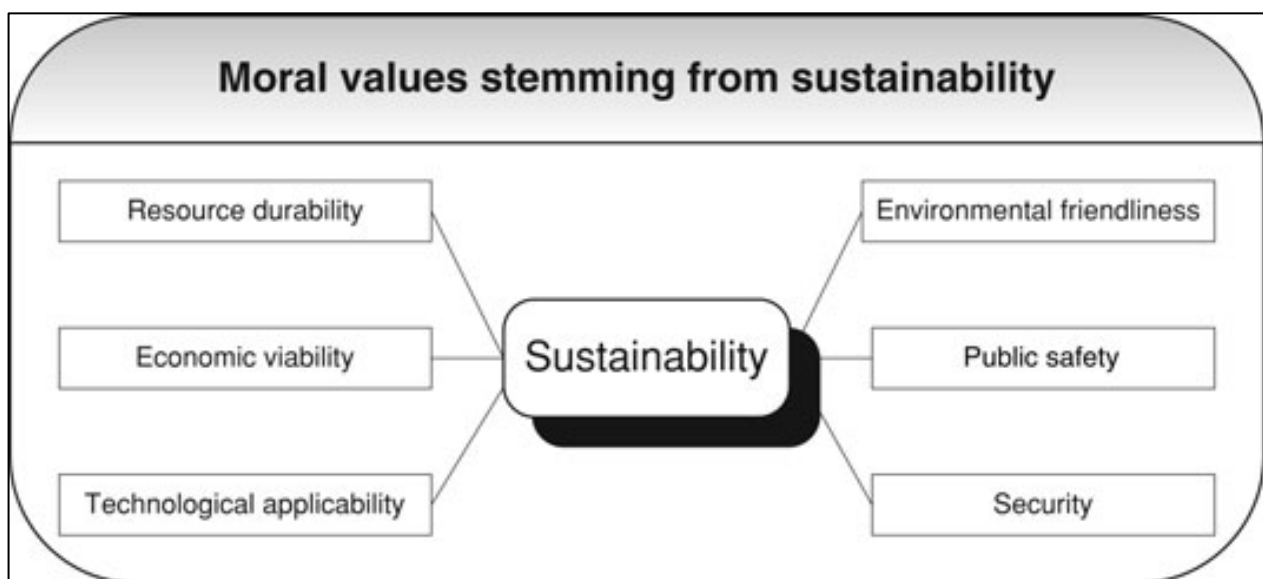

Figure 20. On the right side you will see Environmental Friendliness, Public Health and Safety, and Security. Later, we discuss why security is being discussed separately, and not included with safety. On the left side of this picture, you see Resource Durability and Economic Viability. It is very important to also 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 it Environmental Friendliness. The question that pops up here is: why should we care about the environment? We can approach this answer with two schools of thinking. One is anthropocentrism - which puts human beings at the sole centre of attention. This school of thinking argues that the environment does not have a value as such, so it could only have an instrumental value. The second school of thinking is nonanthropocentrism, in which we argue that the environment has an intrinsic value which may not necessarily relates to what it means 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 nor in the future. A noble goal, but the question here is, how far in the future should we consider and how should we offer protection? And this is the question that very much relates to tangible policy questions.

Let's consider the example of the Yucca Mountain Repository, the world's first and biggest repository being built

in the United States, over the last couple of decades, seen in the image below.

Figure 20: Yucca Mountain Repository (source:
https://en.wikipedia.org/wiki/Yucca_Mountain_nuclear_waste_repository)

When introducing radiation standards for the Yucca Mountain Repository, the US Environmental Protection Agency (EPA) presented several standards. The EPA argued that for the next 10,000 years, we need to 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 health impact of radioactivity or what we call radiotoxicity.) Beyond that period, EPA guarantees a much lower level of protection. Consider that the first proposal was to have 350 millirem and only later, after a lot of public debate, they adjusted the figure to 100 millirem per year. So now, beyond 10,000 years, we are offering a level of protection that is six times less than what we offer the present generation.

This again goes back to the fundamental question. How far in the future should we care and how far 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 we talk about non-proliferation. Non-proliferation again 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. So, safety and security are two notions that are being discussed separately in nuclear energy discussions.

The next value that relates to sustaining human wellbeing is the value of Resource Durability. So here we talk about 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 and to which future generations will not have access?

And the last key value we will discuss here is Economic Viability. For an energy source to be sustainable, it needs to be economically durable. There arise many moral questions yet again. 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 the present ones? And if not, how do we value the future interest compared to the current interest?

In economic studies, this notion of future evaluation becomes important. The value of future interest would actually be discounted for a certain percentage against the present value. Discounting is a very important aspect of a cost-benefit analysis (CBA) and we can perform this in the interests 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. It is important to know about the fuel cycle 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.
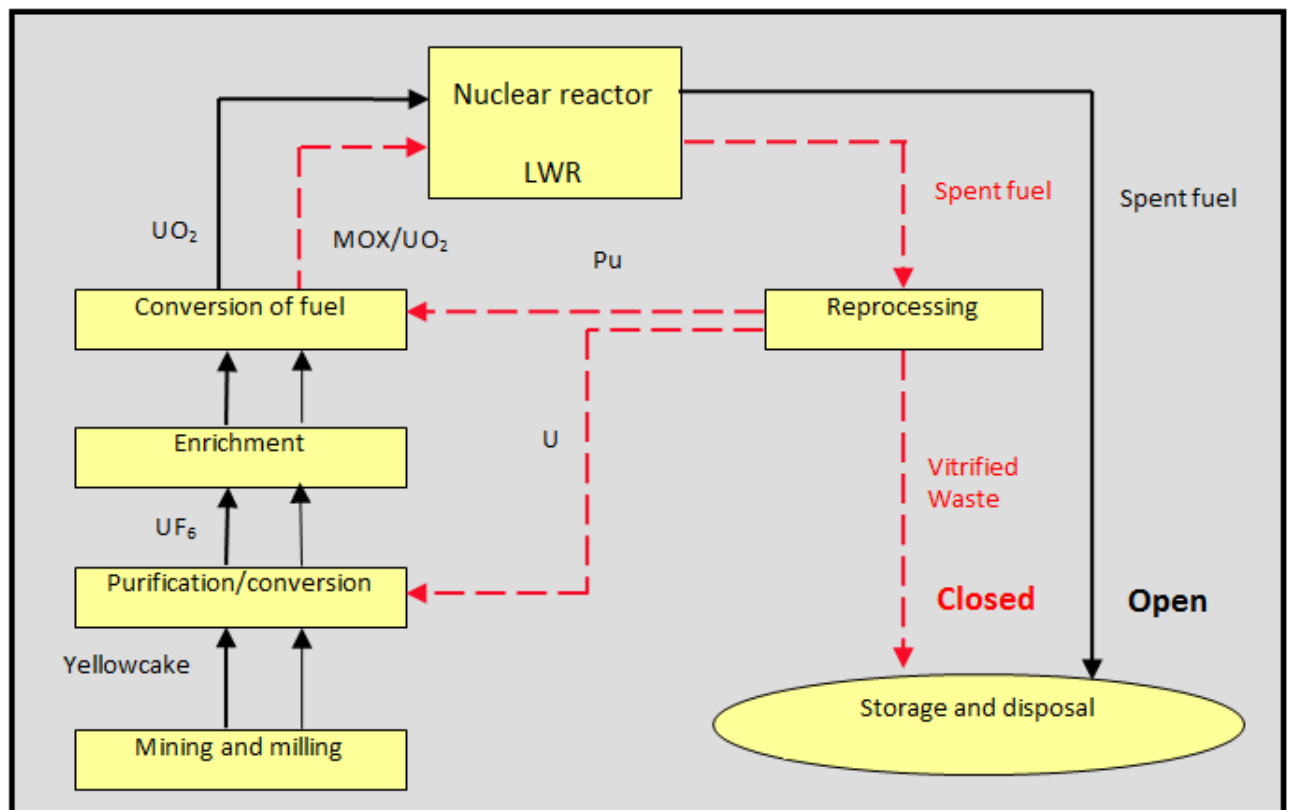

Figure 21: two methods for nuclear power production

The black arrows denote what we call open fuel cycles, commonly used in the US, in Sweden and some other countries. So, first mining and milling uranium, then purifying, converting and subsequently enriching it. The processed uranium oxide will pass through the nuclear reactor. What comes out of the reactor is not 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 from 200,000 years up to a million years.

However, spent fuel could also be recycled in a closed fuel cycle. We call the process of recycling spent

fuel reprocessing. The greatest benefit of reprocessing is that we can salvage still usable material uranium and plutonium - extract and re-insert them into the fuel cycle. A second important benefit of recycling or reprocessing is that we can drastically reduce the waste lifetime. 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. So, reprocessing also brings a very important security risk.

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

Figure 22: relating values to fuel cycles

In short we can relate each of these values to the underlying nuclear fuel cycle, the open and closed cycle. So, the argument presented here is that the open fuel cycle is particularly good for the present generation because it introduces the least burdens for the present generation. The closed fuel cycle on the other hand is best for future generations, because it could help us reduce the waste lifetime significantly, thus benefiting future generations. We can reduce the duration from 200,000 years - the waste lifetime of an open fuel cycle - to just about 10,000 years using a closed fuel cycle. 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

Speaking of responsible innovation, we can also 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 those values, and to address those conflicts prior to developing new technologies.

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

Probabilistic Risk Assessments are being introduced for reducing the probability of a melt-down 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 the 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 - did anticipate that the risk of meltdown of a reactor would be $5\times10-5$, which would be one every 20,000 reactor-years. Be aware that this figure is not years but rather, years of reactor operation. Hence, we call them reactor-years. So, based on 500 reactors, there could be one accident every 40 years. That is, or that was, how the argument went back then, and it was deemed a fairly acceptable risk.

However, it was decided to adjust the reactors because there was considerable growth anticipated. From 500 reactors at a time to an expected 5,000 reactors, which meant 10 times more reactor-years. In turn, this means that any accident, should it occur, would be 10 times more likely. So, going from 500 reactors to 5,000 reactors - based on the same calculation - would imply an accident would occur once in 4 years. That was absolutely

unacceptable, and it motivated serious change in the design of nuclear reactors.

There are two different approaches for making reactors safer. Firstly, we can make incremental changes to the safety. That would mean that we take the current design as the point of departure, and then we add safety features or equally, we remove unsafe elements. The second approach would be a radical approach of changing 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 that do not exist anymore. The second generation (Generation II) of nuclear reactors are the ones with operable reactors right now, all around the world. Beyond Generation II, we talk about Generation III, III+ and Generation IV reactors. Some Generation III reactors are operational right now. But Generation III+ and IV are still being developed.
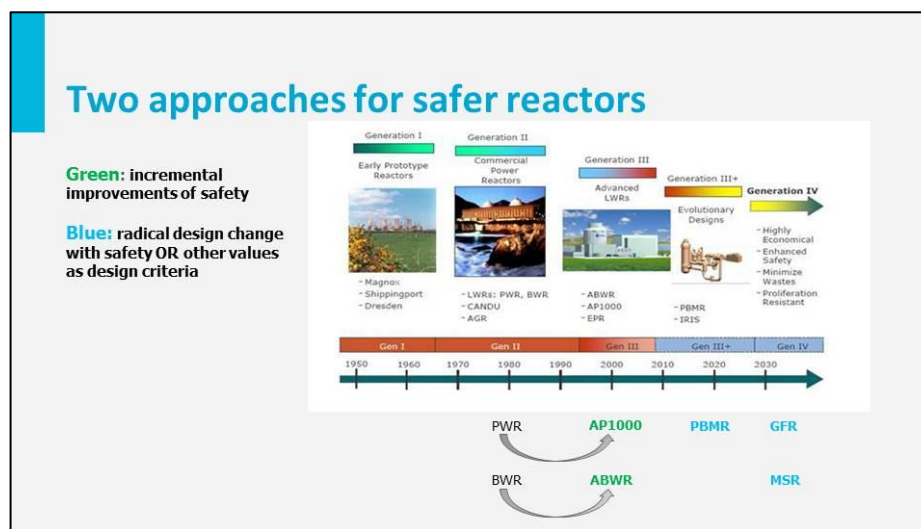

Figure 23: two approaches for safer reactors

As mentioned earlier, there are two different approaches to improvement of safety. One would be by incremental improvement of safety, which is indicated in green. But we could also have a radical change in design and those are indicated in blue. For the radical designs, it is not only safety that is relevant; there are also other values at stake which we should also design for. Generation IV reactors are supposed to be highly economical, and they are also supposed to enhance safety, minimize wastes and be proliferation-resistant; they are in effect designed to accommodate a multitude of values.

So, the paradox of reactor safety is the following. Most of the reactors operational right now are the Generation II reactors - namely, the Boiling Water Reactor (BWR) and the 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 proposed 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 and other kinds of safety features. In turn, these additional features made the design even 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 of the Boiling Water Reactor (BWR) from Generation II. The primary reason for designing the Advanced Boiling Water Reactor (ABWR) was to make the BWR safer. There were many additional safety features, like 10 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.
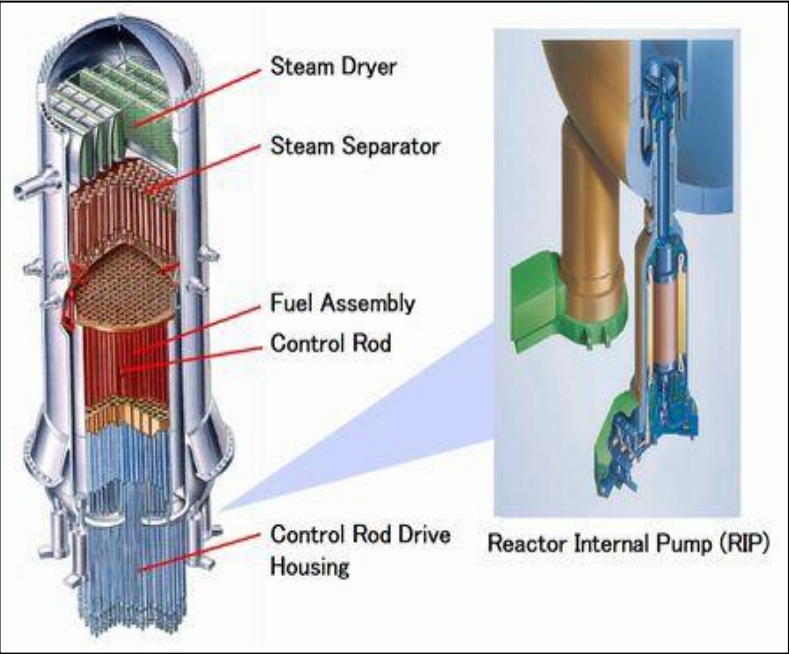

Figure 24: reactor vessel of an ABWR

The image depicts how an ABWR looks. On the right side, we find that the internal pumps are placed under the reactor so less piping is needed; so the complexity decreases, and hence the reactor is safer. There is also redundancy, a key feature of safety.
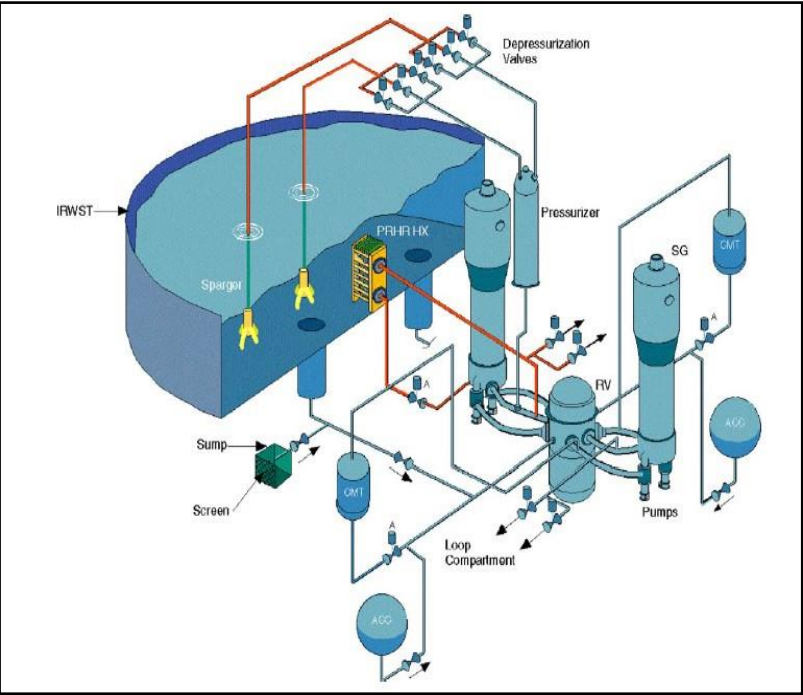

Figure 25: passive core cooling system of the AP1000

The AP1000, shown above, is a Generation III+ reactor. AP1000 is a substantially simplified version of a PWR. PWRs as Generation II reactors are already operational all around the world. AP1000 substantially reduces the complexity of reactor design with fewer valves, pumps, cables and significantly less piping, all of which reduce the possibility of meltdown substantially.

Another important feature of the AP1000 reactor is the introduction of the passive safe system. Here, a passive core cooling system with three sources of water is implemented, and cooling works based on gravity and natural circulation. There are water basins located at a higher altitude than the reactor core. In case of an accident, even without the pumps or electricity supply, water could still flow to the reactor and cool it down. Hence, we call it a Passively Safe Reactor, meaning that there is no human intervention needed.
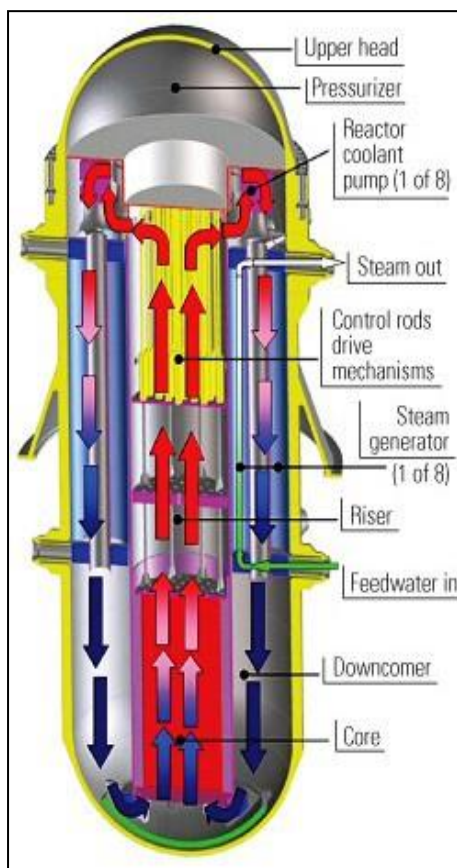

Figure 26: vessel of Pebble Reactor

Next, the Generation III+ Pebble Bed Reactor (PBR). The PBR is a radical design change with two leading criteria: safety and economic viability. This reactor also moves towards another safety paradigm, called Inherently Safe Reactors. So even in the event that cooling fails, temperature in the reactor will remain under 1600°C. There is no need for emergency cooling. Naturally, this goes even one step further than Passively Safe Reactors.

This image depicts the vessel of a Pebble Bed Reactor, and we see that there will be natural cooling during production and, in case of an accident, the same natural cooling will also reduce the temperature inside the reactor vessel. There are some interesting trade-offs here between safety and economics. Due to the inherently safe design, even during normal operation, this reactor loses quite some heat, making it slightly more inefficient. However, the choice is made consciously here to ensure safety and avoid an accident if the reactor keeps heating up.

There is also another essential safety feature in this type of reactor; namely, the type and the shape of fuel, the so-called pebbles.

The name Pebble Bed Reactor comes from these pebbles. These pebbles will never melt because they are made of ceramics that do not melt at reactor temperatures. Even if cooling fails, the pebbles will just never melt - it is physically impossible. However, they could end up damaging the silicon carbide (SiC) coating by temperatures exceeding 1200°C, in turn increasing the risk of radioactive exposure. Accordingly, in the Probabilistic Risk Assessments, we calculate that the possibility of a meltdown is impossible here. However, we do need to introduce a different unit for assessing the risk of PBR, namely to account for the release of radioactivity into the environment.

Let us now look at a Generation IV reactor, the Gas-Cooled Fast Reactor. The leading design criterion for this one is Resource Durability. We make optimal use of a major uranium isotope (238U). First, the uranium is converted to plutonium and then, the plutonium is used for energy production. You might have heard the term "plutonium economy", and it relates usually to this type of reactor.

There is a second important advantage working with this type of reactor. By using them, we could get rid of the long-lived waste after reprocessing. So, they are compatible with an extended fuel cycle. However,

it goes without saying that the use of plutonium raises serious security issues, such as the risk of proliferation.

The last reactor we will discuss here is a Generation IV reactor, the Molten Salt Reactor (MSR). This represents a revolutionary design change with the primary design criterion being Resource Durability. It was first proposed as an aircraft propeller in the United States. An important feature of MSRs is that it could
potentially use thorium (Th) as a fuel source. As a matter of fact, it is one of the very few reactors that can use thorium. Thorium happens to be more abundant as a resource than uranium.

Another important feature of MSRs is that the fuel is circulated and serves also as a coolant. This adds an important safety feature, namely that, in case of accidents, fuel can be drained and dumped into tanks, thus preventing the reactor from overheating.

Indeed there are various R&D investments needed before this reactor can be operational. This is how a MSR might look.
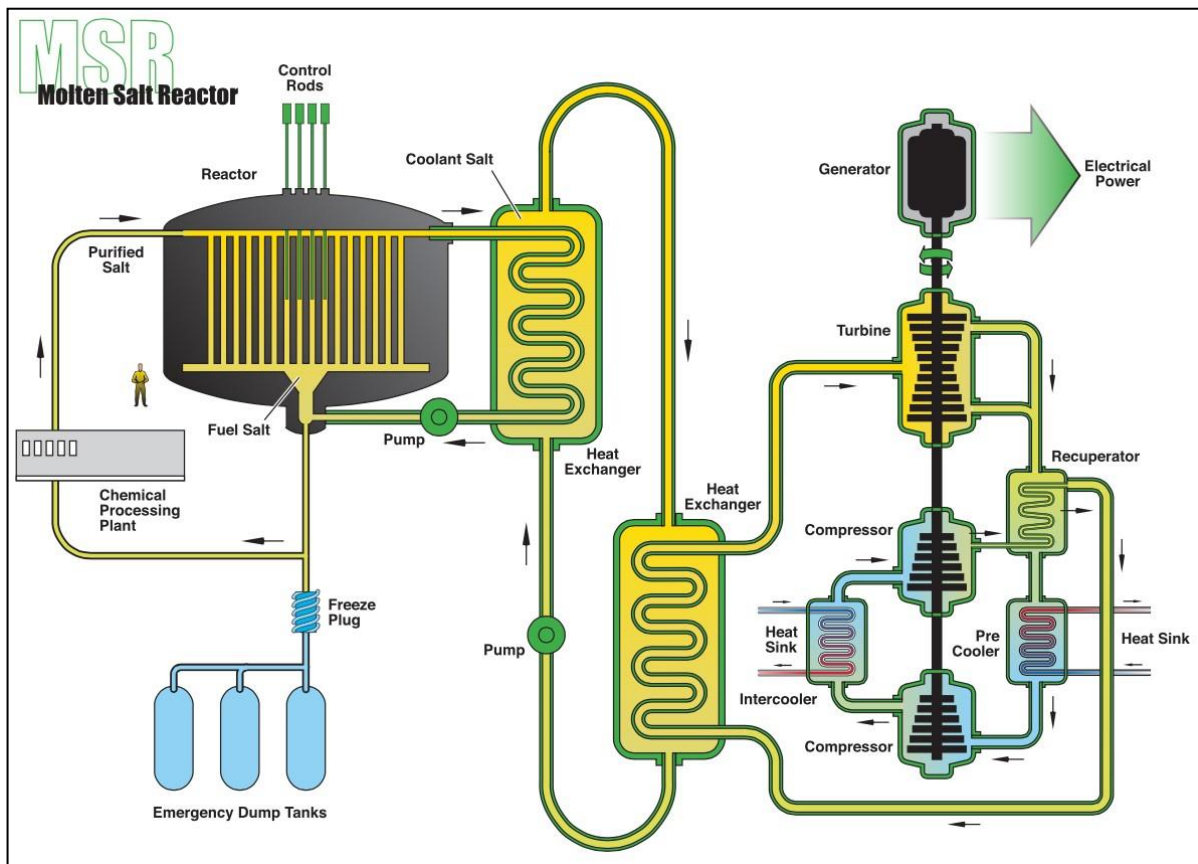

Figure 27: Molten Salt Reactor

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. The table below shows a quick comparison.

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

Table 3: trade-offs in reactor design

The PBR will be the safest choice because it is physically incapable of a meltdown. Meanwhile the MSR would be optimal if we seek to maximise resource durability, since it offers the possibility of using thorium (Th), which is found in greater abundance in nature than uranium. So, when designing reactors, we are actually designing for a variety of those values.

We will attempt to summarise this long but insightful case study. First, we saw that sustainability is best understood in terms of several moral values at hand - namely 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, and 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 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. So, these trade-offs need to be known and addressed for the responsible innovation of nuclear reactors.

7. Risk Management and Safety Engineering

7.1 Cost-Benefit Analysis

In the previous chapter, we learnt a bit more 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 where 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 a safe technology are clear; we effectively minimise 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 implement features that uphold other values as well. Similarly, any benefits from implementing safety features would
also 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, also accounting and quantifying the risks that come with each option.

A cost-benefit analysis 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 really 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 fixed or limited resources 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 different types of incidents/events

Decisions may be straightforward in some cases, but this may not always be true. 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, among others.
- Type II events on the other hand can be categorized as 'major accidents' - for instance 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.

For Type I and Type II events, the economic considerations can be somewhat different. Specifically for the latter, a disproportion factor can be used, as we will see.

Net Present Value

One important concept in CBA is that of the Net Present Value (NPV). A safety-related investment project represents an allocation of means and resources - such as money or time - in the present, such that it 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 costs of safety. With this information as a guide, a safety-related investment decision can be made in a more objective way.

The analysis' role is thus 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 upon 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 a discount rate in the calculations 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 the one you see below.

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

Where Xt 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 and not as certain 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 thus 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 thus 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? Well, we can say that the purpose of safety investments is to reduce present and future accidents. So the benefits are hypothetical, since the accidents - or rather, the accident scenarios - never actually occurred. They are defined then by the difference in consequences with and without a particular safety investment, and, if applicable, taking into account the difference in likelihood as well.

Since there are usually a large number of accident scenarios 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 thus 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, 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. Different approaches can be used to deal with this fact. For instance, cash flows can be expressed as expected values, taking the uncertainties in the form of probabilities into consideration; also 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, are related to extremely low frequencies and 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 above 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 values, 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.

7.2 Introduction to Risk Analysis

When considering large projects, risk and safety come very quickly to the forefront. Risk and safety should be words that are quite familiar by now, and especially for this context, we need to define them precisely and consider them closely to make them useful to us.

Risk, safety and security

Let us 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.

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 to 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, represented by s of scenario; b) how likely is it, or p of probability and; c) what are the consequences, represented by c. We will be referring to this triplet quite often.

There are, of course, different types of risk and so, 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.) Different consequences can be caused in many different ways. It is therefore important that when we talk about a consequence, that we also specify how this consequence could happen.

Safety is defined as a state - a 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, or flying airplanes. The latter is both a hazard for the people in the airplane, and for the people and property on the ground.

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 any plausible scenarios and a set of control measures. With security however, the focus is on intentional actions aimed at creating large consequences.

Quantifying and comparing risks

One of the problems with quantifying risks is that they do not have a common denominator. This is important, because otherwise we would be comparing apples with bananas. So, we should 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. Here we have some numbers but we are not sure what the numbers 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 $5\times10{-3}$. So of a thousand smokers, five of them would 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 to the table. Actually, winning a lottery is not very probable; in fact, it is actually more probable to die from a bee-sting, than winning the biggest prize in the state lottery.

The question is often asked why we use 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, it has proven to be a good proxy for total damage and several studies have shown this result. However, it is not a good proxy for disaster abatement where the numbers of those 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 happening, the probability? And finally, which 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 a 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 and that we go to several lengths to prevent. In a probabilistic approach on the other hand, we consider probabilities of particular accidents.

A risk analysis is part of what is called a risk-based decision analysis. It is shown in this diagram below.
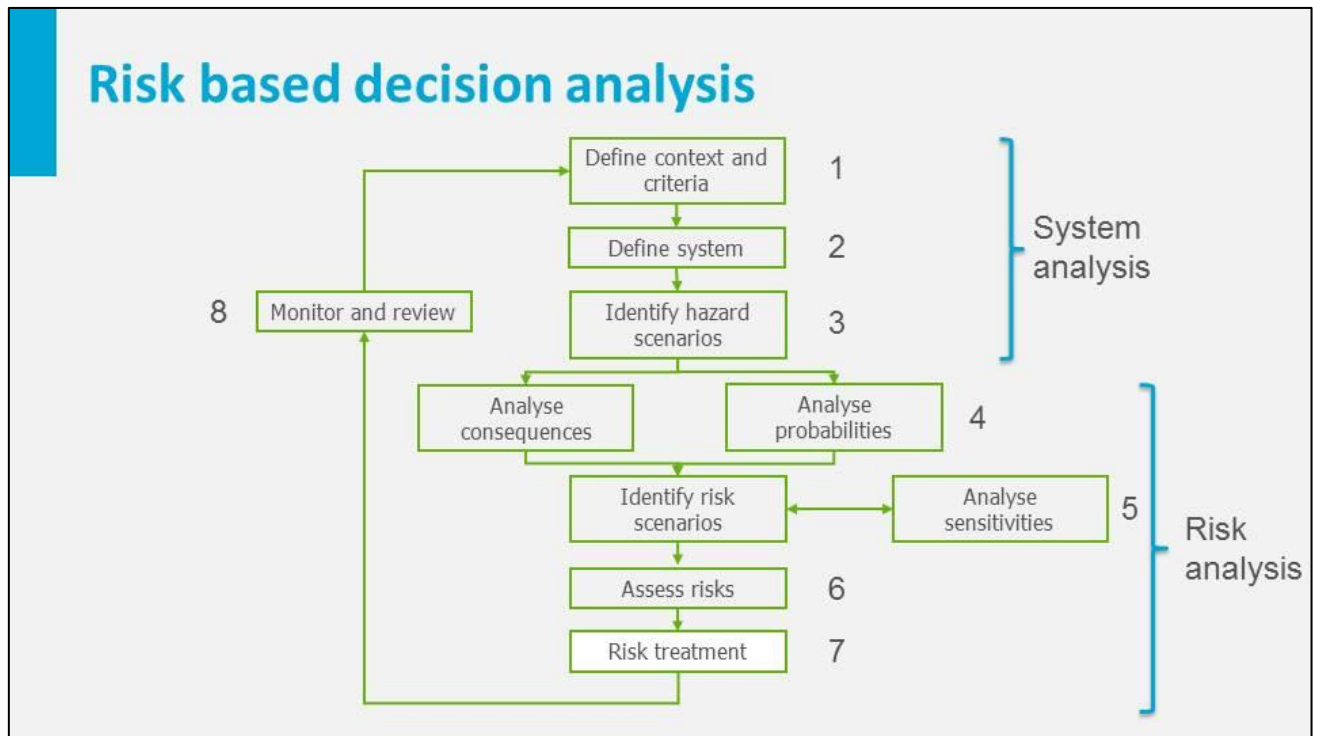

Figure 29: risk based decision analysis

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, and in the later steps, we actually carry out the risk analysis.

First, we define the context and the criteria for our 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 on a new technology, but we want to know where the risks are. And we also want to know who we need to involve, perhaps in order to control the risks, or to convince on the choice of 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. With internal safety, we basically mean occupational safety - the safety of the people at work in the plant or in the field. With external safety, we mean everything around the plant, the reactor, or the activity.

For now, we will only focus on external safety. This is an arbitrary decision, as occupational safety is equally important, looking at the many occupational accidents we have each year. However, these are more likely to specific to each industry, and as such, a comprehensive discussion of such methods would be beyond the scope of this book.

When we talk about external safety, there are different types of risks defined. We have individual risk, which was later redefined as localised risk. This is the probability that one person is being killed in a year at a particular place because of some hazardous activity. We also have group or societal risk, which is about a particular number of people being killed per year with a certain probability. Finally, we have the expectation value, which is the average number of people killed per year.

Risk contours

The probability that a/some person(s) in one particular year will be killed can be put on what we call risk contours. These contours, seen on a map, surround a potential hazardous place - say, a chemical plant or a nuclear power plant.

Figure 30: risk contours:
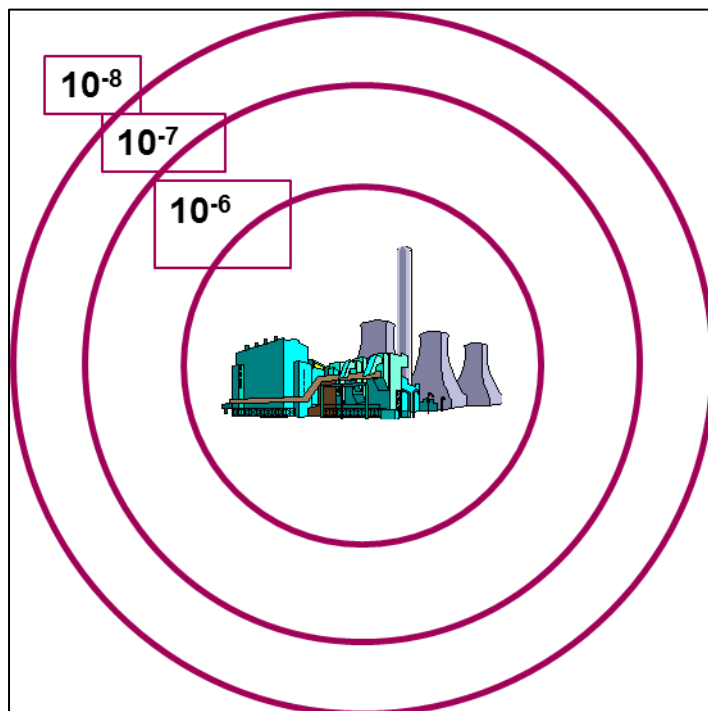
All points on the same contour line have the same likelihood. It is customary to draw these lines with (negative) exponentials of 10, like 5×10-5, 5×10-6, and so on.

Group risk on the other hand is usually represented by a graph, an FN-3curve, 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 amount of casualties.
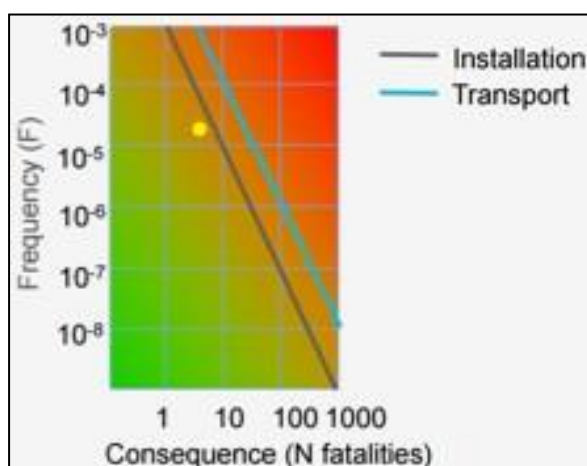
.

Figure 31: FN-curve

For this particular installation, for example, the frequency is about once every 80.000 years, and we can expect between 6 or 7 casualties. In the graph, two lines are also drawn, which represent an agreement between parties defining how many casualties are 'reasonably permitted' and how often these casualties could occur. We can see that the installation is below this norm.

An interesting phenomenon can occur, which also makes clear that there is a tension between the concepts of localised risk and group risk. Say the situation is like the one shown in the picture below.
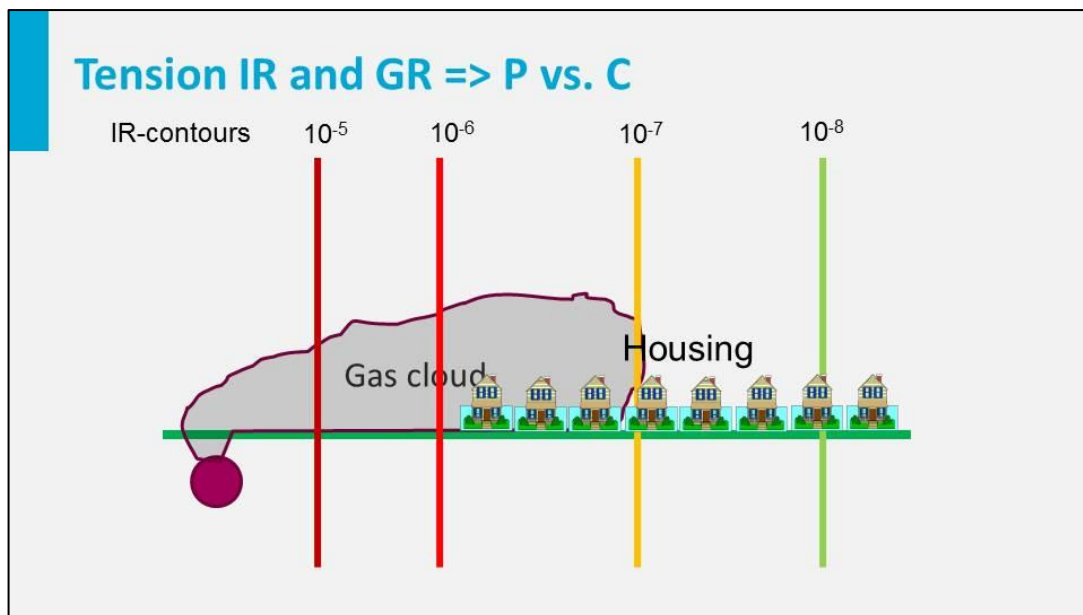
Figure 32: tension between localised risk and group risk

Houses are built beyond the risk contour of 5×10-6. So, once every 10 million years, a certain amount of people might die because of a gas cloud escaping this chemical plant. As you can see, the houses are built beyond the risk contour of one million years, so 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, localised risk and group risk can be at odds, and we have to agree at first on which one we will focus on or prioritise.

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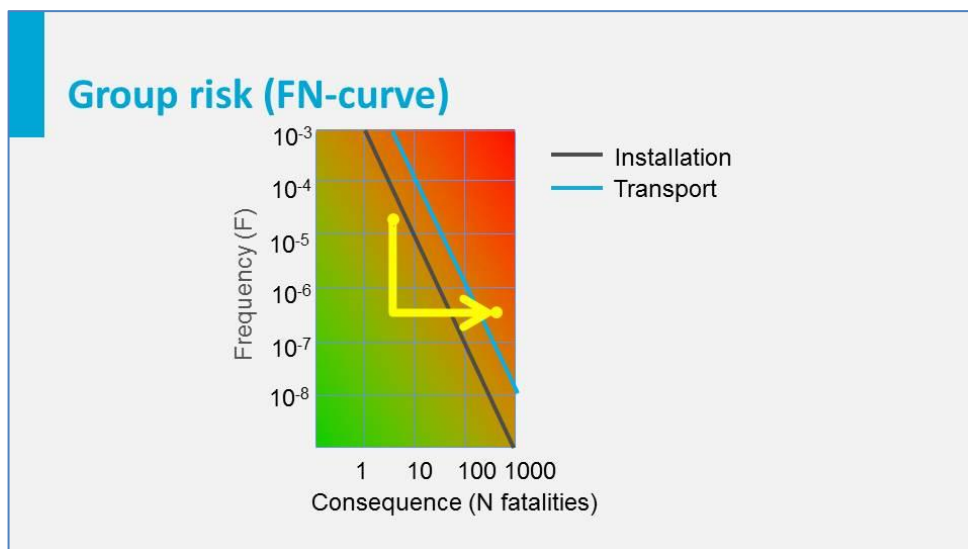

Figure 33: group risk (FN-curve)

It is important to note that risk contours are just theoretical concepts to guide our thinking, but real-world accidents

are not bound to abide by them. On October 4, 1992, a Boeing aircraft crashed into an apartment building in the Bijlmer in Amsterdam.
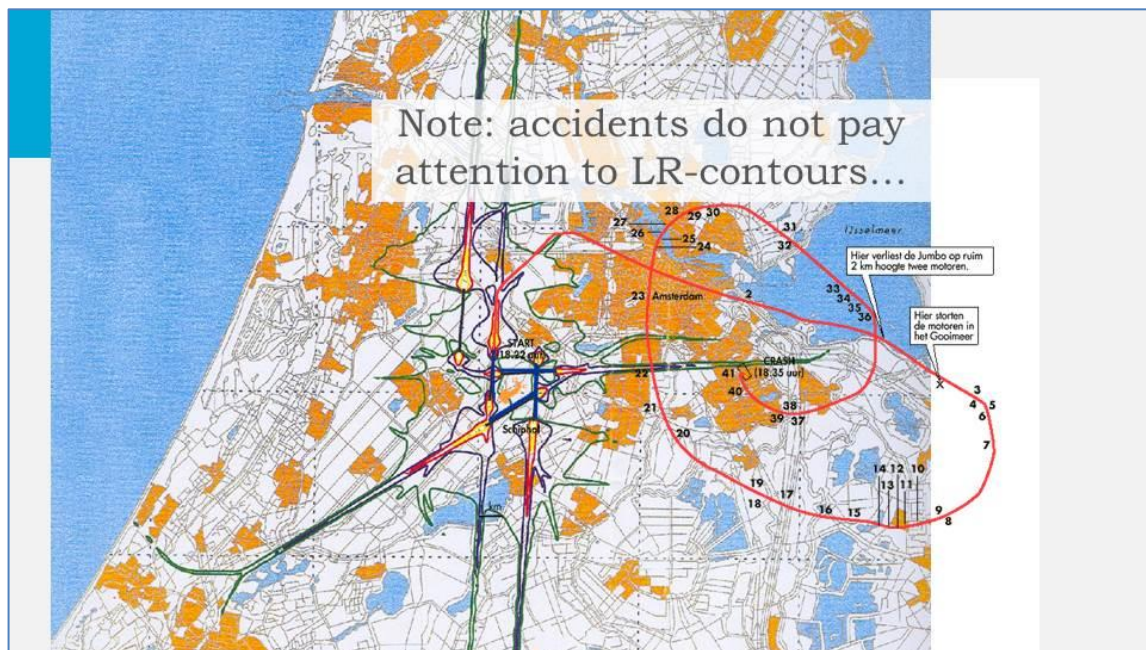

Figure 34: trajectory of plane Bijlmer disaster

You can see its trajectory in red in the image above. The green contours are the low, localised risk contours - the ones where we don't expect a fatality very often. As you can see, the Boeing crashed exactly in such an area!

Defining the system and boundaries

After we have defined our context and criteria, we have to define what the system is and whose 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 worksite 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 pre-empt, 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 due to these being beyond our sphere of control, or the fact that we have only limited resources at our disposal. Also, 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 would it happen; and what measures/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.

There are several different methods for the identification of hazards, listed below.

- 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/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 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.
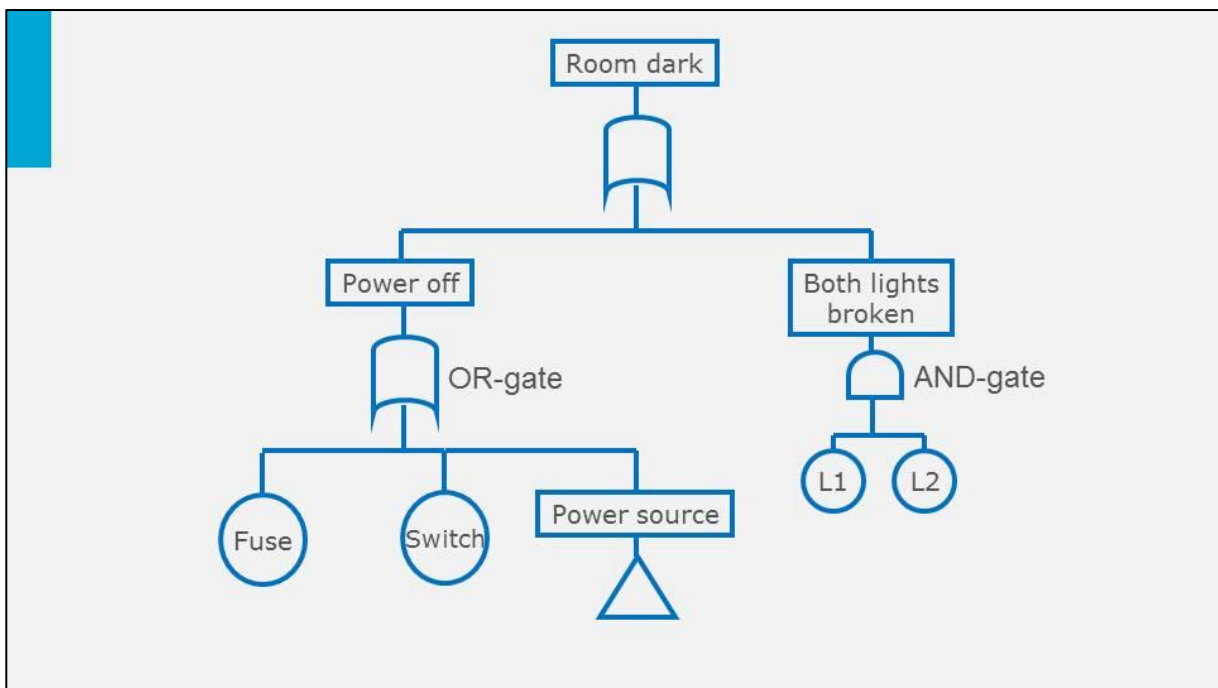

Figure 35: Fault Tree Analysis

Let us take as an example a room that has two light fixtures in it. We enter the room but the lights don't go 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 the 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 ORgate. 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, which has a straight bottom. An AND-gate would mean that both sub-events have to occur for the top event to trigger. However, in our case, we simply have an OR-gate.

Let's now take a closer look at the system itself, as we 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 – (that may

all be faulty) or 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 neighbourhood. We could also include many other ways in which the power supply can falter. Essentially, we have to choose system boundaries depending on what we can effectively prevent or control, and only adapt to or manage any events that fall outside the system boundaries. These decisions are essential for our analysis.

Finally, we can add probabilities in 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 where there are two light fixtures. The Fault Tree also explicitly visualises the different scenarios, which can be understood as individual paths through the tree.

Bow-Tie Diagram

Interestingly, we can combine two Fault Trees at their top events, and put them on their side, as in the image below.
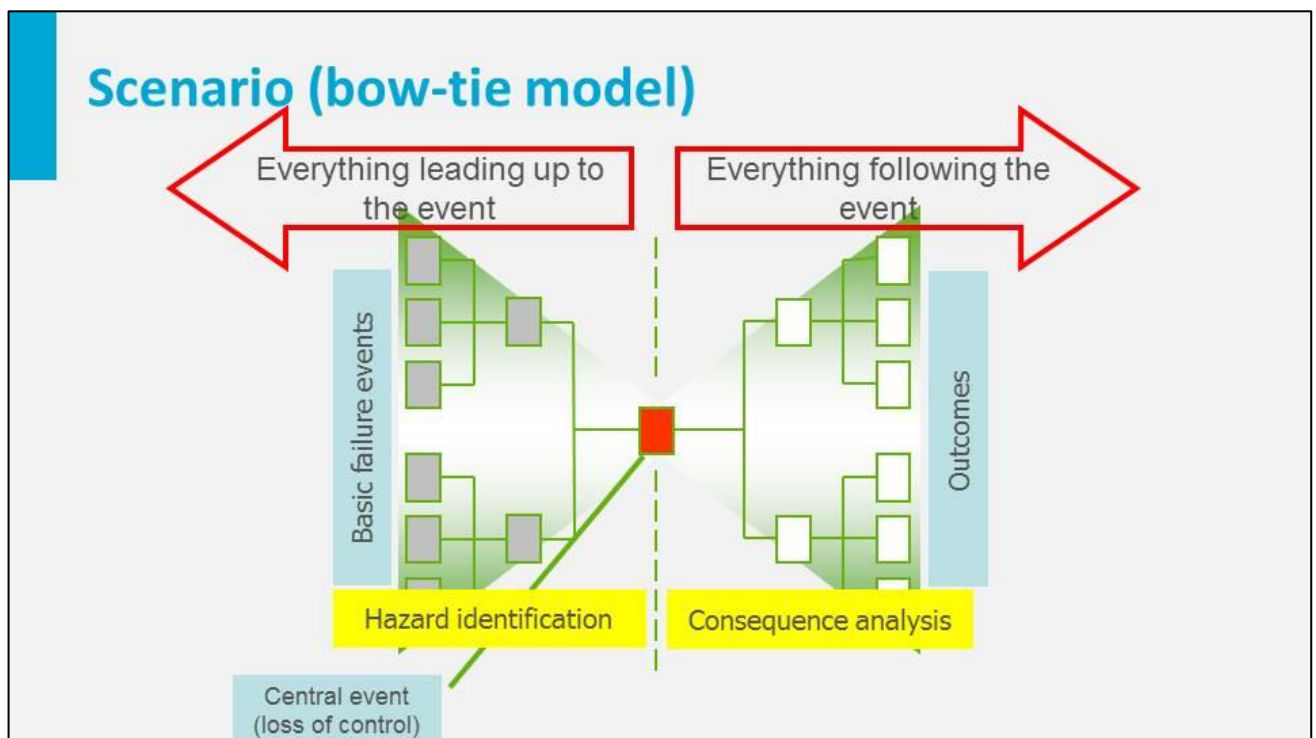
Figure 36: Bow tie diagram

What we then have is a diagram that vaguely resembles a bowtie, and we do call this a bow-tie. Notably, a bow-tie only contains OR-gates, and it is mostly used in a qualitative way with no probabilities.

The central event is often defined as the point of loss of control. On the left-hand side, events are ordered such that they represent the failures leading up to the central event, and on right-hand side, events are ordered depicting the outcomes from 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. 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 take account of that? To answer these questions, one needs to have expertise pertaining to the domain under study.

Different loss categories

	Minor	Critical	Severe	Catastrophic
Plant damage and lost production	Short-term loss of production	Damage to machines. Reparable in short term	Damage to plant. Major repair costs. Serious loss of production	Substantial damage to plant. Potential loss of overall plant
Environment damage	Temporary excursion in emission levels	Significant release. Effluent clean up required	Ecological damage for up to 1 year. Risk of penalties	Ecological damage for more than 1 year. Pressure to cease business
Harm to personnel	Reportable but non-disabling injuries causing over 3 days absence	Disabling injury or severe injury requiring extensive recovery. 1 to 10 chance of fatality	Critical injuries, and possible 1 fatality	One or more fatalities

Figure 37: different loss categories

If we want to make life a bit easier, we can classify consequences into a limited number of categories. This is what companies often do. In this particular example above, four loss categories are defined - namely 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. There are many such matrices that can be used to frame discussions about losses and target groups. Note also there may be many scenarios associated with these consequences.

After we have listed the possible consequences - an exhaustive effort within the scope of the system 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 to use 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 many different questions 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 - that is to say, prevent them altogether - and which ones to approach in a probabilistic manner; preventing them to within a reasonable probability.

We can consider doing a sensitivity analysis as well, to see if our analysis makes sense and also 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 then?

Risk assessment

The sixth step in performing risk analysis is the risk assessment. Again, we can make life 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.

Risk assessment 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

Figure 38: risk assessment matrix

In this particular matrix as seen above, 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. We also have five classes of severity. The colours in the matrix often indicate what will be done about the risk. Some risks in this matrix are accepted, whereas others are prevented to a certain cost, or insured.

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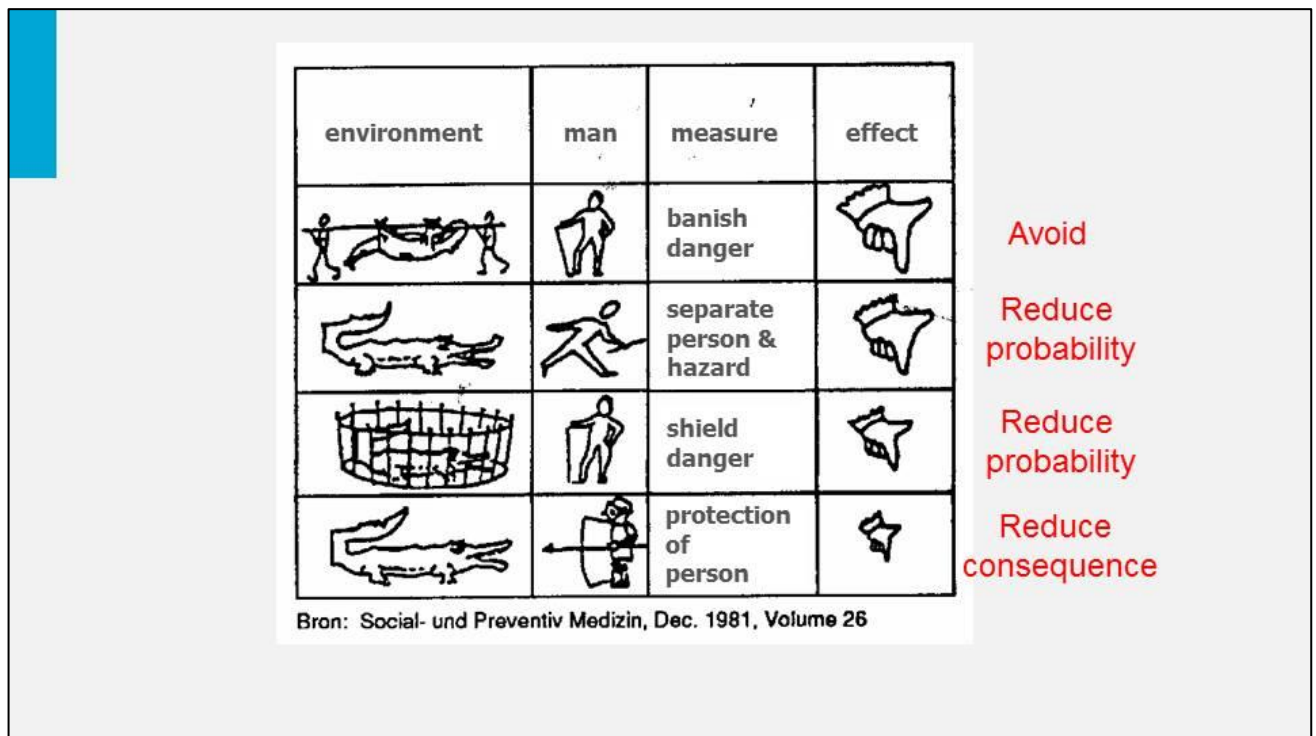

Figure 39: treatment of identified risks (source: Sozial- und Preventiv Medicin, Dec 1981, Volume 26)

We can avoid the risk completely and carry the crocodile away. We can keep a safe distance from the crocodile, we can put the crocodile in a cage, or finally we can put on protective clothing.
It is relevant here to speak of the Haddon Matrix. William Haddon was a medical doctor who tried to think of all possible strategies and ranked them in order of effectiveness, eventually visualising it in the form of a matrix. We find that elimination of hazards is the most effective, followed by minimising exposure to hazard sources. We can also try to prevent the release of hazards, or at least modify the release in order to minimise damage. Haddon thus 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 discussed earlier. We have the time before control is lost to pursue prevention strategies, and after control is lost, we shift to mitigation strategies in order to cope with the consequences. So we can try to 1) avoid negative outcomes, 2) reduce the likelihood, 3) minimise 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 some hazard that presents a possible threat to some target, but which is protected by one or more barriers.
Barriers can be of different kinds; they can be physical, or procedural, or a combination of both. These barriers prevent the unwanted energy flow from reaching the target. In a bowtie, 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. So the open pathways need to be monitored closely.

Risk analysis in practice

So 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/mitigation.

We also need to be 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.

7.3 Case Study: Self-Driving Vehicles

Recently, newspaper articles about Google developing technologies for self-driving vehicles were followed by heated societal 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. This section aims to ask some critical questions about the ethical issues surrounding AVs.

Ethical concerns behind AVs

To begin, we can ask: what are the important ethical concerns about AVs?

- *The problem of "many hands"*

The first area of concern is a problem we have already seen 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 some concerns regarding the accountability of any one actor.

Say an AV causes an accident due to a failed sensor, partly due to bad weather. Who is responsible? Is it the driver? Or the car manufacturer? Or the company providing sensors? How about the company providing the key software? Perhaps the dealer doing vehicle maintenance? Or maybe, the road authority which
allowed these vehicles to be on 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 responsibility, because the driver feels there was no wrongdoing on his or her part.

- *The "trolley problem"*

We can also ask how even the most complex algorithmic intelligence will deal with the "trolley problem". What should the AV do if there is an incoming 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 for alternative (or preferable) accidents 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 how pedestrians would change their behaviours in the (ubiquitous) presence of AVs.

- *Distribution of utility*

Thirdly, there are the potential trade-offs between travel times, safety and sustainability. Any optimization of the system from these one of the three perspectives may not result in equal outcomes. Taking a safety perspective for instance, it is preferable to have longer distances between vehicles, but this arrangement would induce higher fuel consumption due to higher air resistance. Also, the utilised 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 would be 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 given the 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 AVs might be at least €10,000 higher than those of comparable normal cars. We could argue that this is not a problem.

However if the road authority were to allocate dedicated road space - say one lane of the highway - specifically for AVs. This would effectively decrease the available road capacity for normal vehicles, possibly leading to more congestion. Moreover, we could also argue that since only affluent consumers would purchase the expensive AVs initially, such a highway 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 in this sense, serving both as personalised on-demand transportation and freeing the driver from that task.

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 on public transportation and cannot afford AVs in the first place.

- *Transition issues*

It is also important to pay attention to the transition period. Experts think that immediately after the initial market introduction of AVs, the capacity of roads might decrease, not to mention the safety. This could be expected due to the ''growing pains'' so to speak, failures in technology, or a period of real-world learning. Eventually though, AVs would yield better efficiency, address any highway capacity concerns and improve safety. In other words, the initial years of AVs could decrease 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 is said to 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.

Ethical benefits from AVs

Based on the six ethical issues presented above, we could ask: are AVs truly desirable from an ethical perspective? (Not that the issues discussed so far are exhaustive as regards AVs; there could be other ethical issues such as data privacy, security issues, among many others.) It might be premature to decide against AVs. We need to realize there could also be ethical benefits.

- *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 or even social privileges such as maintaining contact with distant family and friends. Even public transport may be too expensive in some cases. In cases like these, AVs - perhaps through schemes like car-sharing - could provide essential paratransit alternatives, reducing levels of 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 or other drivers. Another overlooked area where we would see immediate and significant safety benefits would be from 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 for their fuel-efficiency - after all, what sense is there in having a powerful but inefficient car?

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 currently have AVs only, and no more 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, higher risks, 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.

Part IV

8. Value Sensitive Design

8.1 Introduction to Value Sensitive Design

We are now at the last leg of our introduction to Responsible Innovation. Let us now discuss the relevance of Value Sensitive Design (VSD) to 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.

Cultural developments of IT in society

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 given to 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 - namely 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 define 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 on-going 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 - as 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 initially 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", and 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.

The origins of Value Sensitive Design

There were number of converging lines of thought and research with respect to VSD.

- *Do artefacts 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?", and it drew attention to the fact that values can be manifested in real-world objects and technologies, profoundly shaping behaviour of even large 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 construction of overpasses on New York parkways to be intentionally low, such that it was accessible to 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 neighbourhoods who were more likely
to take the buses could not pass under. 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 baked 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). These 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 artifacts, systems and infrastructures. Researchers like Geoff Bowker, Susan Leigh Starr and Lucy Suchman have contributed much to this body of work.

- *Designing for Values*

Some specialized areas of design and engineering also started to use this basic concept of Design for Values or VSD concretely 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 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, lines of sight among others, were all found to be influential in encouraging or discouraging crime rates. As such, these factors were carefully embedded in the design parameters of buildings and neighbourhoods.

Defining the method of Value Sensitive Design

The most clear and precise formulation of VSD concept originated in a movement at Stanford between the 1970s-80s in the field of Computer Science, and 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. It was developed by Batya Friedman et al., building 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 design of (information) technology.

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 a part and 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 (See Friedman, Kahn, and Borning, 2005 on VSD; or Flanagan, Howe and Nissenbaum, 2005 on VAP). Each of the conceptual, empirical and technical investigations and analyses are carried out iteratively, mutually informing and being informed by the other investigations.

Value Sensitive Design has a number of features that are aligned with Responsible Innovation. Values and moral concerns of all stakeholders need to be articulated at a 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 artifacts need to be evaluated in terms of the values upheld and 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 innovation and design of new technologies, as well as the diffusion and deployment of technological artifacts.

8.2 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 though, we live in a technological world in which technologies shape how we live as well. We have only to think of how different present-day lifestyles are with ubiquitous technologies like the Internet, computers and smartphones compared to the lifestyles of generations that came before us. 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. So, 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: namely, 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 only 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 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 hand than without.

- *Substantivism*

Substantivism takes the position that technology itself is value-laden and that humans have no influence on that. For example, it has been argued that technology embodies values like efficiency, or that technology inherently leads to environmental degradation or to a lack of authenticity or even drives human interactions to a minimum.

A problem with this position that it overlooks the influence that people could 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. Going forward, we will focus especially on the design aspect.

What values should be included in technology design?

A first thing to note is that a whole 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), designers (and their professional communities), users and stakeholders, laws and government policies, technical codes and standards, and codes of ethics and other moral concerns. Listing all these values however, would not tell us which values to include because that 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 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 most important, but they may still agree on the broad spectrum of values which are 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.

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 human well-being, justice, beauty, honesty and truth.

How can we translate moral values into design specifications?

Now we can ask, how can we 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. This following image is an example of a values hierarchy.
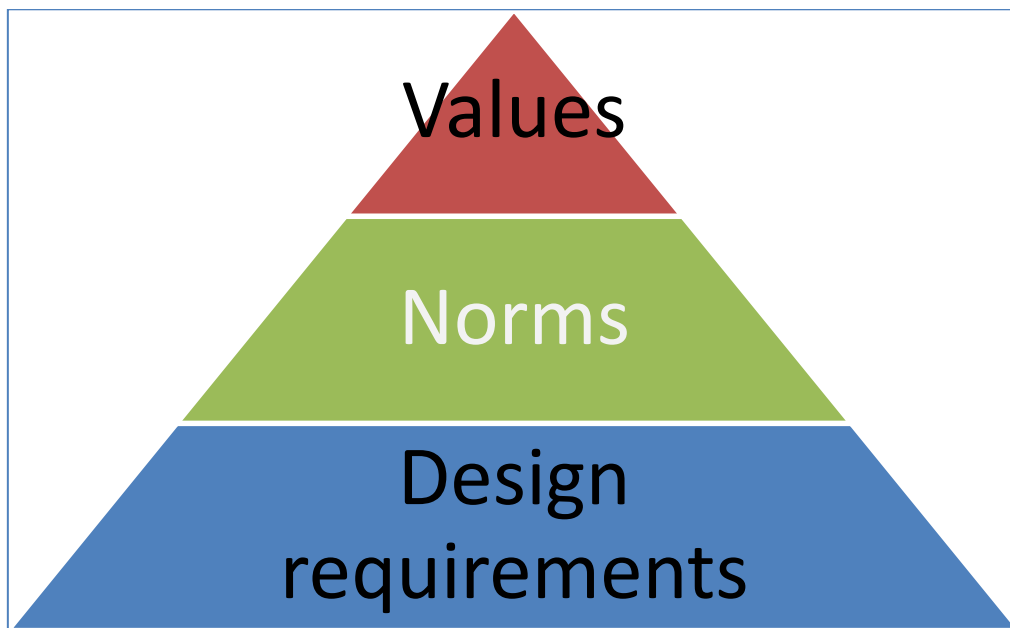

Figure 40: values hierarchy

This table is based on a European directive for the design and production of battery cages for laying hens. The directive was meant to guarantee the value of animal welfare in the design of battery cages. We can see how this value is translated in 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 cm2 of floor area per hen.

In this case, the values hierarchy has been reconstructed on basis of a European law, but we can also make a values hierarchy ourselves. Here is another example; an attempt to make a values hierarchy for biofuels.
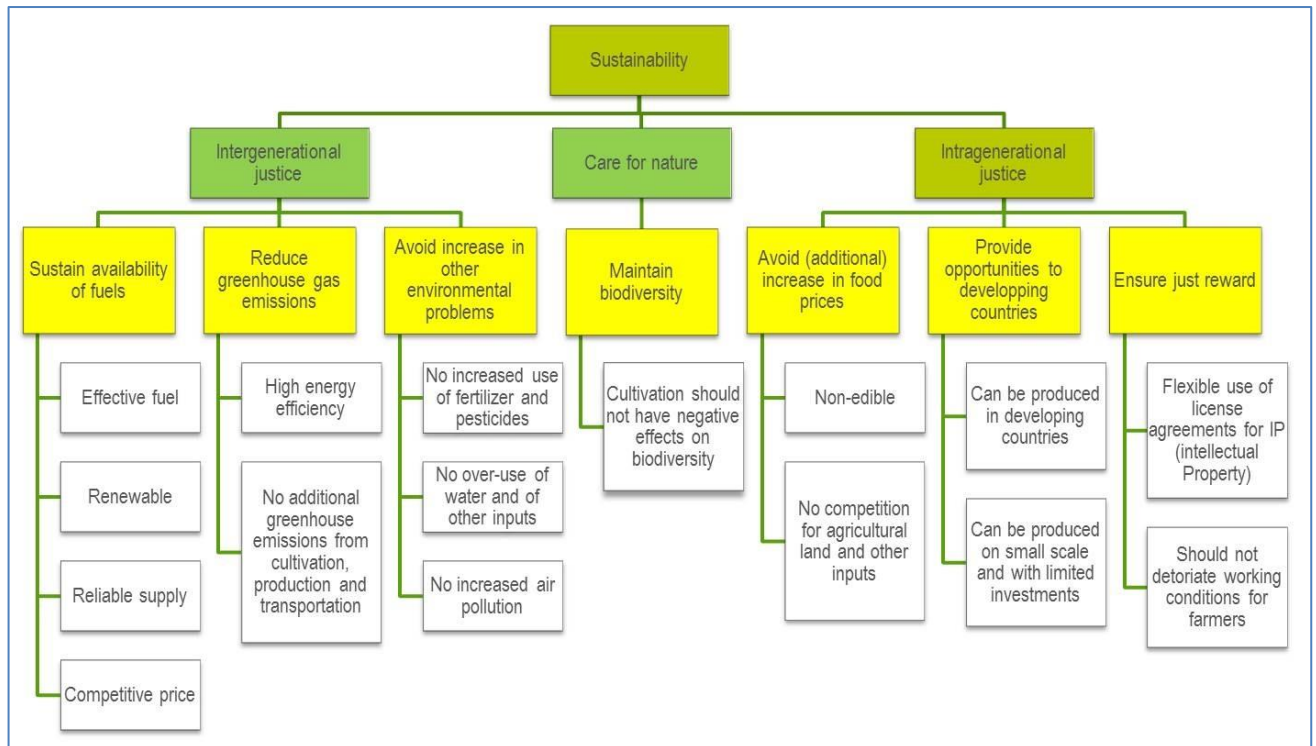
Figure 41: biofuels and value systems design

Biofuels are based on relatively recent lifeless or living biological material. They have been introduced in order to deal with 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 above is an attempt to organize all such concerns into a values hierarchy.

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

With each of these values, a number of norms is associated. 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, 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, these 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 then is: what ultimate goal do these requirements achieve?

An important question is whether a specification of a value in a values hierarchy is adequate. We can question here if meeting lower level design specifications counts towards meeting higher level norm or value, thus going bottom-up?

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. Indeed, the European Union has since changed its laws and formulated more strict design requirements that effectively forbid the battery cage.

So, this is how the values hierarchy helps us to structure and translate abstract moral values into tangible design requirements.

8.3 Case Study: Autonomous Weapons

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

Now this is obviously not entirely helpful because we already have automated weapons. Automated weapons of course could also do those sorts of things, without a human operator. So if we look at missile defense systems for example, they have been automated because speed is a crucial issue in intercepting a hostile missile. And of course, they can also intercept such a missile once they have been programmed without the operator really having to do anything..

So there is really a big question it seems, in the debate and in the academic literature on new weapons systems whether autonomy is in fact something that should be seen as separate from automation. Now, some people say yes to this question.

So 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 this, that it makes a decision by itself, it is actually not clear what decision-making means in this particular circumstance. And that 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. That we could deploy, we could put them into a particular situation, and then they could apply the sort of criteria that regulate the use of real armed forces themselves. 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. So the kind of machines we have in mind cannot make decisions in the sense that they apply intelligent criteria themselves. But they still differ from automated systems. A cruise missile for example, is of course an automated system. We can program it with a certain GPS coordinate, and then it will find and hit the target by itself.

Now with an autonomous system, we would be looking at something more complex. So we would for example be looking at a system that could be deployed in a very complex and challenging environment, in which it could then navigate its own way to a target without you having to be given a particular GPS coordinate.

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

So consider a stealth drone. Drones are currently remote-controlled by an operator. And for that to be possible, there needs to be a link between the operator and the drone so the drone can be controlled. Of course the problem is, if the drone flies into enemy territory, the link between the operator and the drone could potentially be tracked by the enemy. So, 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 it could attack those targets without the operator necessarily having to do anything, apart from initial pre-programming it. Speaking of stealth mission scenarios, that could be very useful to the military.

So these would be very narrowly (and) best described as military advantages. On the other hand, we could say, and some do 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. Now this is a very big if. If that is true, then of course these systems do seem desirable. Who would not want fewer war crimes? Still, it remains a very big if.

So it seems that the burden of proof really falls upon those people who think these sorts of systems can really make that sort of difference on the battlefield. In terms of the risks, the risks are being increasingly highlighted in the literature. There is a very big risk for example, that these machines might not be able to adequately identify their particular targets. That is a very big risk. We are talking about machines that could operate in very complex battlefields and 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.

There would be other technological risks as well. For example, if we consider the possibility of our system being hacked by the enemy, being re-programmed by the enemy and then re-deployed to commit war crimes, or to attack our own troops. That would obviously be a big risk as well.

So there are some advantages, some military advantages. There might be some ethical advantages but there are also significant risks resulting from those sorts of weapons.

So what makes people uncomfortable about the use of robotic weapons? It seems that a lot of the campaigning surrounding robotic weapons does indeed stress 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, these robotic weapons or autonomous weapons are not entirely new. They are precedents of weapons which are very widely accepted, automation within missile defense for example. That's very widely accepted, and people are not really worried about this at all.

So there is the question: does this has something to do with the perception of particular technologies, which seem to raise these big issues? Or does it seem to be something entirely new but which is actually building on what's already there. And so perception is a big issue.

But aside from the perception point, we should also take the worries that people have about these weapons seriously. One worry could be concerning reliability for example, and how such weapons could be deployed in a safe manner. Could these weapons really find targets in a very complex battlefield?

In a way, that 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? People are very worried about this.

There are obviously guidelines when it comes to the development of weapons and new weapon systems. And that applies to autonomous weapons as it would to any other weapon of course. There are also of course guidelines for their deployment, and that needs to be kept in mind. So we are not really operating in a vacuum - a legal vacuum or a moral vacuum - when we consider the development of these kinds of weapon systems.

Now the question always is, would people comply with these sorts of guidelines? And there is a worry that these guidelines might be undercut by some states that are very keen on developing these kinds of technologies. We're already seeing the US's use of drones during counterterrorism operations for example, which are legally ambiguous. It is not really clear how the law applies to those sorts of operations and those sorts of situations. So there is a very real worry here about compliance.

We need then a strong response from civilian society. We also need a strong response from international institutions like the UN, Red Cross and so on, in order to make absolutely clear to all parties that this is a process (speaking of the development and deployment of new weapon systems) that needs to take place in accordance with international law. So we hope that individual militaries do 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.

So a key theme here 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 in particular that these systems are being used in accordance with the law.

Summary

You have come to the end of the MOOC-to-book on Responsible Innovation. We hope you have enjoyed the course, and have gained insights into the ethics behind the technologies we build and use on a daily basis. For your benefit, we have compiled a small 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/confusion of values and responsibilities (Trolley problem, "Many Hands", etc.). We saw that when there are multiple values to uphold, each of them important and desirable in their own way, there can be a sense of moral overload 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 the 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 also be optimistic about the use of innovation to satisfy multiple (conflicting/constrained) values; after all, isn't that what innovation is about?

In Chapter 3, we learned about the institutional context of modern innovation. We discussed how institutions - that is, embedded /explicit social conventions and rules that structure social interactions between individuals and groups - can profoundly preserve and influence favourable values and how they are manifested.

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 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 also not automatically "responsible", and the issue of social standards must be justified before this question may be answered.

In Chapters 6 and 7, 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 learnt 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 very hard 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, so that we can develop new technologies with

pre-emptive safeguards in order to mitigate as much as possible known risks.

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 price the priceless?

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

And finally in Chapter 8, we introduced Value Sensitive Design (VSD) as a framework for operationalising 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 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?

We hope you have enjoyed the content and the discussions as much we enjoyed creating this material.

Thank you.

Question to ponder

Below some question for pondering.

Questions to ponder 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 such that the dilemma effectively disappears?
- Do you think it's useful nonetheless to try and resolve the various thought experiments mentioned above - be it the "Trolley Problem" or the "Fat Man" problem? Why?

Questions to ponder par. 2.3

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

Questions to ponder par. 2.5

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

Questions to ponder 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 to ponder par. 4.3

• Do you know examples of innovations that never appeared in the market, or appeared but did not succeed?

Questions to ponder par. 6.3

- 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 to ponder par. 7.1

- Despite the systematic effort undertaken during a CBA to capture every advantage/disadvantage of a given problem, there are still some ethical concerns about the practice. We can ask how can you price the priceless. 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 to ponder par. 7.3

As we read earlier, in complex socio-technical systems, we inevitably come across the problem of "many hands" - a multitude of actors and stakeholders with varying levels of influence. Not to mention uncertainties, innovations and interdependencies. On top of this, organisations need to be nimble and adapt to competitors, changes in the market, changes in regulation 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?

About the teachers

A large number of teachers provided input for the MOOC on RI. This is an overview (all from TU Delft, unless otherwise stated).

- *Professor dr. Jeroen van den Hoven.*
  Course director is Professor dr. Jeroen van den Hoven. Van den Hoven is Professor of Moral Philosophy at Delft University of Technology, and he is Head of the Department of Values, Technology, and Innovation of Delft University of Technology. His research focuses on responsible innovation. In the past five years, he has worked on three more specific themes in this field. A first theme is the ethical issues raised by Information and Communication Technology (ICT), in particular privacy. He has also done work on the privacy issues raised by nanotechnology, in particular RFID chips. He has proposed and defended a conception of privacy which constitutes privacy in terms of moral reasons for data protection. A second theme he has worked on is Value Sensitive Design (VSD). He has made various conceptual and methodological contributions to VSD and has argued that VSD, and more broadly technical innovation, is a way to avoid moral dilemmas. A third theme Van den Hoven has been working on is the responsibility of system designers. He has argued that the designers of technical systems create the epistemic environment in which the users and operators of these systems can or cannot take their moral responsibility. The system may, for example, provide users and operators with certain crucial information or it may withhold that information. More specifically, he has argued that system designers have a meta-task responsibility to see to it that their systems are designed in such a way that users and operators can take their responsibility.

- *Prof. dr. Ir. Ibo van de Poel*
  Ibo van de Poel is Anthoni van Leeuwenhoek Professor in Ethics and Technology at Delft University of Technology in the Netherlands. Since 1997, he is lecturing in ethics and technology for several engineering course programs at Delft University of Technology. He has published a widely used international textbook on the topic (Ethics, Engineering and Technology, Wiley-Blackwell, 2011, with Lamber Royakkers), and is one of the developers of the online ethics teaching tool Agora (www.ethiekentechniek.nl). He does research on, and has published on responsible innovation, engineering ethics, the moral acceptability of technological risks, values and engineering design, moral responsibility in research networks, and ethics of new emerging technologies like nanotechnology. His current research focuses on new technologies as social experiments and conditions for morally responsible experimentation.

- *Dr. Philip Robichaud*
  Phil did his PhD at the TU Delft and worked as a philosopher on the ' awesome Enhancing Responsibility' project. The project explores the impact that cognitive enhancement technologies have on issues of moral and legal responsibility. He also writes about the epistemic condition of moral responsibility and neuroethics. He teaches in various areas of applied ethics and ethical theory.

- *Dr. Neelke Doorn*
  Neelke Doorn is assistant professor Ethics and Water Governance at the Department of Philosophy of Delft University of Technology and former assistant director of the 4TU.Centre of Ethics and Technology. In 2013, she received the prestigious NWO Veni grant for her project The ethics of flood risk management. Together with Diane P. Michelfelder, Neelke is Editor-in-Chief of the peer reviewed journal Techne: Research in Philosophy and Technology. In 2013, the edited volume Early Engagement and New Technologies: Opening Up the Laboratory appeared, of which Neelke is the leading editor.

- *Prof. dr. Sabine Roeser*
  Sabine Roeser is Professor of Ethics at TU Delft (Antoni van Leeuwenhoek Professor) and at Twente University (Socrates Chair, part-time). She is head of a research group on Moral Emotions and Risk Politics, funded with prestigious VIDI- and Aspasia-grants from NWO (Netherlands Organization of Scientific Research) She previously held a highly competitive VENI-grant from NWO (2005-2009) and a NIAS fellowship from the Royal Dutch Academy of the Sciences (KNAW; 2009-2010). Roeser has published numerous articles and books on risk, ethics and emotion. She is regularly interviewed for popular media in the Netherlands. She is a member of various advisory boards for the Dutch government.

- ***Prof. dr. Rolf Künneke***
  Rolf Künneke is professor in the economics of infrastructures. He holds a masters degree in economics, from the University of Dortmund, and received his doctoral degree from Twente University. His thesis was on the effects of privatization of Dutch energy companies. Since then, he has made a long record of researches on the restructuring of infrastructure industries, with a special focus on the energy sector. His recent research includes themes such as the coherence between institutions and technology in infrastructures, innovations in energy networks, institutional reform, regulation and privatization, ownership unbundling, national reforms in European gas, innovations in gas networks, governance of complex socio-technical and socio-ecological systems.

- ***Dr. Eefje Cuppen***
  A fascination of how science and technology get their meaning in society - and the implications for policymaking. That is the main driver behind Eefje Cuppen's research. Her interests are in the field of sustainability, energy, innovation and climate change, with a particular focus on participatory policymaking, social conflicts, transdisciplinary research, and (quantitative and qualitative) methods for stakeholder analysis and dialogue. She has worked on topics such as shale gas, GM food, bio-energy, water management, and hydrogen. As assistant professor at TU Delft, she is currently involved in a research project on robust long-term sustainability strategies for the Dutch gas sector, funded by the Energy Delta Gas Research Programme (EDGaR). She coordinates a subproject on societal support for energy technologies, such as shale gas, CO2 storage and local energy initiatives. Her teaching concentrates on the interaction between (sustainable) technology and society.

- ***Prof. dr. Cees van Beers***
  Cees van Beers is professor of Innovation Management and chairman of the capacity group Economics of Technology. He holds a doctorate in economics (Ph.D) from the Free University Amsterdam. He also worked at the University of Leiden, the Institute for Research on Public Expenditure in the Hague and as associate professor on innovation economics. He was visiting fellow at the department of economics, University of Adelaide, Australia studying environmental and labour standards. Research focus is on three themes: 1) inclusive and frugal innovations, 2) government policy failures and sustainability, 3) the impact of Information Technology investments and -adoption on business performance. His teaching experience is in Technology and Strategy and Innovation Economics.

  Dr. Roland J. Ortt
  Roland J. Ortt is an Associate Professor of technology and innovation management at the Delft University of Technology, the Netherlands. Before joining the faculty of Technology Policy and Management Roland Ortt worked as an R&D manager for a telecommunication company. He is the author of various articles in journals like the Journal of Product Innovation Management, the Market Research Society and the International Journal of Technology Management and has won several best-paper awards. His research focuses on development and diffusion of high-tech products, and on niche-strategies to commercialize these products.

- ***Dr. André Leliveld (Africa Study Centre, Leiden University)***
  André Leliveld is a development economist and does research on informal insurance and social-support systems and community-based (health) insurance schemes in Africa, particularly in Togo, Uganda and Swaziland. His current research project in this field is a study of the effects of community-based health insurance in Togo on the economic well-being and health status of participants. More recently, his research has also included structural transformation processes in African economies and the effects on (youth) employment and labour protection.

- ***Prof.dr. Peter Knorringa (Institute of Social Studies, Erasmus University Rotterdam)***
  Peter Knorringa holds the chair on Private Sector & Development focuses on the diverse roles and impacts of business on development.
  His past research has focused on small and medium scale enterprises in developing countries, operating in clusters and value chains. His present research agenda also includes the behaviour of large multinational companies, especially from China, India and Brazil. His aim is to investigate to what extent socially responsible and environmentally sustainable forms of development are possible under capitalism, and where and when these might be built upon a strengthened morality in the private sector itself. Core activities include: Leader, Civic Innovation Research Initiative, one of the four ISS research

programs; Principal researcher and coordinator, Leiden-Delft- Erasmus-theme on 'Innovation and Technology Networks with Africa: Re-engineering the Bottom of the Pyramid' and Co-investigator and coordinator Consumer-Civil society work package, ESRC-funded research program on: 'Rising Powers, Labour Standards and the Governance of Global Production Networks'

- ### *Dr. Rafaella Hillberbrand*
  Rafaella Hillberbrand worked as an associate professor working at TU Delft on a prestigious TUD female fellowship. Before joining TU Delft, Rafaela Hillerbrand has worked as a senior research fellow at the University of Oxford on epistemological and ethical problems for decisions under risk and uncertainty. Her research focuses on philosophy of science and technology as regards theoretical philosophy and on problems in applied ethics of technology and environment as regards practical philosophy. In 2011, Hillerbrand was an advisor of the ethics commission for safe energy supply under Klaus Töpfer´s guidance. Subsequent to the accident in Fukushima she gave advice and support to political decision makers in Austria and Switzerland on the change of energy supply. Hillerbrand holds a PhD in theoretical physics as well as a PhD in philosophy. Her book on ethics of technology, which covers aspects of applied ethics just as well as genuine theoretical normative ethics, won the Lilli-Bechmann-Rahn-Preis of the University Erlangen Nürnberg in 2005.

  ### *Dr. Ir. Behnam Taebi*
  Behnam Taebi is an Assistant Professor of philosophy at Delft University of Technology, and a Research Fellow at Harvard Kennedy School's Belfer Center for Science and International Affairs. His research interests are in energy ethics, responsible innovation and nuclear ethics. He studied Material Science and Engineering (2006) and received his Ph.D. in Philosophy of Technology (2010). Taebi is working on his Veni-Research on 'ethics and governance of multinational nuclear waste repositories'. He is the leading editor of a volume on "The Ethics of Nuclear Energy" (under contract with Cambridge University Press) and a special issue of Journal of Risk Research on "Socio-technical Challenges of Nuclear Power Production."
  Taebi is currently based in Cambridge (Massachusetts), where he spends a sabbatical year at Harvard University. He is involved in in the Belfer Center's 'Project on Managing the Atom' and 'International Security Program'.

- ### *Prof. Ir. Bert van Wee*
  Bert van Wee is professor in Transport Policy and Logistics and head of the section at Delft University of Technology, faculty Technology, Policy and Management. His main interests are in long-term developments in transport, the environment, safety and accessibility, and in policy analyses.

- ### *Prof.dr. G.L.L.M.E. (Genserik) Reniers*
  Genserik Reniers obtained a Master's degree in chemical engineering at the Vrije Universiteit Brussel and received his PhD in Applied Economic Sciences from the University of Antwerp. He was appointed in a part-time capacity as a Full Professor at the Delft University of Technology, at the Safety and Security Science Group. At the University of Antwerp, he is a visiting/full professor lecturing amongst others in chemistry, organic chemistry, chemical process technology, and Technological Risk Management. He is also visiting professor Security Management at the Antwerp Management School, Risk Management at ITMMA, and Risk Analysis in a Postgraduate Disaster Management at VESTA. At the KU Leuven campus of the HUB in Brussels he lectures as a professor, amongst others, in prevention management, advanced occupational health and safety management and chemical processes/unit operations. His main research interests concern the collaboration surrounding safety and security topics and socio-economic optimization within the chemical industry.

  ### *Dr. ir. Frank Guldenmund*
  Frank works for the Safety and Security Science Group; this group conducts research into the way in which safety- and security-related risks are assessed and managed. It focuses on developing Safety Science methodologies and transfer of knowledge and methodologies between different application domains like chemistry and chemical engineering, applied mathematics, electrical engineering, systems engineering and policy analysis, transport technology and management, psychology, criminology and law, and organisational sciences. Frank's specialization is safety and culture.

- ### *Prof Dr. ir Nick van de Giesen*
  Nick van de Giesen received his BSc and MSc in Land and Water Management from Wageningen

Agricultural University, The Netherlands, in 1984 and 1987, respectively, and his PhD in Agricultural and Biological Engineering from Cornell University, Ithaca, NY, in 1994. After a postdoctoral position with the West Africa Rice Development Association, Côte d'Ivoire, he was a Senior Researcher for six years with the Center for Development Research (ZEF), University of Bonn, Germany, where he was Scientific Coordinator of the Global Change in the Hydrological Cycle Volta Project. Since 2004, he has been with Delft University of Technology, The Netherlands, where he holds the "Van Kuffeler" Chair of Water Resources Management. Since April 2009, he has served as chairman of the Delft Research Initiative Environment. In 2011 he was science representative in "Topteam Water", which advised the Dutch government on innovation in the water sector.

Colophon

This E-book is based on the Massive Open Online Course Responsible Innovation (for the re-run see: https://www.edx.org/course/responsible-innovation-ethics-safety-delftx-ri101x) which was offered by the TU Delft in November 2014 - January 2015 on the edX-platform. It contains all the content covered by the web lectures.