The Primacy of the Public

The Primacy of the Public

Ethical Design for Technology Professionals

Marcus Schultz-Bergin

MSL Academic Endeavors
CLEVELAND, OHIO

Contents

Preface

The Primacy of the Public is an open educational resource written and compiled by Dr. Marcus Schultz-Bergin of Cleveland State University.

SECTION I

Foundations of Engineering Ethics

1.

An Ethics Primer

Key Themes & Ideas

- Ethics and ethical decision making involve reasoning about values and principles
- Engineering & Design Ethics and Technology Ethics are two closely related sub-fields of ethics
- We should understand technology as encompassing any human designed tool or technique used to change or interact with the world

1. What is Ethics?

> *The most important thing is not life, but the good life* – Plato (*Crito* 48b)

Watch the following two short (~5 minute) videos to get a sense of what *ethics* and *ethical decision-making* involves. You will also be introduced to a few key ethical concepts that will be explored in more detail later in this book. Give particular focus to the concepts of **values**, **principles**, and **purpose** as well as the **4 key drivers of moral decision-making**.

1. What is Ethics?
2. Ethical Decision Making

2. Engineering & Technology Ethics

Ethics, in general, is the study of how we ought to live. But this book is not about ethics in general. Rather, it focuses on two related subfields of ethics:

- **Engineering & Design Ethics**: A subfield of professional ethics that focuses on the ethical conditions of successful engineering and design. Focuses on the purpose of engineering and design as well as the values and principles that should guide engineering and design.
- **Technology Ethics**: A field of ethics focused on the ethical investigation of technology and its effects on society. Focuses on the purpose of technology in society as well as the values and principles that should guide the introduction and use of technology in society.

Put another way, Engineering & Design Ethics focuses on the question "how ought we act as professional engineers and how ought we design technology?" while Technology Ethics focuses on the question "how ought we integrate technology into society?" While these questions certainly do not exhaust the field of ethics, they are becoming increasingly central to the basic ethical question of how we ought to live for the following sorts of reasons:

- Technology is increasingly the medium through which we act with & toward others
- Technology increasingly shapes the social, political, economic, biological, psychological, & environmental conditions in which humans strive to flourish
- Technology makes us more powerful as a species but more vulnerable and interdependent as individuals
- Technological design and implementation decisions are concentrated in the hands of an increasingly elite few who do not embody the interests/needs/values of all
- Technology in our global economy manifests an impersonal drive to efficiency, optimization, measurement, control, & other machine values, often at the expense of humane values such as justice, compassion, nobility, freedom, and leadership
- Technological choices now have *existential* implications for future generations, for the survival/flourishing of humanity & others with whom we share the planet
- For humanity to have a future worth wanting, the growing power of technology must be matched by growth in human *wisdom & responsibility*; our efforts must be rebalanced to fuel the latter kind of growth that is presently in neglect

For our purposes, the term **technology** refers to any human designed tool or technique through which humans change, transform, use, or engage with the environment. In a later chapter we will explore the concept of technology in more detail, but for now the key is to recognize that all fields of engineering are united in their creation of technology. This is in contrast to the narrow use of the term to refer specifically to electronic tools like computers (which we may group under something like "digital technologies").

In the chapters that follow we will explore theoretical, conceptual, and practical issues in engineering and

technology ethics. Throughout that process, our central aim will be to develop the skills and tools necessary to be an ethical engineer and technological designer as well as a more informed citizen, capable of engaging in social debates about technology and its place in society.

Check Your Understanding

After successfully completing this chapter, you should be able to answer all the following questions:

- What is **ethics**? What question does it aim to help us answer?
- What are **values**? How do they relate and differ from **principles**? And what role do both values and principles play in ethical decision-making?
- What are the 4 key drivers of ethical decision-making?
- What is **engineering ethics** about? How does it relate to **technology ethics**? What are some of the key questions each of these fields focuses on?
- What is **technology**?

2.

Ethics & Professional Responsibility

Key Themes & Ideas

- Engineering is a profession, characterized by self-regulation and a commitment to protecting and promoting the public good
- As professionals, engineers have special moral responsibilities that go above and beyond what the law or common morality require
- The core moral commitment of engineers and computer scientists is to protect and promote the welfare of the public
- Professional responsibility is a multifaceted concept incorporating at least 4 different meanings, all of which are important to engineering
- Professional codes of ethics outline many of the specific professional obligations of engineers and computer scientists, but are not exhaustive
- Engineers are sometimes understood to have a special concern for safety, which underwrites their obligations to adhere to a standard of care and to sound the alarm if they are aware of potential harm or unethical behavior

---

This narrative was originally published by Michael Davis in "Thinking Like an Engineer: The Place of a Code of Ethics in the Practice of a Profession"[1]

1. Michael Davis (1991),"Thinking Like an Engineer: The Place of a Code of Ethics in the Practice of a Profession," *Philosophy & Public Affairs* 20:2, 150-167. He notes that his narrative is derived from testimony contained in The Presidential Commission on the Space Shuttle Challenger Disaster (Washington, D.C.: U.S. Government Printing Office, 1986).

On the night of 27 January 1986, Robert Lund was worried. The Space Center was counting down for a shuttle launch the next morning. Lund, vice-president for engineering at Morton Thiokol, had earlier presided over a meeting of engineers that unanimously recommended against the launch. He had concurred and informed his boss, Jerald Mason. Mason informed the Space Center. Lund had expected the flight to be postponed. The Center's safety record was good. It was good because the Center would not allow a launch unless the technical people approved.

Lund had not approved. He had not approved because the temperature at the launch site would be close to freezing at lift-off. The Space Center was worried about the ice already forming in places on the boosters, but Lund's worry was the "O-rings" sealing the boosters' segments. They had been a great idea, permitting Thiokol to build the huge rocket in Utah and ship it in pieces to the Space Center two thousand miles away. Building in Utah was so much more efficient than building on-site that Thiokol had been able to underbid the competition. The shuttle contract had earned Thiokol $150 million in profits.

But, as everyone now knows, the O-rings were not perfect. Data from previous flights indicated that the rings tended to erode in flight, with the worst erosion occurring on the coldest preceding lift-off. Experimental evidence was sketchy but ominous. Erosion seemed to increase as the rings lost their resiliency, and resiliency decreased with temperature. At a certain temperature, the rings could lose so much resiliency that one could fail to seal properly. If a ring failed in flight, the shuttle could explode.

The Challenger on lift-off

Unfortunately, almost no testing had been done below 40°F. The engineers' scarce time had had to be devoted to other problems, forcing them to extrapolate from the little data they had. But, with the lives of seven astronauts at stake, the decision seemed clear enough: Safety first.

Or so it had seemed earlier that day. Now Lund was not so sure. The Space Center had been "surprised," even "appalled," by the evidence on which the no-launch recommendation had been based. They wanted to launch. They did not say why, but they did not have to. The shuttle program was increasingly falling behind its ambitious launch schedule. Congress had been grumbling for some time. And, if the launch went as scheduled, the president would be able to announce the first teacher in space as part of his State of the Union message the following evening, very good publicity just when the shuttle program needed some.

The Space Center wanted to launch. But they would not launch without Thiokol's approval. They urged Mason to reconsider. He reexamined the evidence and decided the rings should hold at the expected temperature. Joseph Kilminster, Thiokol's vice-president for shuttle programs, was ready to sign a launch approval, but only if Lund approved. Lund was now all that stood in the way of launching. Lund's first response was to repeat his objections. But then Mason said

something that made him think again. Mason asked him to think like a manager rather than an engineer. (The exact words seem to have been, “Take off your engineering hat and put on your management hat.”) Lund did and changed his mind. The next morning the shuttle exploded during lift-off, killing all aboard. An O-ring had failed.

Should Lund have reversed his decision and approved the launch? In retrospect, of course, the answer is obvious: No. But most problems concerning what we should do would hardly be problems at all if we could foresee all the consequences of what we do. Fairness to Lund requires us to ask whether he should have approved the launch given only the information available to him at the time. And since Lund seems to have reversed his decision and approved the launch because he began to think like a manager rather than an engineer, we need to consider whether Lund, an engineer, should have been thinking like a manager rather than an engineer. But, before we can consider that, we need to know what the difference is between thinking like a manager and thinking like an engineer.

Reflecting on the *Challenger* disaster, and especially Mason’s request that Lund “take off [his] engineering hat and put on [his] management hat” provides a useful starting point for thinking about the relationship between engineering and ethics. For it is typical, in reviewing the case, to agree that Lund should have stuck to his initial judgment and not approved the launch. That the correct judgment was made with his “engineering hat” on and the incorrect judgment with his “management hat” suggests that engineers think differently from managers. And, more importantly, that the way in which they think differently has important moral implications; after all, the switch to management thinking cost of the lives of 7 astronauts. Thus, in figuring out the difference between thinking like a manager and thinking like an engineer, we can come to understand the professional responsibilities of engineers.

1. Engineering as a Profession

“Engineering is an important and learned profession.” Thus, reads the first line of the National Society of Professional Engineers (NSPE) *Code of Ethics*. While the use of the term “profession” may initially seem unimportant, for in everyday speech we talk about “professional athletes” and “professional attire”, the choice of that term is deliberate and important. For there is a more technical meaning of the term **profession**. And understanding *that* meaning of the term, and why engineering is a profession in *that* sense, will go a long way to helping us understand how thinking like an engineer is different from thinking like a manager and why ethics is so central to engineering.

Engineering is not the only profession, in the sense we are using the term here. The two most commonly identified professions are doctors and lawyers. Knowing this, and knowing what is special about doctors and lawyers, can help us start to understand what is special about engineering as well. Our health and our freedom are both absolutely vital aspects of our lives. And keeping them in good shape – keeping our minds and bodies healthy and keeping us out of confinement – takes a high level of expertise that not just anyone has. Moreover, because

doctoring and lawyering require a high degree of expertise, as a society we expect the practitioners to keep each other in check. Although we, as a society, do impose legal requirements on doctors and lawyers, we also hand over a great deal of power to their associations or societies for establishing rules and regulations that either become a part of the law or go beyond it. Cases of "disbarment", for lawyers, are cases where a practicing lawyer has been stripped of their ability to practice law by the relevant "bar association", which is an association of other practitioners, not judges or lawmakers.

This discussion of what is special about doctors and lawyers should help us see what is special about *all* professions. Drawing from the above examples, it is typical to identify **3 key features of a profession**:[2]

1. *Advanced Expertise*. Professions require sophisticated skills ("knowing-how") and theoretical knowledge ("knowing-that") in exercising judgment that is not entirely routine or susceptible to mechanization. Preparation to engage in the work typically requires extensive formal education, including technical studies in one or more areas of systematic knowledge as well as broader studies in liberal arts (humanities, sciences, arts). Generally, continuing education and updating knowledge are also required.
2. *Self–regulation*. Well-established societies of professionals are allowed by the public to play a major role in setting standards for admission to the profession, drafting codes of ethics, enforcing standards of conduct, and representing the profession before the public and the government. Often this is referred to as the 'autonomy of the profession', which forms of the basis for individual professionals to exercise autonomous professional judgment in their work.
3. *Public Good*. The occupation serves some important public good, or aspect of the public good, and it does so by making a concerted effort to maintain high ethical standards throughout the profession. For example, medicine is directed toward promoting health, law toward protecting the public's legal rights, and engineering toward technological solutions to problems concerning the public's well-being, safety, and health. The aims and guidelines in serving the public good are detailed in professional codes of ethics, which, in order to ensure the public good is served, need to be taken seriously throughout the profession.

In short, professions involve extensive and uncommon knowledge that is essential to some vital public good. But because the knowledge involved is uncommon, society in general must trust professionals to use their expertise for the good of the public and hold each other accountable for doing so. Without expertise in structural engineering, for instance, I cannot personally determine whether a bridge is safe for me to cross. Instead, I am forced to trust that the engineers who designed the bridge and those who regularly check it for integrity are competent, conscientious, and committed to using their power responsibly. To paraphrase Peter Parker's (Spider-man's) uncle Ben, we may say, that *advanced expertise* generates the power while *self-regulation* and a commitment to some aspect of the *public good* establish the responsibility. We will, therefore, focus on those latter 2 features for understanding professional responsibility and engineering ethics.

Now that we know what a profession is and have some idea of what makes engineering a profession (although

2. This particular framing of the features comes from Mike W. Martin & Roland Schinzinger (2005). *Ethics in Engineering*, 4th ed. McGraw Hill Education.

we'll detail that more below), we can make a first pass at understanding what changed when Lund "put on his manager hat". In doing so, first, he ceased to exercise autonomous professional judgment. Instead, he allowed his judgment to be controlled by the needs of the company and the pressures of the management team. This is a failure to properly *self-regulate*. Second, he failed to take seriously his obligation to serve the public good, putting the interests of the company above the well-being of the public (in particular, the astronauts). This, clearly, was a failure to serve the important *public good* that engineering is supposed to serve. But to understand this further, it'll be helpful to dig deeper into the relationship between engineering and the public good.

2. Engineering and the Public Good

All professions involve a commitment to serve some aspect of the public good, above and beyond what is required by law or ordinary morality. In this way, professions create their own role-specific ethical code: ethical requirements that apply only to professionals and to them only in their professional capacity. For doctors (and other medical professionals) this is a commitment to the vital public good of *health*. For lawyers (and other legal professionals) this is a commitment to the vital public good of *justice*. For engineers and computer scientists (as well as other technological professionals) the commitment is to the vital public good of *welfare*. Or, as it is typically stated by many of the engineering professional societies, the "safety, health, and welfare of the public".

But what does it mean to be committed to the welfare of the public? Like medicine and law, this is partially answered by the specific **action-oriented obligations** that engineers have. These are obligations to do (or not do) specific things under specific circumstances. For instance, medical professionals are ethically required to render medical aid to those in need whereas a member of the general public is not. Engineers, too, have a number of action-oriented obligations and we will examine some of them later in this chapter. But, to best understand a professional's commitment to the public good, and so engineering's commitment to the welfare of the public, it is better to focus on how the commitment changes a person's *reasoning* and *judgment*. Because all professions involve special expertise, they all involve a significant amount of high-level reasoning and the exercise of judgment. Many of the issues professionals face aren't answered by simply consulting an action handbook, but instead require considering a complex set of information to render expert judgment. A core part of this judgment involves bringing the technical expertise to bear on the issue. But the other aspect of this judgment is about what values, interests, or goals are to be privileged above others. For instance, when our doctor is deciding on a treatment plan, we expect her to make that decision on the basis of what is *best for our health* rather than, for instance, what is best for her schedule or what will make the hospital the most money.

Thus, on top of the action-oriented obligations, we can say that a profession's commitment to the public good involves a special **reasoning requirement:** a requirement to reason in particular ways to privilege certain interests above others in exercising professional judgment. For the doctor, it is privileging the health of the patient; for the lawyer, the best interests of the client. The engineer, then, is required to *privilege the welfare of the public* in her professional judgments and design decisions. As wrong as it would be for the doctor to privilege the hospital's finances above her patient's health, it is similarly wrong for the engineer to privilege the company's finances above the public's welfare.

This focus on privileging the public welfare in all professional reasoning and judgment is also captured in the NSPE Code of Ethics (as well as the engineering and computer science codes of ethics produced by other professional societies). As the NSPE puts it: "Engineers, in the fulfillment of their professional duties, shall hold paramount the safety, health, and welfare of the public." Like the best interests of the patient or client, exactly what it means to hold paramount the public's welfare is a matter of judgment in particular contexts. And there will certainly be room for disagreement among different engineers. But what must hold constant in those disagreements is putting the primary focus on the welfare of the public rather than personal or business concerns.

Engineering and "Public Welfare"

What do the engineering codes of ethics mean by "public welfare"? Like *profession*, **welfare** is a term that is used in a variety of ways in society but has an important meaning in the context of ethics. For our purposes, welfare is equivalent to **well-being** and it encompasses a wide variety of conditions, activities, and more that contribute to a life going well. This is why "public good" is a useful equivalence, for in ethics when we think of welfare we are thinking about what is "good for" a person (or potentially a group or system).

So what is "good for" a person? Various theories of well-being exist in the philosophical and psychological literature. Rather than wading into those debates, we can take a step back from the question "what is well-being?" and instead ask "what conditions are necessary for the realization of some of the most commonly recognized elements of well-being?" This is, in part, because although various competing theories of well-being exist, they all still typically identify many of the same elements. Their disagreement is more over what makes something an element of well-being rather than the elements themselves.

So what are those conditions necessary for the realization of well-being? Well, those most relevant to engineering include things like having food, shelter, and water, having satisfying human relationships, having free movement and expression, and having a satisfactory relationship to the natural world. Engineering and technology can affect all of these conditions both positively and negatively. Pollution from engineering projects can make access to (clean) water difficult while technological advancements in water purification can make access easier. Recent technological creations like the internet have greatly altered how we relate to each other, making it both easier and more difficult to have satisfying human relationships. Planes, trains, and automobiles all greatly enhanced freedom of movement while the internet has certainly enhanced peoples' ability to express themselves (for better and for worse!).

3. Being a Responsible Professional

The concept of **responsibility** shows up regularly in discussion of ethics, including professional ethics where it often shows up as **professional responsibility**. Similarly, engineering ethics discussions often take on the form of reviewing past engineering disasters (like the *Challenger*) and asking "who was responsible?" But the question "who was responsible?" only captures one small portion of the concept of responsibility. And, it is likely the

least important aspect of all. Thus, to further deepen our understanding of engineering as a profession we can dig into the various dimensions of professional responsibility. While we may say that Lund is responsible for the *Challenger* disaster, we could mean that in a variety of ways.

3.1. Backward-looking & Forward-looking Responsibility

The first distinction we can make is between Backward-looking Responsibility and Forward-looking Responsibility. **Backward-looking Responsibility** captures the "who is responsible?" sort of question and is often associated with the idea of *blame*. We are looking back at something that has already occurred and asking "who is to blame here?" or "who should be held accountable?" To avoid confusion, in fact, it is common to reframe backward-looking responsibility as **accountability**. So, instead of saying "Lund is responsible for the disaster" we might say "Lund is accountable for the disaster"; similarly, instead of "Lund should be held responsible for the disaster" we might say "Lund should be held accountable for the disaster". This sort of responsibility is of course important; the self-regulating aspect of a profession means, among other things, that it takes seriously identifying who is accountable for engineering errors (if anyone) and holding such people accountable for their failures. But it would be a mistake to focus on it too much, as it is effectively the "final form" responsibility will take, once all other dimensions have already been flouted. We only have to ask "who should be held responsible" for a failure once that failure has occurred. Other dimensions of responsibility can help us prevent the failure in the first place. Moreover, as hinted above, there is not always someone who should be held accountable for a failure; honest mistakes can happen. But that does not mean professional responsibility does not apply in any sense.

And so we can turn to **Forward-looking Responsibility**. This is the sort of responsibility we have in mind when we say things like "who is going to take responsibility for getting the job done?" or, more generally, "whose responsibility is it?" These questions are asked prior to carrying out some activity and so prior to any potential failure that may occur. And if we take the issue of forward-looking responsibility seriously then it significantly decreases the chances of failure, or at least of failure that anyone would be accountable for. We should also note the relationship between forward-looking and backward-looking responsibility here: If someone recognizes and accepts that they are responsible (in the forward-looking sense) for some aspect of a project and then they fail to execute (or they execute it poorly) that is typically good evidence that they should be held responsible (in the backward-looking sense) for that failure.

While it is not uncommon to think of responsibility almost exclusively in terms of its backward-looking dimension, professional responsibility is much more about the forward-looking dimension. It is about consciously recognizing what needs to be done, how it needs to be done, and then "taking the responsibility" to do it correctly. When we focus too much on backward-looking responsibility, we tend to focus on how we can *avoid* responsibility. We don't want to be "held accountable" and so we may simply not act where action is required. This may help us not feel guilty, but it does little to improve the world. Since, in that case, either the thing that needs to be done simply does not get done or it gets done by people with less expertise. A true expert professional accepts, to repeat the line from Spider-man again, that with great power comes great (forward-looking) responsibility.

3.2. The Responsibility to Think & The Responsibility to Act

Cutting across the forward-looking/backward-looking distinction, we can also distinguish between the responsibility to think and the responsibility to act. This neatly tracks the distinction mentioned previously between action-oriented obligations and reasoning requirements. Sometimes we may say someone "fulfilled their professional responsibilities" or more generally talk about what our responsibilities are. Typically, in this case, what we mean is that they successfully carried out the actions that they were obligated to carry out. This, then, is about the **responsibility to act**. It is the responsibility to meet your action-oriented obligations. Notice, that this can be discussed in both a forward-looking and a backward-looking way. We can ask "what are my professional responsibilities in this situation?" This is effectively asking "what should I do?" and is thus forward-looking. But we can also suggest that someone should be "held responsible" for a failure because they failed to meet their action-oriented obligations. That would, of course, be backward looking.

But, as we have already suggested, being a professional is not just about doing the right actions in the right circumstances. Much of being a professional involves thinking in the right sorts of ways. And so we can also talk about the **responsibility to think**, by which we mean consider the right things in the right circumstances when exercising professional judgment. This is especially important for professions given that professionals can be held (legally or professionally) responsible for a failure to properly consider factors in making their judgments, in those judgments lead to harm. And this can be true even if they did not intend to cause harm nor could they have foreseen that harm would occur. One useful way to distinguishing the responsibility to think from other uses of the term responsibility is to frame it as being **conscientious**. Conscientiousness just means successfully considering the right things in the right way in our thinking and decision-making. Thus, we can understand the requirement to hold paramount the welfare of the public as a form of professional conscientiousness.

3.3. Codes of Ethics & Professional Responsibility

The final thing to say about responsibility isn't so much an additional dimension beyond the four identified above. Rather, it is to note the relationship between Professional Responsibility and the **Codes of Ethics** produced by various professional societies. These Codes of Ethics are written documents that state (some of) the action-oriented obligations and reasoning requirements of professional engineers. Importantly, although they are put out by various societies that engineers may choose to be a part of, the responsibilities outlined in the codes of ethics apply to all practicing engineers, even those who are not members of the society. These societies are not *creating* the responsibilities; instead, they are making them explicit and organizing them. This fact partly explains why the codes of ethics of all engineering and computer science societies look very similar, as evidenced by their common language about the commitment to public health shown below.

Importantly, while Codes of Ethics can be great initial resources for identifying professional responsibilities, especially action-oriented obligations, they are not exhaustive lists. For one thing, many of the responsibilities outlined are quite vague – they are principles rather than rules – and so require judgment to fully flesh out. Moreover, given how Codes of Ethics are compiled, they represent the bare minimum set of responsibilities that large groups of professionals agree on. Finally, Codes of Ethics as quite slow to change and so do not always clearly keep up with social and professional changes. Given all that, a responsible professional will be

conscientious of the value and the limits of Codes of Ethics and appeal to them accordingly. They will be a resource, but not a holy book; a philosophical text, but not a lawbook.

The Primacy of the Public in Engineering Codes of Ethics

"Fundamental Canon 1" of the *National Society of Professional Engineers' Code of Ethics* states that "Engineers, in the fulfillment of their professional duties, shall hold paramount the safety, health, and welfare of the public."

The first principle of the Software Engineering Code of Ethics and Professional Practice, produced jointly by the Association for Computing Machinery & the Computer Science division of Institute of Electrical and Electronics Engineers (IEEE), states that "Software engineers shall act consistently with the public interest."

The first principle of ACM's Code of Ethics and Professional Responsibility, applicable to all computing professionals, states that "A computing professional should contribute to society and to human well-being, acknowledging that all people are stakeholders in computing" while the second principle requires that computing professionals "avoid harm".

The first requirement of the IEEE Code of Ethics similarly states that (electrical) engineers should "hold paramount the safety, health, and welfare of the public, to strive to comply with ethical design and sustainable development practices, to protect the privacy of others, and to disclose promptly factors that might endanger the public or the environment."

The American Institute of Chemical Engineers' Code of Ethics states that (chemical) engineers shall "hold paramount the safety, health and welfare of the public and protect the environment in performance of their professional duties."

Finally, the American Society of Mechanical Engineers' code of ethics echoes earlier codes of ethics with the first fundamental canon that "engineers shall hold paramount the safety, health and welfare of the public in the performance of their professional duties" while also stating, as the first principle of the code of ethics, that "engineers… [use] their knowledge and skill for the enhancement of human welfare."

4. A Special Concern for Safety

Recall that the central professional responsibility of all engineers and computer scientists is to hold paramount the safety, health, and welfare of the public. We suggested earlier that that general responsibility is best understood as a reasoning requirement indicating how technology professionals should think and exercise professional judgment. But there are also a number of action-oriented obligations that follow from this responsibility. Many of

these are outlined in the various codes of ethics, and we will not canvass them all here, but instead focus on a few core ones.

Many of the core action-oriented obligations of engineers can be captured under the broad heading of **A Special Concern for Safety**. This is, broadly, just another way of stating the same commitment to public welfare, but it narrows the focus a little bit by emphasizing *avoiding harm* over (for instance) *doing good*. When we have a special concern for safety, we are most interested in ensuring our products and projects don't impose unacceptable risks on the public. This is in contrast to a *doing good* approach where we may be thinking much more about the potential benefits of our products and projects. Of course, both approaches are important, and we will want to keep them both in mind in general. However, the majority of the action-oriented obligations outlined in codes of ethics are focused on ensuring safety rather than using engineering skills to do good. And so, for the purposes of examining some of those obligations, having the "safety first" mentality in mind can be helpful.

4.1. The Standard of Care

To have a special concern for safety is to treat safety as an especially important consideration in decision-making. One way that this comes about in professions like engineering is through professional standard setting. Recall that because professions involve a high level of expertise professions are granted a degree of autonomy in regulating their own practitioners. Sometimes these regulations are purely professional and have no legal status. But sometimes they rise to the level of legal requirements. In engineering, this happens with some engineering standards like building codes. But a more general legal standard, rooted in professional self-regulation, that all professions have is known as the standard of care.

The **standard of care** is defined as "that level or quality of service ordinarily provided by other normally competent practitioners of good standing in that field, contemporaneously providing similar services in the same locality and under the same circumstances."[3] The first thing to notice is that this is relatively vague as it does not tell us exactly what that level or quality of service is. Instead, it points us to how it gets determined. This is, to be clear, the legal definition of the standard of care. But what it does is offload the determination of whether the standard of care has been violated in a given case to the profession itself. The only way we find out, for sure, whether the standard of care has been violated, is by a court judgment after the court hears from a variety of expert witnesses – namely, those competent practitioners of good standing in the field.

So, although adhering to the standard of care is an action-oriented obligation as it defines a requirement to conduct proper tests, use proper materials, exercise proper oversight, etc. we cannot say definitively what it requires. And, of course, it changes over time as standards and practices change. This is part of what makes it so important to discuss, as it shows us one clear way in which even if your only goal was to not break the law, and you did not care at all about broader professional ethics, you would still need to be conscientious and exercise responsible judgment. For otherwise you may fail to exercise the standard of care.

The standard of care plays an important role in establishing professional responsibility. As indicated earlier, engineers are certainly not accountable for *every* bad thing that happens as a result of their work. Bad things can

3. This definition is found in the US court case *Paxton v. County of Alameda* (1953) 119 C. A. 2d 393, 398, 259 P. 2d 934)

happen even if no one is responsible for them happening. But, of course, all people are held accountable for bad things that happen as a result of deliberate attempts to cause bad things as well as reckless activities that result in bad things. And this is true of professionals as well: If they intended to cause harm with their designs, then they are responsible for the harm caused; if they reasonably could have known that their design would result in harm, even if they did not intend for it to happen, then they are still responsible for the resulting harm. But what about harms that were not intended and could not reasonably have been foreseen?

We use the standard of care to answer *that* question. Legally speaking, the standard of care establishes the line between "honest mistake" resulting in harm and *negligent* harm. If an engineer is found to have failed to adhere to the standard of care – perhaps they did not run a standard test or failed to consider an aspect of the project that competent professionals would consider – and that resulted in harm, then the engineer is legally and professionally accountable for that harm. However, if they are found to have adhered appropriately to the standard of care then, even though harm resulted, they are not accountable.

In short, the standard of care indicates that engineers must "do their due diligence" in all their work, especially in terms of ensuring safety as the focus here is on whether the engineer is responsible for resulting harms. This means keeping up to date on standards and practices of the relevant field, conducting all tests and using appropriate materials, and whatever else the profession decides is part of being a competent member of that profession.

A number of the specific obligations listed in codes of ethics correspond to this idea. Engineers are required to only work in fields in which they are competent, they are required to adhere to all engineering standards, and they are required to continue their education to stay up-to-date in the field.

4.2. Sounding the Alarm

The standard of care and related obligations all speak to what an individual engineer should do or not do in her own work. But a number of the action-oriented obligations that are most commonly associated with engineering are of a different sort. They still emphasize a special concern for safety, but they all in different ways require engineers to take action when others have acted (or failed to act) in particular ways. We can group these related obligations under the heading of **Sounding the Alarm**, as they all involve speaking out about potential harm or wrongdoing.

Perhaps the most well-known obligation in this category is **whistleblowing**. In general, whistleblowing involves going outside standard communication channels to inform relevant authorities of potential harm or wrongdoing. Engineering codes of ethics include a number of specific obligations which are all forms of whistleblowing. Canon 1 of the IEEE code of ethics states that engineers have an obligation to "disclose promptly factors that might endanger the public or the environment" while one of the NSPE's "rules of practice" states that "If [an engineer's] professional judgment is overruled, under circumstances where the safety, health, property, or welfare of the public are endangered, they shall notify their employer or client and such other authority as may be appropriate." Similarly, another rule of practice states that engineers must "notify the proper authorities and withdraw from further service on the project… if the client or employer insists on… unprofessional conduct." Finally, the NSPE code of ethics also states that "Engineers having knowledge of any alleged violation of this

Code shall report thereon to appropriate professional bodies and, when relevant, also to public authorities, and cooperate with the proper authorities in furnishing such information or assistance as may be required."

There are a few things to notice about these statements. First, whistleblowing can be either "internal" or "external", depending on the circumstances. It may involve notifying your boss's boss or it may involve a call to the federal government. It leaves to the judgment of the engineer what contact is appropriate to responsibly report the issue. Some organizations, including some parts of the US government, have "whistleblower hotlines" that are especially set up to handle such things. Other organizations do not and may actively discourage such behavior. Nonetheless, as this makes clear, engineers and computer scientists are professionally (and, as it happens, legally) obligated to whistle blow under the relevant circumstances. Second, the obligation to whistle blow applies in cases of potential harm – thus fitting nicely with the special concern for safety – but also in cases of other ethical violations that may not result in harm. Since not all the NSPE code of ethics directly corresponds to safety, but engineers are required to report *any* violations, that means violations that may not risk any harm at all. Finally, an engineer cannot simply whistle blow and continue to work on a project they have identified as ethically problematic; instead, as the code states, they must resign from the project as well as report the issue. A very strong obligation indeed! One that can, as several historical cases of whistleblowing make clear, risk significant personal cost. Nonetheless, being a responsible professional means putting safety first, even above one's job. As we learned in the Nuremberg Trials of the Nazis, "I was just taking orders" is not a valid defense for wrongdoing.

The other, related, obligation is known as the **Duty to Warn**. Like whistleblowing, this involves going outside of standard channels to report on potential harm. But, unlike whistleblowing, the duty to warn *only* applies to potential harm (including property destruction) and it is limited to cases of *imminent harm*. There is a similar duty for other professionals. If a doctor or lawyer has good reason to believe that her patient or client poses a genuine and imminent threat to another person then she is professionally and legally obligation to break confidentiality to provide adequate warning. For the engineer, the way this often comes out is with structural inspection. If an engineer has been hired to inspect a building, bridge or similar and finds that it poses imminent risk then regardless of her broader obligation of client confidentiality, she must take the appropriate actions to warn of the harm. That may mean warning legal authorities, but it could mean warning the residents of the building, if for instance collapse or some other potential harmful event is imminent.

Whistleblowing and the Duty to Warn are specific action-oriented obligations of professional engineers. But they also illustrate more broadly what it means to have a special concern for safety and to hold paramount the safety, health, and welfare of the public. In both cases, engineers are expected to violate a number of other professional obligations – most notably the obligation of employer/client confidentiality – for the sake of protecting the welfare of the public. Similarly, these obligations illustrate the expectation that a professional engineer will be prepared to sacrifice her personal well-being for the sake of the public's welfare. Just as medical professionals put themselves at severe risk of infection from COVID-19 to treat patients throughout the pandemic – and a number of them lost their lives in the process – responsible technological professionals should be prepared to put themselves at risk if doing so is necessary to fulfill their obligation to hold paramount the safety, health, and welfare of the public.

Check Your Understanding

After successfully completing this chapter, you should be able to answer all the following questions:

- What is a **profession** and what are its key features? How does engineering fill out those key features?
- What are **action-oriented obligations** and what are some of the action-oriented obligations of professional engineers?
- What is a **reasoning requirement** and what is the central reasoning requirement of engineers?
- What is **backward-looking responsibility** and how does it compare to **forward-looking responsibility**?
- What is the distinction between **responsibility to think** and **responsibility to act**? How does it relate to the distinction between action-oriented obligations and reasoning requirements?
- What is a **code of ethics**? Who do **codes of ethics** apply to?
- What is the **standard of care**? What feature(s) of a profession does it relate to and how?
- What is an engineer's obligation to **sound the alarm**? What are the two specific ways an engineer may sound the alarm?

References & Further Reading

Bazerman, M.H. & Tenbrunsel, A.E. (2011). *Blind Spots: Why We Fail to Do What's Right and What to Do About It* (Princeton University Press).

Curd, M. & May, L. (1984). *Professional Responsibility for Harmful Actions* (Kendall/Hunt Publishers).

Davis, M. (1998). *Thinking Like an Engineer: Studies in the Ethics of a Profession* (Oxford University Press).

Davis, M. (2003). "Whistleblowing," in *The Oxford Handbook of Practical Ethics*, Hugh LaFollette (ed.) (Oxford University Press).

DeGeorge, R.T. (1981). "Ethical Responsibilities of Engineers in Large Organizations," *Business and Professional Ethics Journal* 1:1, pp. 1-14.

Harris, C. E., Pritchard, M.S., James, R.W., Englehardt, E. E., & Rabins, M.J. (2019). *Engineering Ethics: Concepts and Cases*, 6th ed. (Cengage Learning).

Janis, I. (1982). *Groupthink*, 2nd ed. (Houghton Mifflin).

Martin, M. & Schinzinger, R. (2015). *Ethics in Engineering*, 4th ed. (McGraw Hill Publishers).

3.

Technology in Society

Key Themes & Ideas

- Engineering and technology function in a social context
- Technology is best understood as a techno-social system
- Technology mediates our experience of and interaction with the world
- The Control Dilemma shows us that our ability to control the effects of technology are inversely related to our ability to anticipate those effects
- In mediating our world, technology can also change what we value or how we understand our values and principles
- Technology can intentionally and implicitly mediate perceptions and actions
- Technological mediations can take many forms, but the three most common are guidance, persuasion, and coercion

---

As it departed on its maiden voyage in April 1912, the *Titanic* was proclaimed the greatest engineering achievement ever.[1] Not merely was it the largest ship the world had seen, having a length of almost three football fields; it was also the most glamorous of ocean liners, complete with a tropical vinegarden restaurant and the first seagoing masseuse. It was touted as the first fully safe ship. Since the worst collision envisaged was at the juncture of two of its sixteen watertight compartments, and since it could float with any four compartments flooded, the *Titanic* was believed to be virtually unsinkable.

1. The discussion that follows was originally presented by Mike Martin and Roland Schinzinger in *Ethics in Engineering* (1989)

Buoyed by such confidence, the captain allowed the ship to sail full speed at night in an area frequented by icebergs, one of which tore a large gap in the ship's side, flooding five compartments. Time remained to evacuate the ship, but there were not enough lifeboats to accommodate all the passengers and crew. British regulations then in effect did not foresee vessels of this size. Accordingly, only 825 places were required in lifeboats, sufficient for a mere one-quarter of the *Titanic*'s capacity of 3547 passengers and crew. No extra precautions had seemed necessary for an unsinkable ship. The result: 1522 dead (drowned or frozen) out of the 2227 on board for the Titanic's first trip.[2]

The *Titanic* remains a haunting image of technological complacency. So many products of technology present some potential dangers that engineering should be regarded as an inherently risky activity. In order to underscore this fact and help to explore its ethical implications, we suggest that engineering should be viewed as an *experimental* process. It is not, of course, an experiment conducted solely in a laboratory under controlled conditions. Rather, it is an experiment on a social scale involving human subjects.

There are conjectures that the *Titanic* left England with a coal fire on board, that this made the captain rush the ship to New York, and that water entering the coal bunkers through the gash caused an explosion and greater damage to the compartments. Others maintain that embrittlement of the ship's steel hull in the icy waters caused a much larger crack than a collision would otherwise have produced. Shipbuilders have argued that carrying the watertight bulkheads up higher on such a big ship instead of allowing less obstructed space on the passenger decks for arranging cabins would have kept the ship afloat. However, what matters most is that the lack of lifeboats and the difficulty of launching those available from the listing ship prevented a safe exit for two-thirds of the persons on board, where a *safe exit* is a mechanism or procedure for escape from harm in the event a product fails.

1. The Social Context of Engineering & Technology

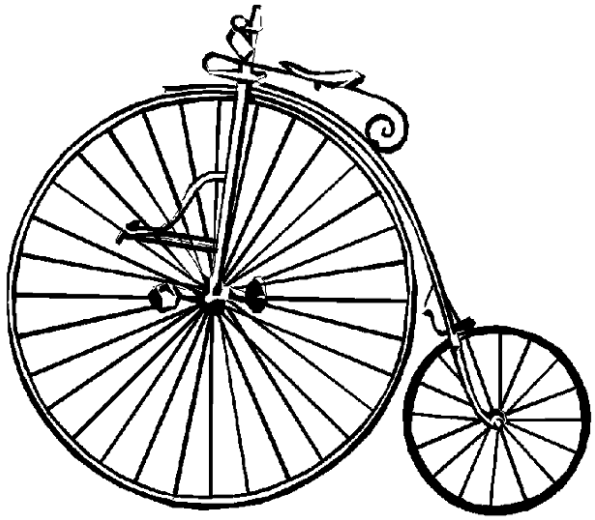
Engineering does not occur in a vacuum and technology is neither created nor maintained in a vacuum. Every individual engineer and every engineering firm or company is embedded in society. As such, both engineers as people and engineering as a profession are influenced by their society while also having the power to influence that society going forward. Consider, for instance, the lack of lifeboats on the *Titanic*. The engineers who signed off on the designs with such limited lifeboats did so, in part, because that is all society required of them via its regulations.

Not only do the law and official regulations influence engineering decisions, but so does public perception. Consider the idea of 'efficiency', a common term in engineering and technology. It is generally agreed that it is better to make an activity more efficient and, within the technical sciences, 'efficiency' is often understood in a purely quantitative way: it is a ratio of energy input to energy output. However, there are a variety of likely 'efficient' changes we could make (or previously had) which are nevertheless excluded from consideration: child

2. Lord, 1976; Wade, 1980; Davie, 1986.

labor was in many ways more efficient than adult labor and yet no one currently builds machinery that is reliant on a small child being able to fit into certain crevices.

Other examples abound. For instance, Trevor Pinch and Wiebe Bijker showed that social forces directed the development path of bicycles in their early history.[3] Early on, there were two types of bicycles: a sporting bicycle with a high front wheel that made it fast but unstable and a more "utilitarian" design with a smaller front wheel to promote stability at the expense of speed. Although originally designed for different purposes – the sporting bicycle for athletes and the other for ordinary transportation – the sporting bicycle never really caught on and eventually disappeared. Society, rather than the designers, determined that the sporting bicycle was unnecessary.



The "Penny-farthing" sporting bicycle

In the other direction, we can see the way technology has influenced and changed society, often in unexpected or unpredictable ways. There are obvious cases: speed bumps effectively force people to drive more slowly and walking paths and streetlamps encourage foot traffic to follow specified paths. But there are also less obvious cases: the invention of the printing press revolutionized European civilization in many ways, including being a major contributing factor in the Protestant Reformation, thus drastically changing the religious landscape of the entire world. More recent technologies like cell phones and social media have also heavily influenced social relationships by encouraging instantaneous and constant communication and altering what it means to call someone a "friend".

And then there are the more general ways technology often influences society: changing what we consider possible, required, or impermissible. Before advances in medical technologies, we would simply expect someone to die from cancer but now because it is possible to treat many cancers we find it morally problematic when those treatments are not available to someone. Technology also changes what we can expect from our lives: what sorts of interactions are possible, where it is possible to live and work, what sorts of jobs even exist.

In short, all of these examples come together to show that **engineering and technology function in a social context:** technology simultaneously exerts influence over society and is influenced by society. This, then, implies that technology professionals both exert influence over society through their designs and, in turn, are influenced by various social factors in the creation and deployment of their designs.

Objects make us, in the very same process that we make them.[4]

Engineering's influence on society is most evident in the context of **disruptive technologies:** technologies that significantly alter the way individuals, industries, businesses, or society at large operate. The internet and,

3. Trevor J. Pinch & Wiebe E. Bijker (1984). "The social construction of facts and artefacts: or How the sociology of science and the sociology of technology might benefit each other," Social Studies of Science 14.3: 399-441.

later, smartphones, are both prime examples of disruptive technologies that have substantially changed our social world. Looking ahead, we can predict that technologies like Generalized Artificial Intelligence, Virtual Reality, and perhaps Blockchain will all end up, if embraced by society, radically transforming the social world as well. But even less "innovative" technologies can be disruptive: autonomous vehicles are not that far removed from existing vehicles in many ways and yet we can imagine that a society which has embraced autonomous vehicles may look very different from our own. For instance, imagine the changes to peoples' daily lives and physical and mental health if traffic jams and collisions were a thing of the past!

Examples of the Social Context of Technology

- Although human and non-human animal cloning is technically feasible, it is widely opposed by (nearly) every society and thus has barely progressed
- We have substantial power to genetically modify crops but genetic modification is heavily opposed in certain areas of the world, such as Europe and some areas of Africa, thus stifling development
- Many towns and cities in the United States are designed around the personal automobile, indicating the impact personal automobiles have had on American society
- More and more of our social interactions are now technologically mediated: we speak via phones, use video chat, text message, play online video games with text and voice chat, etc.
- A significant portion of contemporary occupations only exist because of various enabling technologies such as computers and automobiles

2. Techno-Social Systems

There is a deeper way in which technology functions in a social context. Our previous examples of technology affecting society and society affecting technology largely treated society and technology as distinct even if interactive. But that is not quite an accurate picture, for in reality technology is *embedded* in society and it only functions as part of a system composed of both human/social elements and technological elements. Whether a piece of technology does what it was designed to do is not merely a matter of its technical design, but also how it ties into relevant social structures.

To put this another way, if we really want to understand the function of any piece of technology, we cannot simply examine the technological artifact in isolation. There is no sense to be made of what a smartphone does without reference to the broader social world in which it is embedded. This

We use the term **technological**

4. Daniel Miller (2010), *Stuff.* Polity Press.

artifact (or simply artifact) to refer to any object produced by human craft for some purpose. This can include basic products but also large-scale engineering projects.

includes the background conditions that make its functions possible, such as computer chips and cellular data towers. But it also includes relevant social conditions: people having a desire to communicate with each other or have instant access to distraction.

All this is to say that when we speak of technology, for the purposes of understanding it, we are really speaking of **techno-social systems**: the complex interactions between technological artifacts and aspects of the social world. Understanding this can open up new avenues for exploration and development: for it encourages us to pay attention to how actual people actually use things, rather than merely focusing on how we as technological designers may intend for things to be used.

3. Technological Mediation

We can deepen our thinking about techno-social systems and the social context of technology by exploring the **Theory of Technological Mediation**.[5] Technological mediation is a way of thinking about technology that aims to take technological artifacts seriously by putting our focus on what these artifacts *do*. This can sound like an odd approach, for most of us our used to purely *instrumental* ways of thinking about technology. On these instrumentalist approaches, technological artifacts do not *do* anything. Rather, they are simply passive tools (objects) used by humans (subjects) to achieve human ends. Put another way, the instrumental approach regards technological artifacts as dead matter upon and through which humans can exercise their will.

Mediation theory, however, orients us away from this instrumentalist view of technology as passive. Instead, it holds that technological artifacts play an active role in our lives by always mediating the way we engage with the world around us. And while this may seem uncontroversial when it comes to technological artifacts that seem to ‘act’, like robots or digital avatars, mediation theory suggests that this is true of *all* technological artifacts. Even simple artifacts like hammers, glasses, or the fountain pen mediate our lives.

Technological artifacts mediate two important aspects of our existence in the world: our **perception** of the world, and our **actions** in it. In so doing, technological artifacts change who and what we are.[6]

5. This section is inspired by, and portions of it taken from, Dr. Jan Peter Bergen's "Technological Mediation and Ethics", published by the 4TU Centre for Ethics and Technology under a CC-A-SA license.
6. In mediation theory, this is an aspect of the "co-constitution of humans and artifacts". Without human subjects, there would be no artifacts. However, without technological artifacts, human subjects as we know them would likewise not exist. There is an ongoing co-shaping of humans and technology.

Technology as mediating our being-in-the-world (from Hauser et al., 2018)

When technology mediates our perception of the world, it may amplify or reduce certain aspects of the world to be experienced. When technology mediates our actions in the world, its design or implementation is such that it is inviting, discouraging, or inhibiting certain actions. We can see this at play with a simple technological artifact: the hammer. The hammer invites certain actions like using it to drive nails (rather than attempting to press them with your bare hands) while discouraging others such as fastening two boards with glue.[7] The simple hammer also changes our perception of the world. There is, of course, the old adage that "when all you have is a hammer, everything looks like a nail". But more precisely, we can notice that our focus is drawn to the head of the nail, searching out a proper contact point between hammer and nail as much of the rest of the world fades out of view. When holding a hammer, the world becomes more "hammerable" than it was before: certain features stand out as more or less inviting to my hammer use. In both subtle and less subtle ways, engaging with the world in a way mediated by the hammer changes that engagement in profound ways.

Ethics is traditionally concerned with what humans, or other moral agents, *do*. In the context of technology, this often involves asking how humans should or should not use technological artifacts. But if technology mediates our interaction with the world in the ways just discussed, then we should expand our ethical thinking beyond what humans do and ask: *what do technological artifacts do*? Asking this sort of question can open up two useful avenues for social and ethical reflection.

First, mediation theory can enhance our moral perception and imagination. It does this by helping us identify and describe morally salient aspects of human-technology relations that would otherwise remain hidden. Mediation theory encourages us to ask ethical questions about the way technology enhances and diminishes our perception of the world and the ways it invites and inhibits our actions in the world. Some technologies inhibit actions we want inhibited and direct our focus in helpful ways. But other technologies may encourage actions we would rather people avoid or discourage us from paying attention to morally important issues.

Second, mediation theory can enhance our moral reasoning abilities. Once we become aware of the ways that technology mediates our perception of and interaction with the world, we are now in a position to take seriously our responsibilities to design, implement, and interact with technology with that mediation in mind. Thus, once we realize that technology *can* be used to discourage or even outright forbid certain actions, we can now ask under what conditions technology *should* discourage or forbid such actions. Thus, now when we reason about how to design or implement technology, we are in position to pay closer attention to its design features rather than largely off-loading our ethical thinking to the mere *use* of the technology.

7. The hammer may invite certain other actions as well, such as the use of it as a paperweight. This connects to the idea of multistability, which indicates that specific artifacts invite a specific, varied, but limited set of actions, some of which the artifact was not specifically designed for.

In short, mediation theory suggests that it is the moral responsibility of technology designers to predict and anticipate the mediating effects of their designs and to design their technologies to mediate well.

Example of Technological Mediation

In *Moralizing Technology: Understanding and Designing the Morality of Things*, Peter-Paul Verbeek illustrates technological mediation by detailing how an obstetric ultrasound alters our moral perception and reasoning.

From a purely instrumental perspective (i.e., absent the benefits of mediation theory), we may simply describe an ultrasound as a tool for providing visual access to a fetus *in utero*. In this way it enhances our perception and can provide us with access to certain types of potentially relevant information, such as the fetus's health.

But once we consider the ultrasound with mediation theory, asking what an ultrasound *does* as a piece of technology (rather than just what do humans do with it as a tool), we gain new insight.

First, consider perception: an ultrasound casts the fetus as a distinct individual unborn baby. It does this, first, by encouraging us to focus on the display screen rather than the mother, thereby separating the fetus from the mother in our minds. Additionally, the ultrasound screen enhances the size of the fetus such that it appears to be much closer to the size of a newborn infant. This all suggests how an ultrasound changes our perception of the world (and, in particular, the fetus and its relation to the mother). But it also changes our moral perception: the existence of the ultrasound and its ability to detect features of the fetus encourages the further medicalization of pregnancy and establishes the womb as a site of surveillance. Finally, by establishing the fetus as an independent unborn baby, the ultrasound contributes to a shift in how we think about the responsibilities of (prospective) parents. For even before the child is born it is now an independent being that must be regarded wholly independently from the mother.

Second, consider reasoning: being able to see the biological sex of the fetus as well as potential defects now opens up new avenues of moral thinking that are simply eliminated without the technology. Decisions about whether to continue with the pregnancy can now be made on the basis of the biological sex of the child or the presence of birth defects. Similarly, in contributing to the further medicalization of pregnancy, parents now become responsible for their choice of birthing methods in a way that is not possible if there simply are no choices.

This mediation analysis of the ultrasound suggests that designers should be asking a variety of questions that are likely currently ignored:

- Could the visual presentation of the fetus be altered to better reflect scale or the fetus's interconnections with the mother?
- Could ultrasounds be designed in such a way that parents could use them at home,

rather than necessitating involvement in the hospital system?

4. Technological Mediation & The Control Dilemma

Technological mediation theory tells us that technology influences people and society. And keeping that in mind can allow us to potentially predict what those influences may be prior to the (wide-scale) release of the technology into society. This is certainly a key reason to be familiar with the theory. However, it would be too quick to believe that we always can, in fact, predict how technology will affect society (or how society will affect technology). And a key explanation for this was first developed by philosopher David Collingridge and is now known as **The Control Dilemma** (or sometimes *The Collingridge Dilemma* after David Collingridge):[8]

> **The Control Dilemma.** When technology is at an early stage of development we have the power to control it, but we don't know what its social impacts will be. When technology is at a late stage of development, we know what its social impacts are, but we lose the power to control it.

The Control Dilemma, like all dilemmas, establishes two paths ("di-"), both of which are problematic in some way. We want to be able to control the introduction of technology into society to limit negative effects, but we cannot really know the effects until we introduce the technology into society. But if we introduce the technology into society to figure out the effects, then we are not in as good of a position to control the technology and its effects.

Thus, even as technological mediation theory and general awareness of the social context of engineering and technology empower us to better direct the development of technology and its introduction into society, the control dilemma suggests our power will always be somewhat limited.

The classic control dilemma is predominantly focused on "hard impacts" of technology: effects on health, safety, and the environment, etc. But technological mediation theory suggests that technology can also have "soft impacts" – it can effect, over time, our social values and principles, thereby not just affecting what we already care about (as in the hard impacts) but also changing what we care about. As such, philosophers Olya Kudina & Peter-Paul Verbeek have recently argued that there is a second version of the control dilemma that emphasizes these "soft impacts":

> **Moral Control Dilemma.** "[W]hen we develop technologies on the basis of specific value frameworks, we do not know their social implications yet, but once we know these implications, the technologies might have already changed the value frameworks to evaluate these implications." (Kudina & Verbeek 2019, 293)

8. David Collingridge (1980). *The Social Control of Technology*. London: Frances Pinter.

In recent times, this moral control dilemma is perhaps best illustrated by the changing understanding of the value of privacy in light of recent digital technologies with substantial surveillance possibilities. Before the widespread existence of cell phones with cameras, most live music venues banned photos and videos. Now with the increasing use of video doorbells we are forced to ask what counts as a reasonable expectation of privacy on your own property. The increasing availability of consumer-level drones similarly raises questions about what privacy requires and why privacy is important. The upshot, then, is that although we may have started the development of these technologies with one framework for understanding privacy, by the time the technologies are created and introduced into society, they effectively have forced us to change our framework.

Refer back to the ultrasound example of technological mediation and you can similarly see how technology can force (or at least strongly encourage) us to change our moral frameworks. But, again, to reassert the dilemma: although we can know, in general, that technology may have this effect, it is unlikely that we can wholly anticipate the change or wholly control it.

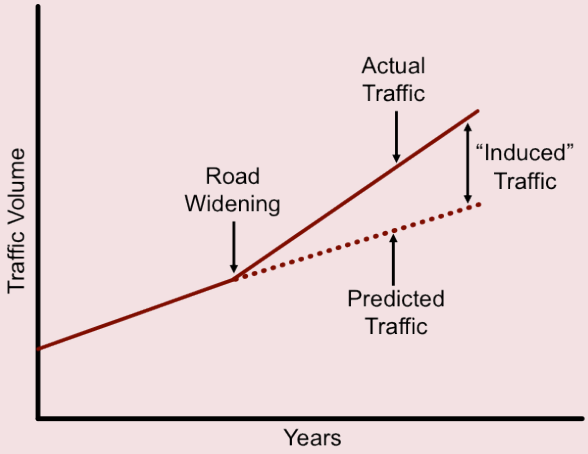
The Control Dilemma in Action: Rebound Effects & Induced Demand

Compact Fluorescent and LED lightbulbs were introduced largely as a means of reducing energy usage related to lighting. But, in fact, they have had the opposite effect, increasing light-related energy usage.[9] Similarly, the typical response (in the United States at least) to traffic congestion is to increase the available lanes of traffic. The expected response would be a reduction in congestion as vehicles now have more space to spread out. But, in reality, the typical result is the opposite: an *increase* in congestion.[10]

These are both examples of two related phenomena which show the importance of reflecting on implicit mediations and their related unintended outcomes. In the first case, we are dealing with the *Rebound Effect*: The fact that increases in efficiency of a technology will often not result in (as much of) a reduction of resource usage because the increased efficiency leads people to see usage as “more acceptable”. The person who buys a more energy efficient vehicle and tells themselves, as a result, “I can now take those long road trips!” is evidencing the rebound effect. Importantly, the rebound effect describes both a situation (like the lightbulbs) where the resulting effect actually makes things worse than they were before as well as a situation where the resulting effect simply reduces the hoped for benefits. Thus, the rebound effect can both produce negative mediation outcomes as well as simply reducing the positive mediation outcomes.

9. Joachim Schleich, et al. (2014). A Brighter Future? Quantifying the rebound effect in energy efficient lighting, *Energy Policy* 72: 35-42.
10. Todd Litman (2022). "Generated Traffic and Induced Travel: Implications for Transport Planning," *Institute of Transportation Engineers Journal* 71(4): 38-47.

An increase in vehicle congestion resulting from more traffic lanes is a type of rebound effect known as *Induced Demand*. The basic idea, in the context of traffic engineering, is that traffic congestion almost always maintains a particular equilibrium: traffic volume increases until congestion delays discourage additional peak-period trips. Adding additional lanes, therefore, simply provides an opportunity for increased traffic volume until that same equilibrium is reached. Thus, although the goal of expanding the roadway was a reduction in congestion the result is no such reduction, and potentially an increase. Although induced demand is most commonly discussed in the context of traffic engineering, the phenomenon can apply to any sector of the economy. In other contexts it often just describes a situation where demand for some good increases as cost of the good decreases.



Induced demand

Both the *rebound effect* and *induced demand* provide striking examples of unintended technological mediations. Additionally, because understanding both requires understanding social dynamics, economics, and human psychology, they show the importance of broad understanding of the humanities and social sciences to effective technological design.

5. Assessing and Designing Technological Mediations

Although the Control Dilemma should humble us in believing we could ever fully know the mediating effects of a technology, that does not mean we cannot become better able to anticipate mediating effects as well as design for desired mediating effects. To do that, however, we must develop and deploy our moral imaginations. We do this, partly, through developing a broad understanding of people and society via humanistic and social scientific inquiry. But we can also do this through knowing what questions to ask. Since the broad understanding of people and society goes well beyond the scope of this book, we will focus instead on developing a quasi-systematic "framework" for asking the right questions.

To craft this framework, we can divide mediation analysis into four elements:

1. *Intended Mediations:* The influences on perception and/or behavior that the designers explicitly intend and hope to design into the artifact
2. *Implicit Mediations:* The influences on perception and/or behavior that the artifact may unintentionally have due to its design and/or its use context

3. *Forms of Mediation:* The specific methods used to mediate perception and/or action
4. *Outcomes of Mediation:* The actions, decisions, and broader social changes that result from the mediations

Each of these four elements are of course related. To see how, consider the case of the speed bump. In designing and installing a speed bump, we *intend* to influence behavior, specifically by making vehicles slow down (an *intended mediation*). To do this, we use a technology that effectively *forces* the driver to slow down (or risk damaging their vehicle). This is the *form* of mediation (more on forms below). So we hope to force people to slow down, presumably with the broader outcome being that people slow down in important areas and therefore safety is enhanced. That is at least one description of the *outcome* of mediation. But, of course, whether it is indeed the outcome depends on the actual social context. Just because we *intend* to mediate behavior in a certain way to produce a certain outcome does not mean we will succeed. Or, even if we do succeed, it does not mean we will not also produce other outcomes as a result of our intended or implicit mediations.

So, to more fully develop our analysis, we would also want to think about potential *implicit mediations* as well as, to whatever degree possible, actually study the artifact (or similar artifacts) in the actual world to know actual outcomes. When it comes to implicit mediations, though, we can try to anticipate them by thinking slowly and deeply about how interaction with the artifact may influence perception or behavior. For instance, if we assume some people will not like to have to slow down due to the speed bump, then we can anticipate that a potential implicit mediating effect of our speed bump will be to divert traffic to another street (of an alternative exists). In this way, a speed bump *encourages* (some) drivers to alter their commute. The inevitable outcome of that, beyond merely increasing traffic on some other roadway, will be a matter of a variety of contextual factors and so would require some specific research. But, by engaging our moral imagination we were able to set ourselves up for that additional research. Perhaps the street that will see increasing traffic is well under capacity and so we are fine with creating an increase. But perhaps it is an area with a school and so the increase capacity would heighten risks for pedestrians. Determining this sort of thing is where the technical competencies of the relevant engineers or computer scientists become important.

We can reconstitute the above analysis into our 4 elements in the following way:

1. *Intended Mediation:* Decrease automobile speed on the street
2. *Implicit Mediation:* Shift more traffic to an alternative street
3. *Forms of Mediation:* Intended mediation is *coercive* while the implicit mediation is *persuasive*
4. *Outcomes of Mediation*: Although the precise outcomes we cannot know until later, the predicted outcomes are an increase in safety for all modes of transport on the street with the speed bump and an increase in traffic (requiring further study) on an alternative route

This is, of course, only a partial analysis. We may be intending other mediations with our speed bump and there will almost certainly be other implicit mediations and therefore other outcomes. But, this partial analysis should provide a useful example of how to engage in a **Mediation Analysis**. To further enhance our abilities, however,

we can examine each of the four elements in a bit more detail to identify common questions we may want to ask and common results we may want to be on the lookout for.

5.1. Intended & Implicit Mediations

Whether a mediation is intended or not, we will tend to ask many of the same questions in order to identify and understand it. Most broadly, we are aiming to answer the following sorts of questions:

- How will this technology alter the perceptions of those who use or interact with it? Where does it direct their attention during use or interaction? What does it direct them away from paying attention to during use or interaction?
- How will this technology alter the actions of those who use or interact with it? What sorts of actions does it make more likely? What sorts of actions does it make less likely?
- How will this technology alter the way individuals, groups of people, or society at large interpret and perceive their world? How might it change their understanding of themselves or their world through the changes in perception and/or action it generates in users?

It is important to notice that the first two sets of questions direct us to focus on those *interacting* with the technology. Mediation always begins with those in contact with the technology, even as it will typically end up having mediating effects even for those who never interact with it. The third set of questions are placed third for a reason, then, as they encourage us to reflect on how interactors will be mediated in order to think about broader changes that we may produce. It is important to focus first and foremost, and most heavily, on the mediating effects for interactors, since those tend to be the result of specific design decisions and are thus most directly open to change through re-design.

5.2. Forms of Mediation

In the speed bump example above we saw two different forms of mediation: coercion and persuasion. Broadly speaking, there are many different forms of mediation and so we cannot exhaustively list them all. Instead, we can identify three broad forms mediation may take. All forms of mediation function as “scripts” that make certain perceptions or actions more or less likely. In this way, they all attempt to generate certain types of perceptions or actions. However, it is worth noting that some forms are more “heavy handed” than others. Thus, in laying out the three general forms below, we will work from the “least heavy handed” to the “most heavy handed”. Thinking in this way is helpful since one major concern raised by the fact of technological mediation is that it reduces human freedom. This is certainly true, in much the same way that laws or social norms also reduce human freedom, but the more “heavy handed” forms of mediation are heavy handed precisely in the fact that they reduce freedom more (or more strongly reduce freedom).[11]

11. The tool and techniques described here are derived from the De-scription activity created by Jet Gipson and the Product Impact Tool created by Steven Dorrestijn & Wouter Eggink. Further inspiration is taken from Peter-Paul Verbeek's *Moralizing Technology: Understanding and Designing the Morality of Things*.

5.2.1. Guidance

Much of our technologically mediated world involves forms of **guidance**: we use images or other tools to indicate to people how to interact with a product or how to use the technological artifact to achieve their ends. Consider all the signage in a building: bright exit signs, labels on doors that say “push” or “pull”. Many signs include images, text, and braille to ensure people are guided in multiple ways.

Car handles are designed to guide a user’s hand to open the door. Textured floors guide people to avoid harm

When we use **Guidance Design**, our goal is simply to facilitate people in doing what they want to do. We are not encouraging them to do anything in particular, but rather making it easier for them to do what they already want to do. In some cases we are guiding interaction with the technological artifact itself: labeling a button ‘power’ or using a symbol that designates power simply guides the user who may want to power on the device. In other cases we are guiding interaction with the larger social world: textured walking strips at curbs help those who are sight-impaired navigate the world.

The most interesting forms of guidance mediation come when an artifact is designed in such a way as to guide usage without explicit signage. So, consider, that in lieu of putting “push” or “pull” on a door, the door is designed in such a way that for most people they immediately know whether it needs to be pushed or pulled to open. Car door handles, as illustrated above, function in this way: there is no sign telling you how to interact with the handle; instead, the handle is structured in such a way that you are guided to use it for its proper function.

As the least “heavy handed” form of mediation, guidance is generally considered the most open to failure. Each of you has probably had the experience of attempting to open a door in the way you thought appropriate only to find out it opens the other way. However, if “failure” is not a big deal (for instance, you just attempt to open the door the other way) then we may prefer guidance over more heavy handed approaches.

5.2.2. Persuasive Design

Sometimes our goal is not simply to help people do what they already want to do, but to actually *encourage* them to do something they otherwise would not do (or, at least, would be less likely to do). For instance, in the Netherlands some of the live speed checking signs on roads output a “sad face” when the driver is exceeding the speed limit. The goal, of course, is to encourage the driver to slow down.

Persuasive design engages peoples’ minds just like guidance design, but does it with more of a ‘push’. Whereas in guidance we assume the person already wants to do the thing we are guiding them to do with persuasion our

default assumption is the person would not otherwise do the thing we are now persuading them to do. Of course, that doesn't mean everyone who interacts with the technology actually needs persuasion, but in general we assume people will need some sort of 'push' to engage in and continue the behavior. Nonetheless, it is important to note that persuasive design does not *force* any sort of behavior, it simply aims to make it more likely.

The "sad face" speed checker. And a staircase painted as a piano to encourage use

Persuasive design is sometimes split into two sub-categories: Persuasive design and Seductive design. The relevant distinction here is whether the design attempts to engage the user as a rational person to encourage them to choose to engage in the desired action or whether it engages them non-cognitively as a means of "seducing" them into doing the desired action without any real reflective choice. We can see the difference when thinking about two different approaches to reducing automobile speed. The "sad face" speed checker, illustrated above, is an example of persuasive design. For it to be effective at all the driver must pay attention to it and then process the information it is providing. An alternative approach would be to design the street in such a way that it simply makes driving at the desired speed the most attractive option. For instance, we may add curves to the road, narrow it, or add a tree lawn. Each of these has known behavioral effects on most people: it makes driving slower the more attractive option even as no driver is asked to "decide" to drive slower. As a side-note, to complete the example, if we simply install a sign indicating the desired speed limit, we would be using guidance, largely relying on people to already want to drive the "safe speed" (of course we are assuming that posted speed limits represent the "safe speed", which is dubious).

For our purposes, we will just include seductive design under the heading of persuasive design. Nonetheless, it is worth keeping in mind the difference. Some would suggest that seductive design is "more heavy handed" than persuasive design, since it doesn't attempt to engage our reasoning abilities. However, both forms of design are about *encouraging* certain perceptions or actions, even if their methods are a bit different.

Persuasive design (including seductive design) can be used for good or ill. Some companies have embraced persuasive design to "persuade" users to use devices more: smartphone "addiction" is a real thing. Often these uses of persuasive design aimed at making people engage with the technology more are dubbed "addictive design" to emphasize the nefarious aim of encouraging people to become 'addicted' to the technology.

But persuasive design need not focus on persuading a person to simply interact with the technology more. Instead, as the earlier examples indicate, persuasive design can be aimed at helping people develop good habits like regularly engaging in physical activity or turning off electronics at a certain point in the night.

5.2.3. Coercive Design

In some cases we need to effectively guarantee people will engage in certain behaviors. Although persuasive design makes it more likely than guidance, persuasive design still leaves open the possibility that people do not do what we are wanting them to do. In these cases, we may turn to **coercive design**.

Speed bumps force a car to slow down. Heavy machinery often requires two hands to operate

If it is essential that drivers slow down, a "sad face" speed checker is not going to do the job. Instead, we may install speed bumps as they force a driver to slow down (lest they have no care about their vehicle). Or, if we are designing dangerous machinery and want to ensure it cannot be operated without someone actively present, we might require someone to always have their foot on a pedal and hand on a button for the machinery to work. We could use persuasive design to achieve the same goal, but it is likely a few people would lose fingers and toes in the process. Coercive design, on the other hand, effectively eliminates any thinking or choice by the person interacting with the technology.

To put things more broadly, coercive design involves design features aimed at requiring or eliminating certain behaviors or perceptions. The examples above were all about requiring certain behaviors, but we can also see coercion at play in eliminating certain behaviors (at least among certain people): some of the overpass bridges constructed on Long Island in New York were intentionally designed to be too low for city busses to pass through. This was done by the designer as a means of keeping low-income people away from Long Island beaches. Similar things occur today in the use of *Hostile Architecture*, such as park benches with a bar in the middle that prevents anyone (typically the homeless) from sleeping on the bench.

We should be clear that coercive design need not always necessarily guarantee the desired action. Instead, it may effectively guarantee it by providing some sort of negative feedback. This is, of course, how speed bumps work. Strictly speaking, someone could go over them without slowing down, but we still consider the design coercive since the result of not slowing down is likely to be damage to the vehicle. This is in contrast to persuasive design which may provide some sort of *positive* feedback for compliance but does not involve any sort of damaging or harmful negative feedback for non-compliance.

5.3. Outcomes of Mediation

Now that we know some of the core questions to ask to determine what sorts of mediations may occur, and

we have a taxonomy of different types of mediations, the final step of our analysis would be in identifying the predicted and actual outcomes of the intended and implicit mediations. To do that, we can ask the following sorts of questions:

- If the mediations work as predicted, what are the short- and long-term effects on both those who interact with the artifact and society more broadly?
- Given other things we may know about the use context, are there any undesirable outcomes we can predict as a result of our mediations?
- What values are we promoting or protecting through our mediations? What values are we diminishing or frustrating?
- Is our *form* of mediation proportional to our expected or actual outcome? Or should we use a less "heavy handed" form of mediation?

To put these questions in context, consider a previous example of coercive design: the requirement that a person simultaneously have their foot on a pedal and hand on a button for a dangerous piece of heavy machinery to function. There is likely to be some implicit mediations from this, but we'll focus on the intended one: requiring someone to be present and actively involved in the functioning of the machinery. If that works as predicted, we can expect the short and long-term effects to be increased workplace safety, thus promoting the value of (human) health. We may also predict some undesirable outcomes in the form of "cheating" the system by perhaps putting heavy objects on the pedal and button so it can function unattended. In this case, that would just cut against our improvements in human health, but is unlikely to eliminate those benefits. We are, however, frustrating a value like individual freedom by forcing a person to interact with the machinery in a particular way. Given all this, is coercive design justified? Or ought perhaps re-design the machinery using guidance or persuasion as the means of promoting safety? The answer to that question may be complex, but to begin answering it we would want to be thinking about the trade-offs between the value(s) promoted and the value(s) frustrated, as well as the same trade-off given alternative designs. If guidance or persuasive design would significantly reduce the benefits to human health – they would increase workplace injuries – then coercion may be justified. But this is partly because the value being promoted is human health, a value that is very important and (basically) universal. If we were instead promoting some other value of less importance, then the cost of individual freedom may not be worth it.

At this point we are moving into the realm of **Mediation Evaluation**, but we do not yet have (all the) tools necessary to evaluate mediations fully. Those will come later, but for now we can emphasize that by thinking in terms of the *four elements* of mediation and asking the appropriate questions or applying the appropriate taxonomy we can produce a reasonably comprehensive **Mediation Analysis**. This analysis does not, on its own, fully work out what we ought to do, but it is essential to doing that. And by splitting the analysis off from the evaluation, we are more likely to produce a robust analysis and therefore (later) a better evaluation.

Check Your Understanding

After successfully completing this chapter, you should be able to answer all the following questions:

- What does it mean for engineering and technology to function in a social context? What are some examples of engineering/technology functioning in a social context?
- What are disruptive technologies and how do they illustrate the idea that engineering and technology function in a social context?
- What is a technological artifact? How do they relate to techno-social systems?
- What does Technological Mediation Theory tell us about technological artifacts?
- What is The Moral Control Dilemma? How does it relate to the standard Control Dilemma?
- What is Induced Demand? How does it relate to the Rebound Effect? And how to both relate to Technological Mediation
- What are the four elements of a Mediation Analysis? What sorts of questions might we ask for each element?
- What are the three main forms of mediation? Construct your own example of each

References & Further Reading

Collingridge, D. (1980). *The Social Control of Technology*. Continuum International Publishing.

de Boer, B., Hoek, J., & Kudina, O. (2018). "Can the technological mediation approach improve technology assessment? A critical view from 'within'," *Journal of Responsible Innovation* 5(3): 299-315.

Dorrestijn, S. (2017). "The Care of our Hybrid Selves: Ethics in Times of Technical Mediation," *Foundations of Science* 22(2): 311-321.

Hauser, S., Oogjes, D., Wakkary, R., & Verbeek, P.P. (2018). "An annotated portfolio on doing postphenomenology through research products," *DIS 2018 – Proceedings of the 2018 Designing Interactive Systems Conference*, 459-472.

Kudina, O. & Verbeek, P.-P. (2018). "Ethics from Within: Google Glass, the Collingridge Dilemma, and the Mediated Value of Privacy," *Science, Technology, & Human Values* 44(2): 291-314.

Verbeek, P.-P. (2005). *What Things Do: Philosophical Reflections on Technology, Agency and Design*. The Pennsylvania State University Press.

Verbeek, P.-P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. The University of Chicago Press.

Winner, L. (1980). “Do Artifacts have Politics?” *Daedalus* 109(1): 121-136.

4.

Designing for Values

Key Themes & Ideas

- Technology is always value-laden, resulting from value judgments and affecting values in the world
- Value Sensitive Design is an approach to design that aims to explicitly consider values early and throughout the design process
- In the context of technological design, anyone affected by a technology is a stakeholder
- A Stakeholder Analysis is a Value Sensitive Design technique used to identify stakeholders, values, and arguments related to a technology
- A Value Hierarchy is a Value Sensitive Design technique that can help us incorporate values into our designs, by moving from abstract values to concrete design requirements
- Design Principles are general requirements for a design project that must be met, but can typically be met in more than one way
- Design Requirements are specific requirements for a design project that typically indicate exactly how to design some feature of the project

Previously we saw that technology *mediates* our lives by influencing our perceptions of the world and our actions in the world. More broadly, because of these mediating effects, technology influences society in myriad ways. Once we understand this, it becomes clear that technological artifacts are not mere instruments for human will; they shape human will as well.

But there is further reason to reject the idea that technology are mere instruments. Or, to put the issue in slightly different terms: technologies are not **value neutral**. The instrumentalist view of technology holds that

technologies have no values or orientation toward values "built" into them; technologies do not affect values all on their own. Instead, on the instrumentalist view, technologies only affect values once people start using them for specific purposes, based on whatever value considerations those users have. Although it now serves a variety of social and political purposes in the United States, the phrase "guns don't kill people, people kill people" largely reflects the view that technologies are value neutral. Part of the claim being made when someone uses that phrase is that guns, on their own, do not alter personal or social values and that guns don't encourage or discourage certain types of perceptions, beliefs, and actions.

In contrast to this instrumentalist idea of value neutrality is the claim that technology is always **value-laden**: soaked through with value considerations. There are always values built into technologies and technologies always exert some influence over human thought and action and thus influence what they take to be valuable. We have already seen the outcomes of this in our discussion of technological mediation. But now we can backup a bit and think about how the technological design process itself ensures all technology is value-laden.

All technologies are the products of human minds and actions. They are the result of numerous decisions, big and small, by the people crafting them. And all decisions are made on the basis of *values*, *outlooks*, *perspectives*, *intentions*, and *desires*. In this way, it is impossible to truly and fully divorce a technology from the humans that designed it and its social context. The design process necessarily involves establishing a (loose and often implicit) hierarchy of value considerations. Take a simple consumer electronic device: are we going to prioritize accessibility and therefore sacrifice quality in order to reduce cost? Or are we going "high-end", knowing we will be designing a product that is out of reach for many people? Whatever answer we give, we are now identifying which values we care about more than other values. And whatever design decisions we make after that will be influenced by those value preferences.

So, building technology is all about design choices, these choices are based in (often implicit) value preferences, and these choices impact values in the world. We rarely make technology just for the sake of it; we make technology because we expect to deploy it out into the world, where it will interact with all different kinds of people and be used in many different contexts. Therefore, technology will be *better* when it is designed in a way that is thoughtful about the values it embodies and is informed by the social, institutional, and cultural contexts in which it is to be used. Technologies that are designed in a *value sensitive* way are more intuitive, more accessible, more seamless, and more delightful than those that are not. They are also more likely to do good: promote human flourishing, generate social benefits, and contribute to environmental sustainability. They are more likely to work effectively, and they are more likely to be successful.

In short, by engaging in **Value Sensitive Design** technology professionals can better meet their professional responsibilities, design better technologies, and contribute to positive social change.

Examples of Values in Technology

There is a constant stream of news stories about controversial, ethically-problematic technologies. Below are a few, brief examples. These cases illustrate the importance of designing technologies in a value-informed way that includes careful consideration of the social and institutional contexts of deployment and use:

- Highways have the potential to help people travel faster and increase access to certain areas. However, their construction often alters the social landscape by dividing neighborhoods, increasing noise pollution, and altering traffic patterns in ways that can both increase and decrease business in nearby areas.
- Facial recognition systems have the potential to make our lives more convenient through automatic face tagging on social media, facial unlock on smartphones, etc. but also have profound implications for individual privacy.
- Smart speakers are extremely popular, but users were upset to learn that transcripts of their audio were secretly being reviewed by human beings. The key issues were (1) that users were not informed about this practice, and (2) audio recordings from within our homes are considered by most people to be sensitive data.
- Sharing economy apps that let anyone rent out their car or home are very convenient for those with resources to offer and those looking to rent, but they are also having negative impacts on cities in the form of increased roadway congestion and displacement of local residents.
- You pay for ‘free’ online services by divulging your personal data, which is collected by hundreds of advertising companies. People are often distressed at the extent to which this data can be used to hyper-target advertising, as best evidenced by the Cambridge Analytica scandal.
- The impending rollout of self-driving cars raises challenging questions about how these cars should be designed to ensure human safety, and how these vehicles, their makers, and their operators should be held accountable when accidents occur. Even more challenging ethical questions arise when we consider the development of autonomous military drones.
- Many organizations are adopting machine learning systems in an effort to reduce costs and remove human bias from processes. These systems evaluate whether people are eligible for loans, insurance, employment, social services, and even parole. However, machine learning systems are not neutral, and these systems have been found to exhibit human biases like racism and sexism.
- User interfaces are powerful mechanisms for shaping how users interact with systems. However, some designers adopt intentionally deceptive user interfaces called “dark patterns”.

1. What is Value Sensitive Design

We know that technologies are value-laden and reflect value judgments. This is true whether designers explicitly consider values or not. But, once we know that technologies will be influenced by values and will influence values, it would be irresponsible to not explicitly consider values in the design process. Just like once we know that technology has mediating effects, we ought to deliberately account for them in our design thinking, once we know that technology is value laden we ought to deliberately consider values in our design thinking. And that is precisely what Value Sensitive Design is all about.[1]

Value Sensitive Design (or VSD) is an approach to identifying and grappling with value-laden design decisions. Originally developed by the computer science professor Batya Friedman, VSD has been widely used across technology disciplines, from civil engineering to surveillance to human-robot interactions.

The central goal of VSD is to help technology professionals make socially-informed and thoughtful value-based choices in the technology design process. At a high level, VSD helps us to:

1. Appreciate that technology design is a value-laden practice
2. Recognize the value-relevant choice points in the design process
3. Identify and analyze the values at issue in particular design choices
4. Reflect on those values and how they can or should inform technology design

Batya Friedman, pioneer of Value Sensitive Design

VSD is, in effect, an *outlook* for seeing the values in technology design and a *process* for making value-based choices within design. It encourages us to account for a variety of often conflicting values and concerns. Additionally, it accepts that there are rarely easy answers to moral questions and so does not provide an algorithmic approach to resolution. Instead, it provides a variety of tools and questions that form the process and which, when done conscientiously and through engagement with others, can often lead to improved technological design.

There are far too many tools and techniques under the banner of Value Sensitive Design to cover them all. So, instead, we will focus on two that can be useful in nearly every design context and that have the added benefit of being useful tools for learning about ethics in general.

2. The Stakeholder Analysis

According to the Association of Computing Machinery's Code of Ethics and Professional Conduct, "all people

1. Much of what is said here is derived from or inspired by Christo Wilson's VSD@Khoury website. https://vsd.ccs.neu.edu

are stakeholders in computing."[2] We can reasonably extend this idea to say that *all people are stakeholders in technology*. By this we mean that everyone is affected, to some degree, by technologies. This is obvious for those who engage with the technologies, but it is also true for those who may never engage with them. Consider, for instance, the massive piles of e-waste in Western Africa. Or, similarly, consider that air pollution cannot be contained to the areas in which it was produced and so anyone could be negatively affected by it, no matter where they live.

The idea of a stakeholder is familiar to many business contexts, where it typically is focused on those entities that have a financial stake in the company or project. But in Value Sensitive Design, picking up on the technology professional's obligation to the welfare of the public, a **stakeholder** is anyone that may be affected, directly or indirectly, by the technology or technological project under consideration. It is also common to include non-human animals and environmental systems as potential stakeholders. Whether we expand stakeholders in this way or not, however, the key is that for our purposes stakeholders always fit broadly in the category of "the public". We do not, in contrast to the business context, care about the company designing the technology or investors in it. Our aim is to design the best product or project for the public, and so our stakeholders are limited to the public.

One of the core tools of VSD is the **Stakeholder Analysis**. This sort of analysis is typically done very early in the design process – perhaps before any design decisions have even been made – but is continually re-examined and updated as the process continues. The core function of a stakeholder analysis is to provide a thorough understanding of who may be affected (positively or negatively) by the technology as well as the values that are affected, or perceived to be affected, by the technology. Put another way, a stakeholder analysis is a tool for designers to "get out of their own head" in order to more comprehensively understand "what is at stake" with a technology. Any given designer may already be able to identify some relevant values for the project, but they may also miss some, understand some differently from members of the public, or (rightly) believe some values aren't relevant that some members of the public nonetheless do believe are relevant. And it matters that they believe a technology may affect a value, even if it won't. For instance, if people wrongly believe that some new technology will negatively impact their privacy, they may not adopt it. Knowing some stakeholders see it this way can help designers innovate in ways that eliminate the (false) view and therefore increase the success of the technology.

Another way of understanding the aim of the stakeholder analysis is to see it as a tool for mapping the "argumentative terrain" of a technology. We are trying to take account of all the possible arguments in favor and against designing the technology and designing it in particular ways. Sometimes we do this by noting a particular group that may be negatively affected by the technology. That the group would be negatively affected may count against designing the technology in a particular way or may count in favor of designing it in some other way. Other times we may do this by identifying a key value that will be impacted by the design. Once we know that value, we can start to ask how we can design the technology to promote or protect it or to limit negative effects on it.

Importantly, exactly how we divide up "stakeholders" will depend on the context. In all cases, our goal isn't to identify specific individuals – Nancy, Nikea, Bill, etc. – but rather to identify the various ways people may relate

2. Association of Computing Machinery (2018). "ACM Code of Ethics and Professional Conduct," https://www.acm.org/code-of-ethics

to the technology. For instance, if we are conducting a stakeholder analysis for a consumer electronic product we may distinguish between "Power users" and "casual users". People falling into these two groups will have different preferences about the device; power users would benefit from greater customization features while casual users would prefer ease of use even at the expense of capabilities. We might also distinguish between "early adopters" and "late adopters" of the product, if we can identify differences between their preferences or the values they associate with the technology. Importantly, too, notice that the very same individual could be both a "casual user" and an "early adopter". This is a major reason we focus on these categories (or *roles* as they are often called) rather than named individuals; most people will fall into multiple stakeholder categories.

The 4 stakeholder categories mentioned above all fall under the umbrella of **direct stakeholders**. These are stakeholder groups that will typically be using or interacting with the technology. For many (but not all) technologies "direct stakeholder" effectively means "user". But most technologies also affect people who will never themselves interact with the technology (or will only do so minimally or incidentally). Many large-scale public works projects are like this. Very few people are direct stakeholders for a powerplant. But a significantly number of people are affected (or potentially affected) by the powerplant. We could, for instance, identify "nearby residents" as a stakeholder category relevant to the design of the powerplant. The powerplant would likely affect the *health* of these people, which is an important value. These sorts of stakeholder categories are typically called **indirect stakeholders**.

It is important to keep in mind that the distinction between direct and indirect stakeholders is *not* about the degree to which they are affected by a technology. The impact of a technology on indirect stakeholders can be just as great, if not greater, than its impact on direct stakeholders. The distinction is about the *pathway* of impact: does it come from interaction with the technology or does it come from the *others*' use of the technology or, more generally, the mere existence of the technology.

Our central goal in conducting a stakeholder analysis is to produce a reasonably comprehensive picture of all the various ways people may be affected by the technology and how they will be affected. But, of course, there may be an infinite number of potential stakeholder roles relevant to a technology, and even if we could identify them all our list would be so long as to be useless. Thus, there are a few general considerations to keep in mind when conducting a stakeholder analysis that can help increase the likelihood of capturing the important stakeholders and values while not getting bogged down.

First, our goal in identifying a stakeholder group is to recognize the unique way(s) they may be affected by the technology. We are not identifying groups just for the sake of it. So, for instance, if we are unable to find any meaningful difference between the values and preferences of "early adopters" and "late adopters" of the technology then we do not need to identify those stakeholder groups at all. Keeping this in mind can help reduce the size of our list.

Second, if we keep in mind that a stakeholder analysis is a means of laying out the "debate landscape" for the technology – identifying the various arguments in favor and against particular ways of designing it – then that can further reduce the inevitable length of our list. We do care about the stakeholders in their own right, but we are also identifying them as a means of identifying relevant values. I distinguish "power users" from "casual users"

in large part because those groups care about different values (or, more precisely, prioritize different values) when it comes to our technology. And so, although the typical stakeholder analysis process moves from stakeholder to value and argument, we can also work backwards. We might, first, identify a value or an argument relevant to our design context and then try to identify a stakeholder group that the value or argument fits with.

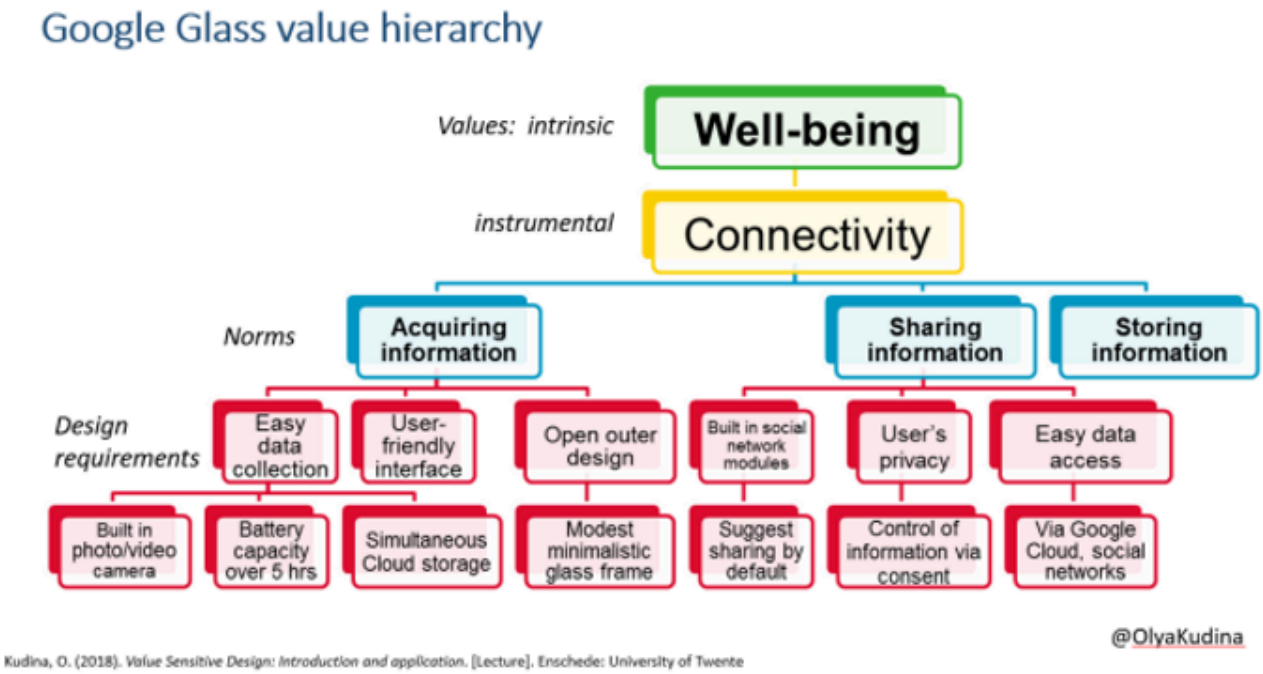
Finally, it is worth noticing the various means by which we may go about identifying stakeholders, values, and arguments. In some cases, we may bring together a "focus group" of individuals who represent various stakeholder groups. Or, similarly, we may talk to various individuals to just get their view. But we may also consult studies and argumentative essays (or videos). Existing attitude surveys can be a useful tool for identifying general perspectives on a technology and argumentative essays can be vital for fully filling out how values or stakeholder groups will be affected and what that means for the design of the technology.

In sum, a stakeholder analysis is a vital tool for taking seriously the responsibility to the welfare of the public. It prompts us to think broadly about the effects of a technology, especially when we consider indirect stakeholders. And it also helps us identify values that may be unintentionally affected by the technology. In these ways, and more, it functions as a tool of responsible technological design and a resource for innovation.

3. The Value Hierarchy

A stakeholder analysis helps us identify relevant values. And, beyond that, we will hopefully already have an idea of what values we *want* the technology to promote or protect. But VSD is not just about initial value identification, it is also about taking values seriously *in the design process*. One valuable technique for doing that is the Value Hierarchy.

A **Value Hierarchy** is a technique we can use to systematically *implement* a value into our designs. It is a means of *operationalizing* values so that they become the sort of thing that a product or project can implement. And they typically result in a visual hierarchy construction, like the one below for Google Glass.



The example hierarchy for Google Glass illustrates the 4 main components of a value hierarchy:

1. **Intrinsic/Overarching Value:** Represented by "well-being" in the example, this top of our hierarchy is where we identify the core value we are trying to design for. The idea of an "intrinsic" value is that it is something we care about *for its own sake*, as opposed to something we care about as a means to an intrinsic value.
2. **Instrumental Value(s):** Represented by "Connectivity" in the example, this section of our hierarchy identifies one or more values that we care about because they can (in the context) help us protect or promote our intrinsic value. Strictly speaking, this level of the hierarchy is not always necessary, but is often helpful.
3. **Design Principles:** Called "norms" in the example and represented by the items in the blue boxes, design principles broadly highlight the sorts of things our technology should be able to do or not do without specifying precisely how they will be designed to do or not do those things.
4. **Design Requirements:** The most "concrete" level of the hierarchy, represented by the many red boxes in the example, design requirements are the (relatively) precise design features of our technology. They provide the sort of information that we can directly use in designing our product or project.

This tells us, in brief, what a value hierarchy contains, but we will go into more detail for each of these components below. Before that, however, it is worth noting that there is one sense in which all technological designers should already be familiar with value hierarchies. As noted in an earlier chapter, it is often suggested that engineers "have a special concern for safety". Thus, engineers are always expected to build a value hierarchy with safety as the value. Of course, this is just typically done without explicitly constructing the value hierarchy. Nonetheless, we could place safety as a value – likely an instrumental value with something like well-being or health as the intrinsic value. In standard engineering thinking, safety is often further specified into context-specific

design principles. For instance, when constructing something like a bridge we must consider the "factor of safety", which is effectively a design principle. It says something like "the bridge should be able to hold five times its expected actual load". Of course, as any engineer familiar with bridge construction could tell you, there are several different ways to construct a bridge to meet that safety factor. Thus, that the bridge must be designed to a 5x safety factor does not immediately give us any concrete design requirements. But it does tell us that we will need to identify design requirements that will implement that design principle.

And so, in a certain sense, the value hierarchy is no different from standard technical design thinking. What is different, however, is that whereas "designing for safety" is the foundational non-negotiable of all engineering, a value hierarchy helps us focus on how to design for other values that are not foundational in the same way and are typically (at least when it comes to the instrumental values) specific to our design context.

To more fully understand how to conduct a value hierarchy, as well as its value, we can turn to examine the components of the hierarchy in more detail.

3.1. Values

The first chapter of this textbook introduces you to the concept of a value. So, we will not cover that again. However, given the centrality of values to a value hierarchy and the fact that values are divided into two different levels in the value hierarchy, it is worth investigating their role in this technique in a bit more detail.

There are some things that we care about for their own sake and other things we care about because they are a means to something else. Well-being is the quintessential **intrinsic value**, something we care about for its own sake. While something like wealth is a prime example of an **instrumental value**. Having money is not valuable in its own right (although some people certainly act as if it is!). Rather, if being wealthy is valuable at all, it is valuable because it can help us protect or promote other values, like well-being. There are, as well, some values which are both intrinsic and instrumental. Health and Education are prime examples. It is good in itself to be healthier and more educated, but also being healthier and more educated is a means to many other good things in life.

There is no easy way to distinguish between instrumental and intrinsic values. People will disagree over some cases. But, in general, when it comes to instrumental values we can ask the question "and why does that matter?" and there should be a good answer (although we may not be able to immediately give it). When it comes to intrinsic values, however, there is no real answer to that question. In response to the question "why does well-being matter?" the appropriate answer just seems to be "because it does" and anyone who says otherwise is either being obstinate or living in a very different world from us.

The good news is that, for our purposes, we do not need to worry too much about the distinction between intrinsic and instrumental values. Not all value hierarchies contain both. Instead, for our purposes, we can often simply focus on what would otherwise be instrumental values. In the Google Glass case, we are interested in designing for "connectivity". That is certainly not an intrinsic value; if it matters at all, it matters because it can promote

well-being (as well as perhaps other intrinsic values). But, in the context of the value hierarchy, connectivity is our focus: we derive our design principles from thinking about connectivity, not well-being (at least not directly).

Thus, just as standard technical design involves thinking about the (instrumental) value of safety and identifying design principles related to it, value sensitive design involves thinking about instrumental values and identifying the design principles related to them. We can, however, keep the idea of intrinsic values in the back of our mind to ensure that what we are identifying as an instrumental value is valuable at all.

3.2. Design Principles

Design principles are perhaps the most difficult component of a value hierarchy. They need to be simultaneously more precise and context-specific than the value(s) we are designing for and less precise and context-specific than design requirements. They are an intermediate category that, broadly, captures a goal we are trying to achieve. Using our designing for safety example, our safety factor design principle can be restated as the goal of achieving a design that can withstand 5x the expected actual weight. Unfortunately, the examples in the Google Glass hierarchy are not the most illustrative, but they can be restated in a better way. To promote the value of connectivity, we should design our product to be able to acquire information, to be able to share information, and to be able to store information.

Another example of a value hierarchy focuses on the design of aviaries for chickens. It identifies "animal welfare" as the value (without worrying about whether it is intrinsic or instrumental) and then provides design principles like "hens should have sufficient living space", "hens should be able to lay their eggs in laying nests", and "hens should have the freedom to 'scratch' and to take 'dustbaths'".[3] These sorts of design principles are better examples of what design principles can be, laying out general requirements for the design of the aviary without demanding precisely how the aviary be built to meet those general requirements. For instance, while "having sufficient living space" does specify a *minimum* space requirement it does not specify a maximum. Thus, aviaries could be designed in various sizes and shapes while meeting the principle.

The example of "sufficient living space" highlights another common feature of design principles that distinguishes them from design requirements. In the context of the aviaries, the minimum requirement for "sufficient" was specified by the relevant EU law that imposed these design principles. But it did need to be specified, and it could have been specified differently. Design principles often (but not always) include terms or are otherwise phrased in such a way that they need interpretation and specification. This is what makes them principles rather than (for instance) rules or requirements. Importantly, of course, while design principles (or parts of them) may be open to some interpretation and specification, that doesn't mean anything goes. Instead, they typically prompt us to conduct appropriate research in the appropriate field. For the aviary, this meant digging into research on animal welfare and cage sizes for chickens. Similarly, we could have a technical design principle that simply says "must meet or exceed the safety factor" without specifying what the safety factor is. In doing that, we are simply prompting the relevant designers to look at the appropriate laws, building codes, and/or engineering standards to make the determination.

3. Ibo van de Poel (2013). "Translating Values into Design Requirements," in *Philosophy and Engineering: Reflections on Practice, Principles and Process*, edited by D.P. Michelfelder et al.

As should be clear by now, specifying values into design principles is a complex process. It requires technical knowledge, social knowledge, and conceptual knowledge. But there are a few things to keep in mind that can help with the process:

- We should always have a relatively precise understanding of the value we are working to specify before we start to craft design principles. Thus, we should not move from "connectivity" to design principles without first clarifying what we mean by "connectivity".
- Design principles are *prescriptive*: they tell us what we should do (or not do) in order to protect or promote the relevant value(s). As such, it can often be helpful to start all principles with something like "the product/project should…" or "the product/project should not…". Principles must provide direction, even as they should not specify precise requirements.
- A design principle should be an appropriate response to the relevant value and it should, perhaps in concert with other design principles, constitute a sufficient response to the relevant value. A sufficient response means we fully responding to the demands of the value, rather than only responding to parts of it.

3.3. Design Requirements

Design principles are much more immediately useful in the technical design process than values. Values can be complex, nuanced, and difficult to apply to design decisions. Design principles, on the other hand, give us direction to our designs. Thus, once we have our design principles, we can move into the iterative design process with those principles in mind. To do that, however, we need to begin translating our principles into design requirements.

As previously noted, design requirements represent the most "concrete" part of our value hierarchy. If our design principle is to "adhere to the safety factor" for our bridge, then the design requirements that implement that principle would be things like "support pillars every 20 feet" and "steel construction" (perhaps specified further to identify relevant properties of the steel). These design requirements are the sorts of things that directly make their way into our product or project. If the Google Glass should store information for easy retrieval (a design principle), then it can implement this principle with the use of a Cloud storage service built into the device (a design requirement). Similarly, in the aviary design example, the principle of "sufficient living space" might by specified into the requirements to have at least 450 square centimeters of floor area per hen. This now provides us a precise requirement for our overall design (even as it admits of allowing us to go beyond it).

There are two important features of the move from design principles to design requirements that we should keep in mind. The first is that just as we typically need multiple design principles to sufficiently specify our value, we may also need multiple design requirements to fully satisfy a single design principle. In the aviary example, in fact, "sufficient living space" is translated into 4 design requirements. Simply identifying the minimum floorspace is insufficient, we also have to think about space at the feeding trough, height of the living space, and slope of the living space. Similarly, to fully specify our design principle of adhering to the safety factor, we probably need

to specify both the number of support pillars and the appropriate construction materials. Just doing one of those things would not fully meet the demands of the principle.

The second feature is that sometimes a single design requirement may be relevant to adhering to multiple design principles. For instance, in the EU's guidelines on chicken aviaries, it not only provides a design principle related to the size of the living space. It also includes the principle that hens should be able to rest on perches. Design requirements specifying the size of those perches simultaneously help us implement the perch principle and the sufficient living space principle. This can obviously be valuable from the perspective of simplicity in design. However, it is typically better to first try to lay out how to meet each principle independently and then, on a later review, identify potential overlaps or synchronicities.

4. Conclusion: Engaging in Value Sensitive Design

This chapter has introduced an approach to design known as Value Sensitive Design. The overall goal of this approach is to account for values early in the design process and throughout the design process. To do that, we examined two important VSD tools: the stakeholder analysis and the value hierarchy. There are many more tools in the VSD toolkit beyond these two, but together they provide useful techniques for taking seriously the value-laden nature of technology and technological design.

Check Your Understanding

After successfully completing this chapter, you should be able to answer all the following questions:

- What does it mean for technology to be value laden? How does that compare to the idea of technology as value neutral?
- As a general approach to design, what is Value Sensitive Design?
- In the context of engineering and technology, what is a Stakeholder? What is the difference between Indirect and Direct Stakeholders? You should be able to provide an example of each
- What is the purpose of a Stakeholder Analysis?
- In general, what is a Value Hierarchy and what is its purpose in technological design?
- What is the difference between an intrinsic and an instrumental value? What is an example of each?
- What are Design Principles? How do they relate to values? What is an example of a Design Principle?

- What are Design Requirements? How do they relate to Design Principles? What is an example of a Design Requirement?

References & Further Reading

Friedman, Batya & David G. Hendry (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

van de Poel, Ibo (2013). “Translating Values into Design Requirements,” in *Philosophy and Engineering: Reflections on Practice, Principles and Process*, edited by D.P. Michelfelder et al. Springer Press.

5.

Experimental Technology & The Principles of Engineering Ethics

Key Themes & Ideas

- Experimental technologies are those technological products and projects for which the social benefits and risks are uncertain and/or unknown
- We can borrow ethical principles developed for human medical research to ethically analyze and evaluate technologies, especially experimental technologies
- The 3 principles of engineering ethics are the Principle of Welfare, the Principle of Autonomy, and the Principle of Fairness
- The principles function as ethical design constraints that must be met by a technology to be morally successful
- Each principle directs us to focus on specific values as well as reason in specific ways about those values and technologies in general

---

The Smartphone Social Experiment[1]

1. This discussion comes from John Danaher's "New Technologies as Social Experiments: An Ethical Framework," *Philosophical Disquisitions*: March 15, 2016. https://philosophicaldisquisitions.blogspot.com/2016/03/new-technologies-as-social-experiments.html

What was Apple thinking when it launched the iPhone? It was an impressive bit of technology, poised to revolutionize the smartphone industry, and set to become nearly ubiquitous within a decade. The social consequences have been dramatic. Many of those consequences have been positive: increased connectivity, increased knowledge and increased day-to-day convenience. A considerable number have been quite negative: the assault on privacy; increased distractibility, endless social noise. But were any of them weighing on the mind of Steve Jobs when he stepped onstage to deliver his keynote on January 9th 2007?

Some probably were, but more than likely they leaned toward the positive end of the spectrum. Jobs was famous for his 'reality distortion field'; it's unlikely he allowed the negative to hold him back for more than a few milliseconds. It was a cool product and it was bound to be a big seller. That's all that mattered. But when you think about it this attitude is pretty odd. The success of the iPhone and subsequent smartphones has given rise to one of the biggest social experiments in human history. The consequences of near-ubiquitous smartphone use were uncertain at the time. Why didn't we insist on Jobs giving it quite a good deal more thought and scrutiny? Imagine if instead of an iPhone he was launching a revolutionary new cancer drug? In that case we would have insisted upon a decade of trials and experiments, with animal and human subjects, before it could be brought to market. Why are we so blasé about information technology (and other technologies) vis-a-vis medication?

Human experimentation is essential to medical innovation. It is through human experimentation that we get the best understanding of the risks (and benefits) of a potential therapy. But, importantly, we also only conduct human experimentation when we are unsure of the potential hazards and/or are uncertain of their likelihood. Strictly speaking, a **risk** is a measure of the *probability* of some hazard occurring multiplied by its *magnitude* – that is, a measure of how bad the hazard would be if it did occur. Sometimes we are **ignorant** of potential hazards – we simply do not know what effects something may have and therefore cannot determine magnitude – and other times we know what may occur but are **uncertain** as to its likelihood and therefore cannot determine probability. And so, when we conduct human medical experiments, we are doing so in order to better understand what hazards are possible and/or how likely the hazards are.

But because human medical experimentation is carried out under these conditions of ignorance and uncertainty, we insist on strict ethical controls. We are likely all familiar with some of these ethical controls: all research subjects must be informed of the potential risks (as well as the lack of complete knowledge) and voluntarily agree to be exposed to them; research can only be carried out when there is good reason to believe the therapy could in fact provide a meaningful benefit. There are others as well, which we need not detail now. The central point is simply that we require strict ethical scrutiny of human medical experiments precisely because they involve exposure to unknown hazards with unknown probabilities.

But isn't the same true of many technological innovations? Recall the Control Dilemma from a previous chapter: it shows us that for many technological products and projects, we find ourselves in the same position as medical researchers: we are simply unaware of some of the potential effects and we are uncertain of the likelihood of other effects that we are aware of. Moreover, just like medical therapies, many technologies can have a deep and lasting

impact on individual lives and on the trajectory of society. Just think which has had a larger social impact: the invention of vaccines or the invention of the internet? Whatever the answer to that question, it is not an obvious one, and that alone should show the affinity between medical therapies and technological innovations. And if there is this sort of meaningful parallel, and we hold human medical research to high ethical standards, ought we not similarly hold the introduction of technology to high ethical standards?

1. Experimental Technology

At the time of its launch, the iPhone had two key properties that are shared with many other types of technology:

1. **Significant Impact Potential**: It had the potential to cause significant social changes if it took off; and
2. **Uncertain and Unknown Impact**: Many of the potential impacts could be speculated about but not actually predicted or quantified in any meaningful way while some of the potential impacts were complete unknown at the time

These two properties make the launch of the iPhone rather different from some other technological developments. For example, the construction of a new bridge could be seen as a technological development, but the potential impacts are usually more easily identified and quantified (although this can vary). Since we have a lot of experience building bridges and long running familiarity with the scientific principles that underly their design and construction, there is a lot less uncertainty involved. But, of course, if the bridge uses new techniques or new materials, or a bridge is used for a new purpose, then even a bridge can start to have properties similar to the iPhone. Nonetheless, what we can say here is that some technologies fall into a special class that merit greater ethical scrutiny:

> **Experimental Technology:** New technology with which there is little operational experience and for which, consequently, the social benefits and risks are uncertain and/or unknown

Experimental technologies, more than others, are subject to the Control Dilemma: they are those technologies where we are substantially ignorant and uncertain of the risks and therefore are not in a position to control them early on. But it is also worth noting that whether a given technological product or project is “experimental” is really a matter of degree: even a new bridge may be experimental, at least to some degree.

So, many technological products and projects are (to some degree) human experiments. When a human experiment involves medical therapies, we insist on strict ethical controls. But these experimental technologies can have at least as great, if not a greater, effect on individuals and society than the medical therapies. Thus, if we think it is right to have strict ethical controls for medical experimentation, we should also have strict ethical controls for engineering and technology experimentation.

But what does that look like?

2. The Principles of Engineering Ethics

A good starting point for developing an ethical framework for thinking about experimental technologies is to consider the ethical framework most commonly used to think about human medical experiments, a framework known as **Ethical Principlism**. In broad outline, ethical principlism is an ethical framework centered around multiple equally important, but sometimes conflicting, ethical principles. In the context of medicine, there are three major principles:

The Principle of Beneficence: Experiments should only be done if there is a potential of individual or social benefit and the human subjects involved should not be harmed

The Principle of Respect for Autonomy: Human autonomy and agency should be respected

The Principle of Justice: The benefits and risks ought to be fairly distributed

These three principles are intentionally general and somewhat vague. The basic idea is that they represent widely shared ethical commitments and provide the basis for establishing more precise and practical guidelines for medical researchers.

By and large, we can apply these same three principles to thinking about experimental technology. But, of course, out context is a bit different (even if similar) and so we will want to tweak things a bit. Additionally, for the sake of simplicity, we can rename the principles. In so doing, we end up with the following three **Principles of Engineering Ethics:**

The Principle of Welfare. Technological products and projects should aim at promoting social welfare and avoid harming or risking harm to individual and social welfare

The Principle of Autonomy. Technological design and implementation should respect the right of all people to make their own informed choices and develop their own life plans

The Principle of Fairness. The benefits and burdens of technology should be distributed fairly

We rename the Principle of Beneficence to the Principle of Welfare to make clearer that its focus is on promoting and protecting welfare, thereby underwriting the most fundamental principle of engineering ethics. Additionally, we spell out what it means to respect “human autonomy and agency” and, finally, rename the Principle of Justice to the Principle of Fairness to avoid confusion with thinking about (for instance) criminal justice or the law more generally.

3. Applying the Principles

We now have three principles for scrutinizing the design and implementation of technology into society. But just as the principles of biomedical ethics are general and require specification for context, so too do the principles of engineering ethics. So, the question arises, how do we work with these principles?

Technological design and implementation already involves working with a variety of necessary but sometimes conflicting **design constraints**. For instance, buildings must be able to resist wind forces up a certain speed and bridges must be able to hold at least a certain weight level. But these considerations often butt up against other constraints regarding, for instance, total weight of the bridge or building. The project must be kept under a certain total weight but also must display a certain level of strength. Both of these are constraints on the design, and must be adhered to, but also must be balanced against one another – you cannot build the bridge maximally strong, for it will be too heavy, and you cannot build it maximally light, for it will be too weak. Similar considerations apply when we aren't talking about safety: an interactive device must balance simplicity with power, other products may need to balance longevity with accessibility.

And so, we can think of the ethical principles as **Ethical Design Constraints** that function in a similar manner. Each principle must be adhered to for a project to be (ethically) successful, but the requirements of the principles will sometimes come into conflict. Indeed, one basic conflict applies to both engineering and medicine and is illustrated with the iPhone example: we could, hypothetically, make significantly greater medical progress if we *did not* inform people they were research subjects and *did not* worry about whether the experiment would be harmful to them. In this way, if we ignore the principle of autonomy and half of the principle of welfare we would be better able to achieve what the first half of the welfare principle requires. But such use of people for the benefit of others should strike us all as morally suspect.

In functioning as ethical design constraints, the principles provide broad guidelines that must always be accounted for in design and implementation. So, then, how do we account for them? To do that, it can help to recognize that the principles really engage us in two ways: they both tell us what to pay attention to as well as what to do with the information we are paying attention to.

3.1. Thinking about Welfare

The Welfare Principle, as its name suggests, directs us to pay attention to the ways in which welfare may be positively and negatively affected. In this way, it functions quite similarly to Canon 1 of the NSPE Code of Ethics (and the first principle of many other engineering codes of ethics). But, importantly, it also directs us to pay attention to two distinct ways that a technology may affect welfare: it may enhance it by either providing a new benefit or by reducing or eliminating existing harms or risks. So, for instance, the introduction of FaceTime provided a new benefit: the ability to visually converse via phone. On the other hand, new water treatment technology can be understood as reducing or eliminating the existing risk of water-borne illness. These are two ways of "doing good". But the Welfare Principle also directs us to "do no harm", which means not *introducing* new harms or risks through the technology. Thus, this second part of the principle directs us to (for instance) not introduce a technology that will pollute the water and thereby create (or increase) the risk of water-borne illness.

This point about introducing new risks and reducing existing risks is important because they influence how we think about what to do once we have identified the potential effects on welfare. There is a millennia old debate over whether it is morally worse to *cause harm* than to *fail to benefit*. On one view, so long as the consequence is identical, then the moral evaluation is identical. Thus, were you to fail to save a drowning child from a pond when you were able to, that would be just as bad as if you had actually drowned the child in the pond. According to this *symmetry thesis*, harms caused are symmetrical with benefits denied – so long as their likelihood and effect are identical, they are morally identical. In contrast, the *asymmetry thesis* holds that causing harm is morally worse than failing to benefit. While it may still be morally bad to fail to save the drowning child from the pond, it is not as morally bad as if you were to have drowned the child.

We will not resolve this debate here (there is a reason it has been raging for thousands of years). Instead, we point it out so that we all may understand the competing perspectives. But what position one takes on this "doing/allowing" debate does often influence how one will apply the welfare principle. If you believe that it is morally worse to cause harm, then you will "err on the side of caution" and be more willing to (for instance) stop a project or reduce the power of a technology so as to limit the introduction of new risks. On the other hand, if you accept the symmetry thesis then you are much more willing to say "the benefits outweigh the risks" and therefore be willing to justify almost any risks so long as the benefits are greater.

Our stance on the doing/allowing debate will influence how we apply the welfare principle to some degree, but there are nonetheless some general guidelines that the principle offers us. These include:

- Only pursue a product or project if it is reasonable to expect social benefits
- Aim to reduce the potential for harm to the public as far as reasonably possible
- Conduct controlled "experiments" of the technology whenever possible
- Provide safeguards that allow a project to be halted or a product to be recalled if excessive harm is discovered
- Avoid products and projects that may produce irreversible harm
- Pursue products and projects that have the potential to improve social or environmental resiliency (the ability to 'bounce back' from bad situations)

Like the principle itself, these guidelines are still a bit general and vague: it leaves open, for instance, what counts as "reasonable possibility". But they are more precise than the principle itself and help us understand, in more detail, how to reason with it.

3.2. Thinking about Autonomy

Autonomy is derived from Ancient Greek and literally means "self-legislate" (auto-nomos). Although it hasn't always been the case, in contemporary society we value individual autonomy, even at the expense of individual and social welfare. For instance, we leave people quite free to make their own decisions even knowing that many of them will make decisions that reduce their welfare. This is because we value people being the authors of their

own lives alongside valuing their welfare. Put another way, we think it important not just that people live good lives but that they control their own lives. Obviously, this control is not absolute, but it matters, nonetheless.

In the medical context, the autonomy principle grounds the practice of *informed consent*: we only allow researchers to research on human subjects who have been properly informed of the risks involved in the experiment and have voluntarily agreed to be exposed to those risks. It is not the role of the researcher to decide whether someone else is exposed to risks.

The principle functions similarly in the technology case. First, it directs us to pay attention to what information people have about the technology and whether they are in control of their exposure to it or not. Relatedly, and as we will explore in more detail in a later chapter, it encourages us to think about their **privacy**, for privacy is a means by which we protect our autonomy. My ability to control what others know about me, for instance, helps me limit the ways in which others may try to interfere with my informed choices and the development of my life plans. Thus, whereas the principle of welfare directs us to focus on all those values associated with welfare, the principle of autonomy directs us to focus on those values associated with autonomy: honesty, adequate information, good reasoning and decision-making, control over our bodies and personal information, and individual freedom more generally.

An important note must be made about respecting autonomy. The principle directs us to respect the informed choices of autonomous people. But not everyone is autonomous all the time. If I have been systematically deceived and am therefore making choices based on false information, then I am not making autonomous choices. Relatedly, even if I have accurate information, if I am (for instance) drunk then I may not be reasoning autonomously. Thus, autonomy can be imperiled both by false information and by various barriers to good reasoning. Respecting autonomy is *not* about just letting people do whatever they want. However, we must also be careful not to judge someone as non-autonomous just because they are making choices we don't agree with. Autonomous choice does not require making the best or right choice either.

Like the principle of welfare, the principle of autonomy also provides a few guidelines for thinking through technology's impact on autonomy:

- Ensure all those exposed to the risks of a technological product or project are informed
- Ensure all those exposed to the risks of a technological product or project have voluntarily agreed to be exposed
- Introduction of experimental technologies should be approved by democratically legitimized bodies
- Those involved in the "experiment" should be able to withdraw from the experiment
- Reduce, as far as possible, any exposure to risk for people who have not been informed and/or have not approved the product or project

Once again, these guidelines are intentionally general so as to be widely applicable. It should also be noted that all these guidelines (including those for the other principles) are often statements of what is *ideal*. As a matter of fact, it may be impossible to "ensure all those exposed to the risks of a technological product or project are informed"

since, unlike medical experiments, technologies are often simply "tested on society" via their introduction as a consumer product or their implementation as a public works project. We simply do not know who all will be affected and therefore cannot inform them in advance. That does not mean the guideline is irrelevant, though. By setting the ideal that all will be informed, we know what we should aim for, even if we cannot ever quite reach it. This is certainly better than not even attempting to meet the guideline at all.

3.3. Thinking about Fairness

The Principle of Autonomy is fundamentally about protecting *individuals*, for it is only individuals that have autonomy. The Principle of Welfare is about both individuals and groups, since we can meaningfully speak about both individual and social welfare. The Principle of Fairness, on the other hand, is fundamentally about groups and not individuals. Whereas the Principle of Autonomy augments our pursuit of social welfare by basically telling us not to sacrifice the individual, the Principle of Fairness augments our pursuit of social welfare by telling us to focus not just on the social benefits but also on how they are *distributed* across society. In this way, both autonomy and fairness help us avoid the sort of "tyranny of the majority" potential that characterized medical research before the introduction of the ethical principles.

The Principle of Fairness, then, directs us to focus on values like fairness, equality, democracy, and the protection of the vulnerable. It is *unfair* for only a select few people to gain all the benefits of a new technology whilst everyone else pays all the costs. More positively, technology has the potential to, and has in fact, promoted greater equality in society and "democratized" society. Various social goods that were once only available to nobility are now available to everyone while the creation of new means of communication have enhanced peoples' abilities to have a say in how their society progresses.

Once we have identified the potential for unfair distribution or potentially affected vulnerable populations, the Principle of Fairness offers us the following sorts of guidelines for moving forward:

- Avoid imposing risks on vulnerable populations or, at the least, offer additional protections and/or compensation if they are exposed
- Avoid imposing risks on specific populations without any corresponding benefit to that same population
- Distribute the benefits and risks of a technology fairly across all those affected by the technology
- Avoid imposing irreversible risks on society

4. The Principles & Ethical Design

As should now be clear, these ethical principles are applicable to our thinking about technology at multiple levels. As policymakers or members of society, we can use the principles to think about existing or emerging technologies and how they are implemented into society. As designers of technology we can also use the principles to help guide the design and implementation of technology in society.

When using the principles to think about design, we should think broadly. That means considering the immediate

risks the technological artifact may pose – for instance, the potential health effects of some new building material. But it also means considering the wider social effects of the use of technology in society – for instance, the potential health effects of the spread of misinformation due to the widespread use of social media as a result of the easy accessibility of smartphone devices. This is why keeping in mind the social context of technology and technological mediation is important – they both help expand our thinking beyond the technological artifact to the techno-social system as a whole.

We should thus endeavor to speculate, within reason, about the potential long-term social effects of a technology on existing social institutions and systems. For instance, the principle of welfare does not only direct us to think about the welfare effects of the technology but also how the technology may affect existing social institutions and practices that protect or promote welfare. The spread of misinformation via social media both directly affects health, a welfare value, and indirectly affects existing health-oriented institutions by, for instance, leading to an increased burden on health resources, healthcare workers, and the health system as a whole. Similarly, the rise of consumer surveillance – Ring doorbells, etc. – raises both immediate privacy concerns with the device itself but also has a long-term effect on how we think about privacy in society. And, finally, existing institutions that promote fairness such as democratic elections can be imperiled both directly and indirectly by technology.

In short, then, a major value of these principles is their wide applicability. They are simply statements of nearly universal ethical principles, with application to thinking about technology. As such, in applying them, we should keep in mind both the narrow applications and the wider applications.

Check Your Understanding

After successfully completing this chapter, you should be able to answer all the following questions:

- What is a **risk**? What are the 2 ways we can lack knowledge of technological risks?
- What is an **experimental technology**? What are the 2 key features of experimental technologies?
- What does it mean for the **principles of engineering ethics** to function as **ethical design constraints**?
- What does the **principle of welfare** focus on? What values does it direct us to focus on and how does it tell us to think about technology?
- How does the “doing/allowing” debate relate to the principle of welfare? What are the 2 main positions (or ‘theses’) related to the debate?
- What does the **principle of autonomy** focus on? What values does it direct us to focus on and how does it tell us to think about technology?

- What does the **principle of fairness** focus on? What values does it direct us to focus on and how does it tell us to think about technology?
- What are the various "levels" of analysis that the principles can be applied to? What are some questions that we might ask at each level?

SECTION II

Issues in Engineering & Technology Ethics

6.

Engineering & the Environment

Key Themes & Ideas

- Engineering and technology have an outsized impact on the environment
- Maintaining an environment sufficiently clean from pollution remains an important but complex issue
- Recent commitments to sustainable development emphasize a more proactive approach to environmental concern
- There are two competing frameworks for sustainable development: ecomodernism and degrowth
- Life cycle analyses aim to take into account the full life of a technological artifact, rather than focusing only on production and use
- Cradle-to-Cradle Thinking encourages an even wider perspective on the life of an artifact, including its impact once it is broken or unusable for its original purpose

---

Engineering and technology have done wonders in improving human welfare. Infant and early childhood mortality have been significantly reduced, average life spans significantly increased, and the overall quality of most peoples' lives are higher now than ever in the past. Just consider that in the 17th and 18th centuries the pineapple was a rare gift reserved only for royalty while now anyone can walk into a supermarket and purchase a can of pineapple for a few dollars. While a pineapple certainly is not essential to a good life, it is representative of the overall shift in quality of life that was only made possible by advancements in technology.

But these improvements in human well-being have had a cost. Pollution, waste, and climate change are all the result of technological advancement. The extraction and use of fossil fuels for electricity, heat, and transportation

is the largest contributor to climate change and among the most polluting sectors of our economy. Industrial manufacturing and construction is not far behind.[1] And these sorts of industries are only made possible due to the past innovations of engineers and the continued support of the engineering profession.

Engineering's impact on the environment raises myriad ethical concerns. Many of these concerns relate to issues of fairness: Is it fair that some people – namely those living in wealthy industrialized nations – reap the benefits of environmental destruction while others – those living in poor and unindustrialized nations – pay the costs? Is it fair that humans have spoiled the natural environment, risking the lives of non-human animals and the existence of species, for their benefit? And is it fair that current generations effectively leverage the livelihoods of future generations for their own benefit?

And it should be noted that engineering's contribution to environmental harms is only increasing with the rise of digital technologies. While such technologies may reduce visible pollution such as paper trash, their existence depends on highly polluting extraction activities. Silicon, lithium, and the other resources that are essential to electronic technology are all in short supply and exceedingly destructive to harvest. And, of course, all these computer systems require massive amounts of energy to run and so the segment of the economy that already most contributes to climate change is only put under more pressure.

All of this has led to increasing interest in how engineers can take seriously their responsibility to hold paramount the welfare of the public, including their ability to engage with nature and the ability of future generations – still members of the public – to live good lives. Of course such concern is not totally new as concern for the environmental impacts of development began in earnest in the 1960s (in the US at least). But increasing concern over climate change as well as greater capability for technological innovation has pushed thinking in new directions.

In this chapter, we'll explore some ways of thinking about environmental impact and approaches engineers may take to improve their environmental consciousness and design approaches. We begin with some initial thinking about air and water pollution before turning to the contemporary focus on **sustainable development**.

1. "Sufficiently Clean": Environmental Laws & Regulation

The first main wave of the environmental movement, especially as it carried over into law and regulation, focused on managing various forms of environmental pollution. The iconic images of Cleveland's Cuyahoga River on fire in 1969 helped birth the Environmental Protection Agency (EPA) in the United States and led to a focus on "cleaning up the environment". Many contemporary policies continue with this emphasis on an "acceptably clean environment". Importantly, however, there is disagreement over how to interpret "acceptably clean" (or "sufficiently clean"), disagreement which has immediate effects on the interpretation and enforcement of these laws. The box below reproduces Harris, Jr., et al.'s summary and assessment of the 6 commonly discussed criteria.

1. The 5th Assessment Report of the Intergovernmental Panel on Climate Change. Report can be accessed at AR5 Climate Change 2014: Mitigation of Climate Change — IPCC

When is the environment sufficiently clean?[2]

1. According to the **comparative criterion**, an aspect of the environment is sufficiently clean if and only if it imposes no greater threat to human life or health than do other risks that most people might consider reasonable. This is a defective criterion. It may often be the case that the public does not understand the seriousness of certain risks they accept. Furthermore, data about comparative risks are often difficult to obtain.
2. According to the **normalcy criterion**, an aspect of the environment is sufficiently clean if and only if any pollutants present in it are normally present in it to the same degree. However, if the pollutants present in a river or the air are "normally" present, they could still pose a threat to human and animal health.
3. According to the **optimal pollution reduction criterion**, an aspect of the environment is sufficiently clean if and only if funds required to reduce pollution further could be used in other ways that would produce more overall human well-being. According to this criterion, if funds necessary to make the Cuyahoga River sufficiently clean (e.g., by one of these criteria) could be better used to remediate an environmental problem somewhere else, the Cuyahoga River should be left in its present condition. This seems unsatisfactory.
4. According to the **maximum protection criterion**, an aspect of the environment is sufficiently clean if and only if any identifiable risk from its pollution that poses a threat to human health has been eliminated, up to the limits of technology and the ability to enforce. This criterion could require all available funds to be spent on a single environmental remediation project if it were serious enough, leaving many other problems unaddressed.
5. According to the **demonstrable harm criterion**, an aspect of the environment is sufficiently clean if and only if every pollutant that is demonstrably harmful to human health has been eliminated. Still stronger than the previous criterion, this criterion eliminates not only considerations of cost but also considerations of technical feasibility. It also requires proof of harm to human health, which is sometimes difficult to obtain. The criterion thus seems to be unrealistic.
6. According to the **degree of harm criterion**, an aspect of the environment is sufficiently clean if and only if cost is not a factor in removing clear and pressing threats to human health, but when the degree of harm is uncertain, economic factors may be considered. This criterion may suggest the best balance of cost and health considerations and seems to be the closest to the position taken by many court decisions.

2. This list and discussion was originally developed Charles E. Harris, Jr., et al. (2019), *Engineering Ethics: Concepts and Cases* (Cengage Learning), 159.

2. Sustainable Development

Environmental law and regulations set minimal standards for thinking about engineering and the environment. But recall that part of what makes an occupation a profession is that its practitioners hold themselves to a higher standard than what is required by law. As such, while environmental law and regulation lags behind in its narrow focus on cleaning up and limiting pollutants, the engineering profession has moved to broader questions about the long-term sustainability of human and nonhuman life. Various engineering codes of ethics include a requirement to "adhere to principles of sustainable development". But what exactly is sustainable development?

There exists, as one might expect, disagreement over the meaning of sustainable development and the principles that support it. But the best-known contemporary definition comes from the World Commission on Environment and Development (WCED):

Sustainable development is "development that meets the needs of the present without compromising the ability of future generations to meet their own needs."[3]

The WCED further identified five goals for sustainable development:

1. Economic growth
2. Fair distribution of resources to sustain economic development
3. More democratic political systems
4. Adoption of lifestyles that are more compatible with living within the planet's ecological means
5. Population levels that are more compatible with the planet's ecological means

Reflection on this definition and these five goals should make clear the tension that sits at the heart of sustainable development: goals 1-3 are all human-focused while goals 4 and 5 are distinctly environmental. Because sustainable development is still about *development*, it makes sense that it would maintain at least some of the human-centered concerns that have led to environmental issues in the first place. But we should ask: is sustainable development possible? Are goals 1-3 compatible with goals 4 and 5? Indeed, some have criticized the very notion of sustainable development on the grounds that it is an attempt to combine two incompatible ideas. We turn, in the next section, to investigate this debate and develop our thinking about sustainable development in more detail.

3. How do we Develop Sustainably?

Even once we have accepted that sustainability is a worthwhile goal and that engineers should adhere to principles of sustainable development, we are still faced with the question of how we best adhere to those principles and achieve that goal. To help us fill in the details, we can examine a contemporary debate between two competing frames for thinking about the relationship between sustainability and technology. This is the debate between

3. World Commission on Environment and Development, *Our Common Future* (Oxford University Press, 1987), cited in Stanley R. Carpenter, "Sustainability," in Ruth Chadwick, ed., *Encyclopedia of Applied Ethics* (San Diego, CA: Academic Press, 1998), 275-293.

Ecomodernism and Degrowth. As should become clear, one of these frames accepts the possibility of truly sustainable development while the other is skeptical.[4]

Ecomodernism is an environmental philosophy that focuses on sustainability through forward technological innovation. Ecomodernists hold that humanity has the power to solve today's major ecological challenges without making fundamental changes to our behavior and social structures. For the ecomodernist, economic growth is not opposed to sustainable development but rather essential to it as economic growth makes possible technological innovation that will allow us to replace older, environmentally destructive technologies with alternatives. Ecomodernists will tend to support innovations like renewable energy, genetically modified organisms, precision agriculture, and synthetic meat and oppose calls to "downsize" or "return to nature".

The **Degrowth** approach takes precisely the opposite view. In contrast to ecomodernism, Degrowth theorists hold that there is an essential tension between economic growth and sustainability. Rather than focus on economic growth, Degrowth theorists emphasize establishing fair social conditions, environmental justice, and re-embedding our 'economic metabolisms' in the natural cycles of the biosphere – even if that means less overall growth. Degrowth theorists believe in the regenerative potential of natural environments and tend to oppose technological innovations as wasteful, unnecessary, decadent, or superfluous.

The above descriptions are brief, and of course both approaches have substantial literature and complexities behind them. Nonetheless, we can take this brief description, along with the comparison table below, as a starting point for thinking through some other elements of sustainable development.

Comparison Table: Ecomodernism & Degrowth

Ecomodernism	Degrowth
Technological optimism	Technological pessimism
Continue growing the economy	Cease (or slow) economic growth
Innovate and improve	Downsize and reduce
Keep current comfort & consumption	Produce less, consume less
Overcome the harsh natural world	Harmonize with nature
Primary problem: Resource shortage	Primary problem: Human ambition

3.1. The Jevons Paradox & "Techno-Fixes"

One central element of the disagreement between Ecomodernists and Degrowth theorists is about whether using technology to increase efficiency of resource use will, in fact, have the effect of decreasing resource consumption overall. Ecomodernists think so, but degrowth theorists will generally suggest that human ambition and avarice will just lead to more use of the resource overall. For degrowth theorists, these innovations are mere "techno-fixes" that only kick the problem down the road.

4. This discussion is reproduced from Roel Veraart's "Thinking Technology – Degrowth vs. Ecomodernism" (2021). It was published under a CC-A-SA license. Original can be found at Degrowth vs. Ecomodernism | Edusources

The work of English Economist William Jevons can provide insight into this disagreement. Working in England in the 19th century, Jevons observed that as coal production and use became more efficient it actually led to an increased consumption of goal, rather than a decrease. This flies in the face of standard economic thinking, which holds that supply and demand should inevitably balance out. In the case of coal use, increasing efficiency increased supply presumably to meet demand. But demand just continued to rise. **Jevons Paradox**, as we now call it, thus shows that when it comes to the use of (some) resources, consumption rises together with production instead of balancing out. In observing this paradox, Jevons concluded that in contrast to standard economic views of the time (and still today) technological progress *does not* guarantee reduced consumption.

Contemporary applications of the Jevons Paradox focus on two related concepts: Rebound effects and "techno-fixes". The **rebound effect** makes sense of the Jevons Paradox: much of the time when the harvesting and use of a resource becomes more efficient it also becomes cheaper and, as a result, people are incentivized to use more of the resource. This has been seen across the globe when it comes to electricity and gasoline production. In both cases, as we have become more efficient at production we have driven costs down and, as a result, increased usage. This makes sense in standard economic models, as well, where demand is "elastic" – how much of it people want or use is at least partly a function of how much it costs. In effect, the Jevons Paradox applies to goods with substantial elastic demand but not to goods with little or no elasticity.

Techno-fix is a more recent idea that loosely builds on the Jevons Paradox and rebound effect ideas. Although there is nothing inherently problematic with "technological solutions", the term techno-fix is typically used derisively to refer to attempts to solve a problem created by technology with more technology, only to create new and different problems (or simply not fix the ones they were aimed to fix) while simultaneously giving people the false sense of security that the problem has been solved. This sort of issue is perhaps best illustrated by recycling technology, especially plastic recycling technology. As plastics increased in use in the mid-20th century there became increasing social awareness around plastic waste. Recognizing that such concern could lead to a reduction in demand for plastics, plastic manufacturers and chemical companies (such as Dow) began pitching the idea of individuals "recycling" their plastic. However, despite proclamations from the companies, recycling was never a real solution to the issue of plastic waste – it is inefficient, overly costly, and produces substantial pollution. Nonetheless, individuals became convinced that their plastic use was not environmentally harmful so long as they recycled and the result is a world covered in plastic.

Notice the parallels between the Jevons Paradox and techno-fixes: techno-fixes hypothetically make our resource usage more efficient (for instance by creating a production loop) but the result is people feel better about the use of the resource and therefore increase their use, creating a rebound effect. Thus, while discussion of techno-fixes is not identical to the precise economic discussion of the Jevons Paradox, there is a similarity.

3.2. Life Cycle Analysis & Cradle-to-Grave Thinking

One common criticism of existing environmental law and regulation is that it thinks too narrowly about environmental impacts. By focusing almost exclusively on acceptable levels of pollution, we neglect the environmental effects of the materials or products being used, as well as the environmental effects of the manufacturing processes themselves. To broaden our focus, we can appeal to the method of **Life Cycle Analysis**

(LCA). LCA is a type of "cradle-to-grave" thinking that aims to consider the entire life history of a product or process. This includes the extraction of raw materials from the earth, the manufacture and use, and the final disposal.

It is common for us to be overly narrow in our thinking about environmental impacts. Consider, for instance, the increasing opposition to plastic shopping bags and the concomitant rise of reusable shopping bags. Here, the main case against plastic shopping bags comes from the lack of proper disposability – it is not economically feasible to recycle plastic bags and so they end up in landfills (or on the sides of streets). In virtue of being reusable, reusable bags are seen as less of an issue in terms of disposability. However, the creation of reusable bags – be they plastic, cloth, or other – typically involves much greater resource usage and pollution than the creation of plastic bags. And reusable bags tend to have the same disposability issue, albeit further down the line. Thus, the value of an LCA can be seen in giving us a wider understanding of the issue. It may still be that, all things considered, reusable bags are preferable. And, of course, the specific materials and methods matter a lot. But the value of the LCA is in helping us see that the issue is not as cut and dry as popular discussions would have it.

It should be noted that the LCA method is not without its weaknesses, largely related to the difficulty of collecting the relevant data and of making complex comparisons. Nonetheless, the LCA method and Cradle-to-Grave thinking in general can be valuable in promoting sustainability, particularly by encouraging us to think about the environmental impact that comes *before* manufacturing and use as well as that which comes *after*.

3.3. Biomimicry & Cradle-to-Cradle Thinking

The LCA method and Cradle-to-Grave thinking are fundamentally *linear*: we begin with extraction of raw materials and end with disposal. A more recently developed alternative attempts to adopt the circularity of natural processes. This application of **biomimicry** – the emulation of natural processes in artifactual design – has been dubbed **Cradle-to-Cradle Thinking**.

Broadly, advocates of Cradle-to-Cradle Thinking (C2C) note that natural processes tend to be highly efficient – only using the energy they need – and that there is no waste in natural processes. Following this, advocates suggest that human beings don't have a pollution problem but rather a design problem.[5] Our tendency is to design only for the first use of a product, ignoring potential uses after the product breaks, crumbles, or otherwise becomes (seemingly) useless. C2C advocates William McDonough and Michael Braungart (an architect and chemist, respectively) contrast human design with an ant colony. A colony of ants will handle their waste, grow and harvest their food, build their houses out of recyclable material, and make the soil healthier than it would otherwise be. Ants don't produce waste. McDonough and Braungart suggest, then, that "To eliminate the concept of waste means to design things—products, packaging, and systems—from the very beginning so that, at the end of a product's useful life, the inorganic (or 'technical') components can be separated from the organic components, the former being 'upcycled' into new products and the latter being returned to the earth for reuse in the natural cycle."[6]

5. William McDonough and Michael Braungart (2013), *The Upcycle* (North Point Press), pp. 7-8.
6. William McDonough and Michael Braungart (2002), *Cradle to Cradle* (North Point Press), 104.

As an example of C2C thinking, McDonough and Braungart produced their book *Cradle to Cradle* of compostable and nontoxic materials and ink so that it could simply decompose back into the Earth without adding harmful toxins (notably, the paper is a form of plastic!).

4. Taking Environmental Responsibility Seriously

With the ever increasing threat of climate change, engineer's will feel ever growing pressure to take their environmental responsibilities seriously. While some have proposed exercising their professional autonomy to refuse to engage in engineering practices that further climate change, others emphasize the compatibility of technological innovation and environmental stewardship. It is likely that any comprehensive approach will require a variety of outlooks, methods, and tools. It may be the case that engineers should simply cease to participate in some activities; but, equally, it is clear that the welfare of the public still depends on engineering projects and technological innovation and so what will be important is finding ways to pursue those projects in a way compatible with concern for the environment. The concepts and methods discussed in this chapter can help in finding that compatibility.

Check Your Understanding

After successfully completing this chapter, you should be able to answer all the following questions:

- In what sorts of ways does engineering contribute to environmental degradation?
- What are some of the competing criteria for establishing when an environment is sufficiently clean?
- What is **sustainable development**? What are some of the goals of sustainable development?
- What is the debate between **ecomodernists** and **degrowth theorists** about? How does the **Jevons Paradox** fit into their debate?
- What is a **Life Cycle Analysis**? What are its strengths and weaknesses?
- What is **Cradle-to-Cradle Thinking**? How does it relate to the idea of **biomimicry**?

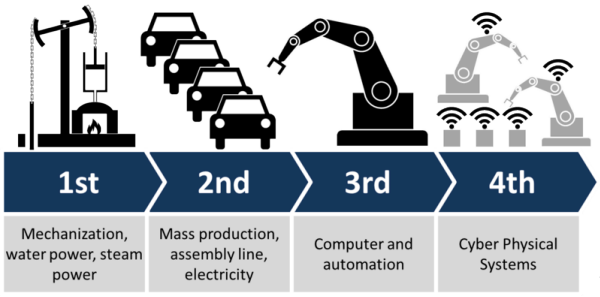
7.

The 4th Industrial Revolution

Engineering Ethics in the Machine Age

Key Themes & Ideas

- The Fourth Industrial Revolution, characterized by increasing interconnectivity and smart automation, is likely to drastically alter our social landscape
- The use of “Big Data”, machine learning, and algorithmic processing presents important social opportunities but also threatens grave personal and social harms
- The ethical use of Artificial Intelligence can help us solve previously intractable human and social problems
- Artificial Intelligence may create ethical concern from misuse, under-use, and over-use
- The Principles of Engineering Ethics, suitably refined and supplemented by a Principle of Explicability, can provide a valuable framework for ethical AI development and implementation



We are in the midst of the Fourth Industrial Revolution. Taking off where the Digital Revolution – the Third Industrial Revolution – ended, this new Machine Age is characterized by increasing interconnectivity and smart automation. This fourth industrial revolution is made possible by advancements in data science, artificial intelligence, and robotics which are beginning to blur the lines between the physical, digital, and biological worlds. According to many, the Fourth Industrial Revolution is especially characterized by an **augmented social reality** where nearly all of our interactions in the world are mediated by some form of digital technology.

This new revolution offers great opportunity but also threatens great peril. Some of what it offers is not new: the looming threat of job loss due to automation has been with us since the First Industrial Revolution. But it does increase this and other old threats, for the greater power of these new technologies make the potential harms more probable. When even creative jobs can be done by machines, what is left for human beings? Still other opportunities and risks are new, driven by wholly new possibilities for knowledge and interaction.

If we are to drive the Fourth Industrial Revolution, rather than simply being caught in its wake, it will be essential that we have a grip on what is possible and what is at stake. In this chapter, then, we examine two of the major drivers of this revolution: Big Data and Artificial Intelligence. As we will see, these drivers overlap as a great deal of what matters about Big Data is the result of at least basic forms of Artificial Intelligence: machine learning, algorithmic processing, etc.

1. Big Data & Algorithmic Processing

In contrast to (for instance) the engineering of bridges and airplanes, data practice has a much broader ethical sweep: it has the potential to significantly impact all of the fundamental interests of human beings. While unethical choices in bridge or airplane design may result in the loss of life or health, unethical choices in data practice can do this and more – it can ruin reputations, savings, or rob someone of their liberty. In contrast, of course, this also means that data practice holds out the hope of benefitting individuals and society in many more ways as well. Nonetheless, what this all suggests is that the ethical landscape for data practitioners, software engineers, and others who work with data is even more complex than that faced by other types of technology professionals.

To start to get a feel for the complexities of data practice, it will be useful to consider some of the major potential benefits and risks of data. We'll begin by examining the promise of a data-fueled world.[1]

1. The following discussion borrows heavily from Shannon Vallor's "An Introduction to Data Ethics" teaching module. The original can be found at https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/an-introduction-to-data-ethics/

1.1. The Promise of Big Data

The promise of big data can be grouped into three main types of benefits: (1) Improving our understanding of ourselves and the world; (2) improving social, institutional, and economic efficiency; and (3) predictive accuracy and personalization.

1.1.1. Human Understanding

Data and its associated practices can uncover previously unrecognized correlations and patterns in the world. In so doing, data can greatly enrich our understanding of ethically significant relationships—in nature, society, and our personal lives. Understanding the world is good in itself, but also, the more we understand about the world and how it works, the more intelligently we can act in it. Data can help us to better understand how complex systems interact at a variety of scales: from large systems such as weather, climate, markets, transportation, and communication networks, to smaller systems such as those of the human body, a particular ecological niche, or a specific political community, down to the systems that govern matter and energy at subatomic levels. Data practice can also shed new light on previously unseen or unattended harms, needs, and risks. For example, big data practices can reveal that a minority or marginalized group is being harmed by a drug or an educational technique that was originally designed for and tested only on a majority/dominant group, allowing us to innovate in safer and more effective ways that bring more benefit to a wider range of people.

1.1.2. Social, Institutional, and Economic Efficiency

Once we have a more accurate picture of how the world works, we can design or intervene in its systems to improve their functioning. This reduces wasted effort and resources and improves the alignment between a social system or institution's policies/processes and our goals. For example, big data can help us create better models of systems such as regional traffic flows, and with such models we can more easily identify the specific changes that are most likely to ease traffic congestion and reduce pollution and fuel use—ethical significant gains that can improve our happiness and the environment. Data used to better model voting behavior in a given community could allow us to identify the distribution of polling station locations and hours that would best encourage voter turnout, promoting ethically significant values such as citizen engagement. Data analytics can search for complex patterns indicating fraud or abuse of social systems. The potential efficiencies of big data go well beyond these examples, enabling social action that streamlines access to a wide range of ethically significant goods such as health, happiness, safety, security, education, and justice.

1.1.3. Predictive Accuracy and Personalization

Not only can good data practices help to make social systems work more efficiently, as we saw above, but they can also be used to more precisely tailor actions to be effective in achieving good outcomes for *specific individuals, groups, and circumstances*, and to be more responsive to user input in (approximately) *real time*. Of course, perhaps the most well-known examples of this advantage of data involves personalized search and serving of advertisements. Designers of search engines, online advertising platforms, and related tools want the content they deliver to you to be the most relevant to you, *now*. Data analytics allow them to predict *your* interests and needs with greater accuracy. But it is important to recognize that the predictive potential of data goes well beyond this

familiar use, enabling personalized and targeted interactions that can deliver many kinds of ethically significant goods. From targeted disease therapies in medicine that are tailored specifically to a patient's genetic fingerprint, to customized homework assignments that build upon an individual student's existing skills and focus on practice in areas of weakness, to predictive policing strategies that send officers to the specific locations where crimes are most likely to occur, to timely predictions of mechanical failure or natural disaster, a key goal of data practice is to more accurately fit our actions to specific needs and circumstances, rather than relying on more sweeping and less reliable generalizations. In this way the choices we make in seeking the good life for ourselves and others can be more effective more often, and for more people.

1.2. The Perils of Big Data

Many of the major potential benefits of big data are quite obvious once we have a basic understanding of what is possible. The risks, however, can be more difficult to see. Nonetheless, we can loosely group them into 3 main categories: (1) Threats to privacy and security; (2) Threats to fairness and justice; and (3) Threats to transparency and autonomy.

1.2.1. Threats to Privacy & Security

Thanks to the ocean of personal data that humans are generating today (or, to use a better metaphor, the many different lakes, springs, and rivers of personal data that are pooling and flowing across the digital landscape), most of us do not realize how exposed our lives are, or can be, by common data practices.

Even *anonymized* datasets can, when linked or merged with other datasets, reveal intimate facts (or in many cases, *falsehoods*) about us. As a result of your multitude of data-generating activities (and of those you interact with), your sexual history and preferences, medical and mental health history, private conversations at work and at home, genetic makeup and predispositions, reading and Internet search habits, political and religious views, may all be part of data profiles that have been constructed and store somewhere unknown to you, often without your knowledge or informed consent. Such profiles exist within a chaotic data ecosystem that gives individuals little to no ability to personally curate, delete, correct, or control the release of that information. Only thin, regionally inconsistent, and weakly enforced sets of data regulations and policies protect us from the reputational, economic, and emotional harms that release of such intimate data into the wrong hands could cause. In some cases, as with data identifying victims of domestic violence, or political protestors or sexual minorities living under oppressive regimes, the potential harms can even be fatal.

And of course, this level of exposure does not just affect *you* but virtually everyone in a networked society. Even those who choose to live 'off the digital grid' cannot prevent intimate data about them from being generated and shared by their friends, family, employers, clients, and service providers. Moreover, much of this data does not stay confined to the digital context in which it was originally shared. For example, information about an online purchase you made in college of a politically controversial novel might, without your knowledge, be sold to third-parties (and then sold again), or hacked from an insecure cloud storage system, and eventually included in a digital profile of you that years later a prospective employer or investigative journalist could purchase. Should you, and others, be able to protect your employability or reputation from being irreparably harmed by such data flows?

Data privacy isn't just about our online activities, either. Facial, gait, and voice-recognition algorithms, as well as geocoded mobile data, can now identify and gather information about us as we move and act in many public and private spaces.

Unethical or ethically negligent data privacy practices, from poor data security and data hygiene, to unjustifiably intrusive data collection and data mining, to reckless selling of user data to third parties, can expose others to profound and unnecessary harms.

1.2.2. Threats to Fairness & Justice

We all have a significant life interest in being judged and treated fairly, whether it involves how we are treated by law enforcement and the criminal and civil court systems, how we are evaluated by our employers and teachers, the quality of health care and other services we receive, or how financial institutions and insurers treat us.

All of these systems are being radically transformed by new data practices and analytics, and the preliminary evidence suggests that the values of fairness and justice are too often endangered by poor design and use of such practices. The most common causes of such harms are arbitrariness, avoidable errors and inaccuracies, and unjust and often hidden biases in datasets and data practices.

For example, investigative journalists have found compelling evidence of hidden racial bias in data-driven predictive algorithms used by parole judges to assess convicts' risk of reoffending.[2] Of course, bias is not always harmful, unfair, or unjust. A bias against, for example, convicted bank robbers when reviewing job applications for an armored-car driver is entirely reasonable! But biases that rest on falsehoods, sampling errors, and unjustifiable discriminatory practices are all too common in data practice.

Typically, such biases are not explicit, but *implicit* in the data or data practice, and thus harder to see. For example, in the case involving racial bias in criminal risk-predictive algorithms cited above, the race of the offender was not in fact a label or coded variable in the system used to assign the risk score. The racial bias in the outcomes was not intentionally placed there, but rather 'absorbed' from the racially-biased data the system was trained on. We use the term **proxies** to describe how data that are not explicitly labeled by race, gender, location, age, etc. can still function as *indirect but powerful indicators* of those properties, especially when combined with other pieces of data. A very simple example is the function of a zip code as a strong proxy, in many neighborhoods, for race or income. So, a risk-predicting algorithm could generate a racially-biased prediction about you even if it is never 'told' your race. This makes the bias no less harmful or unjust; a criminal risk algorithm that inflates the *actual* risk presented by black defendants relative to otherwise similar white defendants leads to judicial decisions that are *wrong*, both factually and morally, and profoundly harmful to those who are misclassified as high-risk. If anything, implicit data bias is *more* dangerous and harmful than explicit bias, since it can be more challenging to expose and purge from the dataset or data practice.

In other data practices the harms are driven not by bias, but by poor quality, mislabeled, or error-riddled data (i.e., 'garbage in, garbage out'); inadequate design and testing of data analytics; or a lack of careful training and

2. Angwin, et al. (2016). "Machine Bias", *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

auditing to ensure the correct implementation and use of the data system. For example, such flawed data practices by a state Medicaid agency in Idaho led it to make large, arbitrary, and very possibly unconstitutional cuts in disability benefit payments to over 4,000 of its most vulnerable citizens.[3] In Michigan, flawed data practices led another agency to levy false fraud accusations and heavy fines against at least 44,000 of its innocent, unemployed citizens for two years. It was later learned that its data-driven decision-support system had been operating at a shockingly high false-positive error rate of *93 percent*.[4]

While not all such cases will involve datasets on the scale typically associated with 'big data', they all involve ethically negligent failures to adequately design, implement and audit data practices to promote fair and just results. Such failures of ethical data practice, whether in the use of small datasets or the power of 'big data' analytics, can and do result in economic devastation, psychological, reputational, and health damage, and for some victims, even the loss of their physical freedom.

1.2.3. Threats to Transparency and Autonomy

Transparency is an important procedural value that emphasizes the importance of being able to see how a given social system or institution works, as well as being able to inquire about the basis of life-affecting decisions made within that system or institution. So, for example, if your bank denies your application for a home loan, transparency will be served by you having access to information about exactly *why* you were denied the loan, and by whom.

Transparency is importantly related to **autonomy** – the ability to govern the course of one's own life. To be effective at steering the course of my own life (to be autonomous), I must have a certain amount of accurate information about the other forces acting upon me in my social environment (that is, I need some transparency in the workings of my society). Consider the example given above: if I know why I was denied the loan, I can figure out what I need to change to be successful in a new application, or in an application from another bank. The fate of my aspiration to home ownership remains at least somewhat in my control. But if I have no information to go on, then I am blind to the social forces blocking my aspiration, and have no clear way to navigate around them. Data practices have the potential to create or diminish social transparency, but diminished transparency is currently the greater risk because of two factors.

The first risk factor has to do with the sheer volume and complexity of today's data, and of the algorithmic techniques driving big data practices. For example, machine learning algorithms trained on large datasets can be used to make new assessments based on fresh data; that is why they are so useful. The problem is that especially with 'deep learning' algorithms, it can be difficult or impossible to reconstruct the machine's 'reasoning' behind any particular judgment.[5] This means that if my loan was denied on the basis of this algorithm, the loan officer and even the system's programmers might be unable to tell my why—even if they wanted to. And it is unclear how I would appeal such an opaque machine judgment, since I lack the information needed to challenge its basis. In this

3. Jay Stanley (2017). "Pitfalls of Artificial Intelligence Decisionmaking Highlighted in Idaho ACLU Case," *American Civil Liberties Union*. https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case
4. Paul Egan (2017). "Data glitch was apparent factor in false fraud charges against jobless claimants," *Detroit Free Press*.
5. Will Knight (2017). "The Dark Secret at the Heart of AI," *MIT Technology Review*. https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

way my autonomy is restricted. Because of the lack of transparency, my choices in responding to a life-affecting social judgment about me have been severely limited.

The second risk factor is that often, data practices are cloaked behind trade secrets and proprietary technology, including proprietary software. While laws protecting intellectual property are necessary, they can also impede social transparency when the protected property (the technique or invention) is a key part of the mechanisms of social functioning. These competing interests in intellectual property rights and social transparency need to be appropriately balanced. In some cases the courts will decide, as they did in the aforementioned Idaho case. In that case, *K.W. v. Armstrong*, a federal court ruled that citizens' due process was violated when, upon requesting the reason for the cuts to their disability benefits, the citizens were told that trade secrets prevented releasing that information. Among the remedies ordered by the court was a testing regime to ensure the reliability and accuracy of the automated decision-support systems used by the state.

However, not every obstacle to data transparency can or should be litigated in the courts. Securing an ethically appropriate measure of social transparency in data practices will require considerable public discussion and negotiation, as well as good faith efforts by data practitioners to respect the ethically significant interest in transparency.

2. Ethics & Artificial Intelligence

Artificial Intelligence (AI) has the power to drastically reshape human existence. Indeed, the discussion of big data and algorithmic processing in the previous sections shows how some lesser forms of AI already have reshaped our lives. But AI is more than machine learning and algorithms (although that is part of it). Given the increasing involvement of AI in our lives, it is more essential than ever that we have a framework for understanding how to develop AI for the social good. And this is precisely the task that the AI4People initiative set for itself: to understand the core opportunities and risks associated with AI and to recommend an ethical framework that should undergird the development and adoption of AI technologies.[6]

In what follows, we will summarize and synthesize some of the key findings of the AI4People initiative and its foundations for a "Good AI Society".

2.1. Risks & Opportunities of AI

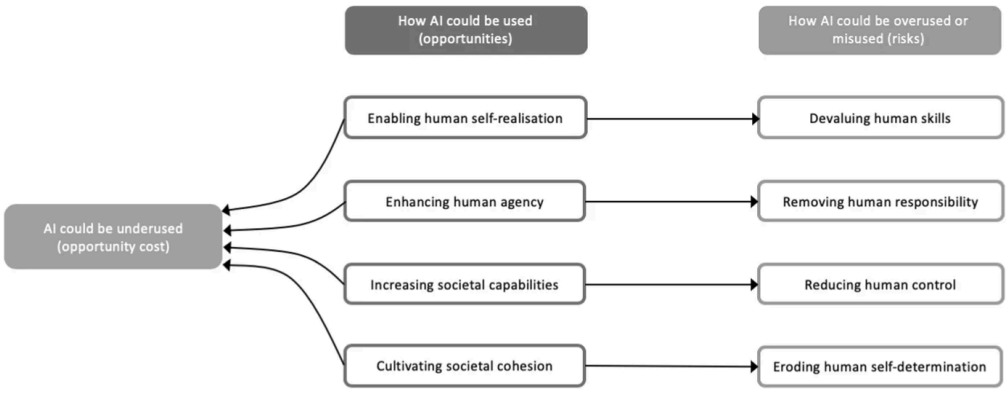
Starting from the position that AI *will* (and already does) have a major impact on society, the focus for ethical investigation is on what sort of impact(s) it will have. Rather than asking *whether* AI will have an impact, the focus is *who* will be impacted, *how* will they be impacted, *where* will we see the impacts, and *when* will we see the various impacts?

To investigate those questions in a more useful way, AI4People identified four broad ways AI may improve or threaten key aspects of society and human existence. They explain their approach thusly:

6. Luciano Floridi, et al. (2018). "AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines*.

> [The four categories] address the four fundamental points in the understanding of human dignity and flourishing: *who we can become* (autonomous self-realisation); *what we can do* (human agency); *what we can achieve* (individual and societal capabilities); and *how we can interact with each other and the world* (societal cohesion). In each case, AI can be *used* to foster human nature and its potentialities, thus creating opportunities; *underused*, thus creating opportunity costs; or *overused* and *misused*, thus creating risks. As the terminology indicates, the assumption is that the *use* of AI is synonymous with good innovation and positive applications of this technology. However, fear, ignorance, misplaced concerns or excessive reaction may lead a society to *underuse* AI technologies below their full potential, for what might be broadly described as the wrong reasons. This may cause significant opportunity costs. It might include, for example, heavy-handed or misconceived regulation, underinvestment, or a public backlash akin to that faced by genetically modified crops (Imperial College, 2017). As a result, the benefits offered by AI technologies may not be fully realized by society. These dangers arise largely from unintended consequences and relate typically to good intentions gone awry. However, we must also consider the risks associated with inadvertent *overuse* or willful *misuse* of AI technologies, grounded, for example, in misaligned incentives, greed, adversarial geopolitics, or malicious intent. Everything form email scams to full-scale cyber-warfare may be accelerated or intensified by the malicious use of AI technologies (Taddeo, 2017). And new evils may be made possible (King et. al, 2018). The possibility of social progress represented by the aforementioned opportunities above must be weighed against the risk that malicious manipulation will be enabled or enhanced by AI. Yet a broad risk is that AI may be underused out of fear of overuse or misuse.

In line with this analysis, they offer us the following graphical representation of the related risks and opportunities. On the back of this graphical representation, we will briefly explore each of the key ideas.



2.1.1. Self-realization without devaluing human abilities

AI may enable **self-realization**: the ability for people to flourish in terms of their own characteristics, interests, potential abilities or skills, aspirations and life projects. Indeed, many technological innovations have done this. The creation of laundry machines liberated people – particularly women – from the drudgery of domestic work. And so, similarly, various forms of "smart" automation present us with the opportunity to free up time for cultural, intellectual and social pursuits.

The fact that some skills will be made obsolete while other skills will emerge (or emerge as newly valuable)

should not be a concern in itself, for that is simply a fact of life. However, this sort of change can be concerning for two reasons: the pace at which it happens and the unequal distribution of benefits and burdens that result. If old skills are quickly devalued or rendered obsolete then we are likely to see significant disruptions of the job market and the nature of employment. For individuals, this could be troublesome as work is often intimately linked to personal identity, self-esteem, and social role or standing. So even ignoring the potential economic harm (which could be accounted for through policy), such disruption raises issues. From a societal perspective, as AI begins to replace sensitive, skill-intensive domains such as health care diagnosis or aviation, there is a risk of vulnerability in the event of AI malfunction or an adversarial attack.

In short, change will occur but the aim is for the change to be as fair as possible.

2.1.2. Enhancing human agency without removing human responsibility

AI provides an ever growing reservoir of "smart agency". In augmenting human intelligence, it will make it possible for us to do more, do it better, and do it faster. However, in giving decision-making over to AI, we risk a black hole of responsibility. One major concern with AI development is the "black box" mentality that sees AI decision-making as beyond human understanding and control. Thus, it is important we think clearly about how much and what sorts of agency we delegate to AI.

Helpfully, the relationship between the degree and quality of agency that people enjoy and how much agency we delegate to autonomous systems is not zero-sum. If developed thoughtfully, AI offers the opportunity of *improving and multiplying* the possibilities for human agency. Human agency may be ultimately supported, refined, and expanded by the embedding of **facilitating frameworks**, designed to improve the likelihood of morally good outcomes, in the set of functions that we delegate to AI systems.

2.1.3. Increasing societal capabilities without reducing human control

AI offers the opportunity to prevent and cure diseases, optimize transportation and logistics, and much more. AI presents countless possibilities for reinventing society by radically enhancing what humans are collectively capable of. More AI may support better coordination, and hence more ambitious goals. Augmenting human intelligence with AI could find new solutions to old and new problems, including a fairer or more efficient distribution of resources and a more sustainable approach to consumption.

Precisely because such technologies have the potential to be so powerful and disruptive, they also introduce proportionate risks. If we rely on AI to augment our abilities in the wrong way, we may delegate important tasks and decisions to autonomous systems that should remain at least partly subject to human supervision and choice. This may result in us losing the ability to monitor the performance of these systems or preventing or redressing errors or harms that arise.

2.1.4. Cultivating societal cohesion without eroding human self-determination

Many of the world's most difficult problems – climate change, antimicrobial resistance, and nuclear proliferation – are difficult precisely because they exhibit high degrees of **coordination complexity**: they can only be tackled

successfully if all stakeholders co-design and co-own the solutions and cooperate to bring them about. A data-intensive, algorithmic-driven approach using AI could help deal with such coordination complexity and thereby support greater cohesion and collaboration. For instance, AI could be used to support global emissions cuts as a means of combatting climate change, perhaps as a means of "self-nudging" whereby the system is set up to monitor and react to changes in emissions without human input, thereby eliminating (or at least reducing) the common problem we face where nations agree to cuts but never carry out the necessary tasks.

Of course, such use of AI systems may threaten human self-determination as well. They could lead to unplanned or unwelcome changes in human behaviors. More generally, we may feel controlled by the AI if we allow it to be used for "nudging" in a variety of areas of our lives.

2.2. Principles of Ethical AI

In light of these opportunities and risks, the AI4People group proposes an ethical framework based on the bioethical principles – the same principles we borrowed for thinking about experimental technology. But they also added a new principle – *explicability* – which becomes particularly important in the context of AI.

The relevance of the Principle of Welfare to AI should be clear from the earlier discussion. AI should be used to improve human and environmental well-being, and it certainly has that potential. But of course, in the process of using AI to improve the public welfare we must be on guard to the variety of harms AI can cause. This includes the harms resulting both from overuse – which may often be accidental – and misuse – which is likely deliberate.

The Principle of Autonomy takes on a new life in the context of AI, as consideration of the opportunities and risks should evidence. Because adopting AI and its smart agency can involve *willingly* giving up some of our decision-making power to machines, it raises the specter of autonomously giving up our autonomy. As such, in the AI context, the principle of autonomy should focus on striking a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents.

The group thus suggests that what is most important in the context of AI is "meta-autonomy", or a **decide-to-delegate model**. They suggest that "humans should always retain the power to *decide which decisions to take*, exercising the freedom to choose where necessary, and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making." And any time a decision to delegate has been made, it should be reversible.

The Principle of Fairness applies to AI in at least three key ways. First, it is suggested that AI should be used to correct past wrongs such as by eliminating unfair discrimination. We know that human beings have biases and that we are often unaware of our biases and how they influence our decision-making. At least in principle, a machine could be immune to the sort of implicit biases that plague human psychology. Second, fairness demands that the benefits and burdens of the use of AI are fairly distributed. And third, we must be on guard against AI undermining important social institutions and structures. For example, the rising use of AI and algorithms in healthcare could be seen, if carried out improperly, as a threat to our trust in the healthcare system.

Finally, the researchers introduce a new **Principle of Explicability**. This principle suggests that the creation and

use of AI must be *intelligible* – it should be possible for humans to understand how it works – and must promote *accountability* – it must be possible for us to determine who is responsible for the way it works. On the group's view, this principle *enables* all the others: To adjudicate whether AI is promoting the social good and not causing harm, we must be able to understand what it is doing and why. To truly retain autonomous control even as we delegate tasks, we must know how the AI will carry out those tasks. And, finally, to ensure fairness, it must be possible to hold people accountable if the AI produces negative outcomes, if for no other reason than to figure out who is responsible for fixing the problem.

Check Your Understanding

After successfully completing this chapter, you should be able to answer all the following questions:

- What is an **augmented social reality** and how does it relate to the Fourth Industrial Revolution?
- What are the main types of benefits and risks associated with **Big Data**?
- In the context of data science, what is a **proxy**? What are its ethical implications?
- What is the value of **Transparency** concerned with? How does it relate to **The Principle of Explicability**?
- What are some of the major risks and opportunities associated with **Artificial Intelligence**?
- What does human **self-realization** involve? How might AI facilitate it?
- In the context of AI, what are **Facilitating Frameworks**?
- What does it mean for a social problem to exhibit a high degree of **Coordination Complexity**? How might AI help with deal with such a problem?
- What is a **Decide-to-Delegate Model**? How does it relate to autonomy?