Certified Ethical Emerging Technologist™ (CEET): Exam CET-110

Certified Ethical Emerging Technologist™ (CEET): Exam CET-110

Part Number: 095029
Course Edition: 1.0

Acknowledgements

PROJECT TEAM

Authors	*Development Support*	*Content Editor*
Marco Meyer	Pamela J. Taylor	Geoff Graser
Theodore Lechterman	Brian J. Sullivan	
Wessel Reijers		
Linda Eggert		

The Principia Advisory Group wishes to thank Sarah Gold, Christine Jacobson, Benjamin Lange, Elisabeth Ling, Roderick Noordhoek, Geoffrey M. Schaefer, Robbie Stamp, Kate Vredenburgh, and Pak-Hang Wong, for their instructional and technical expertise during the creation of this course.

ABOUT PRINCIPIA

Principia is a network that brings together academics, data scientists, practitioners, and thought leaders to strengthen organizational ethics. Our mission is to support organizations to act in the most responsible, principled, and accountable ways with their stakeholders, and in how they consider their interactions with broader society and the environment. We work with the boards and C-suites of our clients to provide rigorous data-led insight, clear strategic direction, and support in implementation. In our technology practice, we work with Silicon Valley's leading players to set up responsible innovation functions and help leaders to navigate new ethical challenges. Learn more at **www.principia-advisory.com**.

Marco Meyer leads a research group on organizational ethics at the University of Hamburg. His research draws on organizational psychology, philosophy, and data science to help organizations avoid misconduct and contribute to society. Marco holds a Ph.D. in Philosophy from the University of Cambridge, and a Ph.D. in Economics from the University of Groningen. At Principia, Marco leads the firm's Innovation and Research, as well as Principia's technology practice. Marco has advised C-suite clients in S&P 500 companies across technology, banking, and professional services on organizational culture, responsible innovation, and ethical decision making.

Theodore Lechterman is a research fellow at the Institute for Ethics in AI at the University of Oxford. His current research investigates the risks and opportunities AI poses for democratic values. He has also written extensively on the ethics of philanthropy and corporate social responsibility. Ted was educated at Harvard and Princeton and completed postdoctoral fellowships at Stanford, Goethe University, and the Hertie School. His work with Principia draws on expertise in political philosophy and applied ethics to help global technology and nongovernmental organizations navigate complex ethical challenges.

Wessel Reijers is a Research Associate at the Robert Schuman Centre, European University Institute in Florence. His research focuses on the impacts of emerging technologies on citizenship, examining cases such as the Chinese Social Credit System. He has published widely on topics including blockchain governance, hermeneutic philosophy of technology, and virtue ethics in responsible innovation. He has also recently published a monograph, Narrative and Technology Ethics, with Mark Coeckelbergh. His work with Principia is driven by questions on how to live well with technology, questions that are increasingly pressing in a world saturated with smart devices, AI, and systems for automated decision-making. Wessel's focus is on turning abstract ideas in ethical and political theory into practical applications for engineers and policy makers.

Linda Eggert is a Technology and Human Rights Fellow at the Carr Center for Human Rights Policy at the Harvard Kennedy School, and a joint Fellow-in-Residence at Harvard's Edmond J. Safra Center for Ethics. Linda carried out her doctoral work in political theory at the University of Oxford. Her research is in moral, political, and legal philosophy. At Principia, Linda's work on normative ethics, human rights, and questions of justice helps guide leading technology companies and nongovernmental organizations through complex ethical terrain.

Notices

DISCLAIMER

While Logical Operations, Inc. takes care to ensure the accuracy and quality of these materials, we cannot guarantee their accuracy, and all materials are provided without any warranty whatsoever, including, but not limited to, the implied warranties of merchantability or fitness for a particular purpose. The name used in the data files for this course is that of a fictitious company. Any resemblance to current or future companies is purely coincidental. We do not believe we have used anyone's name in creating this course, but if we have, please notify us and we will change the name in the next revision of the course. Logical Operations is an independent provider of integrated training solutions for individuals, businesses, educational institutions, and government agencies. The use of screenshots, photographs of another entity's products, or another entity's product name or service in this book is for editorial purposes only. No such use should be construed to imply sponsorship or endorsement of the book by nor any affiliation of such entity with Logical Operations. This courseware may contain links to sites on the Internet that are owned and operated by third parties (the "External Sites"). Logical Operations is not responsible for the availability of, or the content located on or through, any External Site. Please contact Logical Operations if you have any concerns regarding such links or External Sites.

TRADEMARK NOTICES

Certified Ethical Emerging Technologist™ is a trademark of CertNexus. Logical Operations and the Logical Operations logo are trademarks of Logical Operations, Inc. and its affiliates.

Certified Ethical Emerging Technologist™ (CEET): Exam CET-110

About This Course

Mutually reinforcing innovations in computing and engineering are catapulting advances in technological production. From blockchain and artificial intelligence (AI) to gene editing and the Internet of Things (IoT), these advances come with tremendous opportunities for improvement in productivity, efficiency, and human well-being. But as scandals increasingly demonstrate, these advances also introduce new and serious risks of conflict and harm.

Technology professionals now face growing demands to identify and mitigate ethical risks to human rights and the environment, as well as to navigate ethical tradeoffs between qualities such as privacy and accuracy, fairness and utility, and safety and accountability. This course provides the tools to identify and manage common ethical risks in the development of emerging data-driven technologies. It distills ethical theory, public regulations, and industry best practices into concrete skills and guidelines needed for the responsible development of digital products and services. By following the course's practical, problems-based approach, learners will become adept at applying theories, principles, frameworks, and techniques in their own roles and organizations.

Course Description

Target Student

This course is designed for technology leaders, solution developers, project managers, organizational decision makers, and other individuals seeking to demonstrate a vendor-neutral, cross-industry understanding of ethics in emerging data-driven technologies, such as AI, robotics, IoT, and data science.

This course is also designed for professionals who want to pursue the CertNexus Certification Exam CET-110: Certified Ethical Emerging Technologies.

Course Prerequisites

To ensure your success in this course, you should have a genuine interest in ensuring that emerging technologies are ethical, trusted, and inclusive. It may also be helpful if you have an understanding of concepts related to emerging technologies. Practical experience implementing data science, AI, and/or IoT to solve business problems is encouraged but not required. You can obtain the appropriate knowledge of emerging technologies concepts by attending any or all of the following courses:

- *IoTBIZ™ (Exam IOZ-110)*
- *DSBIZ™ (Exam DSZ-110): Data Science for Business Professionals*
- *AIBIZ™ (Exam AIZ-110)*
- *ETBIZ: Emerging Technology for the Business Professional (Exam ETZ-110)*

Course Objectives

In this course, you will incorporate ethics into data-driven technologies such as AI, IoT, and data science. You will:

- Describe general concepts, theories, and challenges related to ethics and emerging technologies.
- Identify ethical risks.
- Practice ethical reasoning.
- Identify and mitigate safety and security risks.
- Identify and mitigate privacy risks.
- Identify and mitigate fairness and bias risks.
- Identify and mitigate transparency and explainability risks.
- Identify and mitigate accountability risks.
- Build an ethical organization.
- Develop ethical systems in technology-focused organizations.

The CHOICE Home Screen

Logon and access information for your CHOICE environment will be provided with your class experience. The CHOICE platform is your entry point to the CHOICE learning experience, of which this course manual is only one part.

On the CHOICE Home screen, you can access the CHOICE Course screens for your specific courses. Visit the CHOICE Course screen both during and after class to make use of the world of support and instructional resources that make up the CHOICE experience.

Each CHOICE Course screen will give you access to the following resources:

- **Classroom**: A link to your training provider's classroom environment.
- **eBook**: An interactive electronic version of the printed book for your course.
- **Files**: Any course files available to download.
- **Checklists**: Step-by-step procedures and general guidelines you can use as a reference during and after class.
- Social media resources that enable you to collaborate with others in the learning community using professional communications sites such as LinkedIn or microblogging tools such as Twitter.

Depending on the nature of your course and the components chosen by your learning provider, the CHOICE Course screen may also include access to elements such as:

- LogicalLABs, a virtual technical environment for your course.
- Various partner resources related to the courseware.
- Related certifications or credentials.
- A link to your training provider's website.
- Notices from the CHOICE administrator.
- Newsletters and other communications from your learning provider.
- Mentoring services.

Visit your CHOICE Home screen often to connect, communicate, and extend your learning experience!

How To Use This Book

As You Learn

This book is divided into lessons and topics, covering a subject or a set of related subjects. In most cases, lessons are arranged in order of increasing proficiency.

The results-oriented topics include relevant and supporting information you need to master the content. Each topic has various types of activities designed to enable you to solidify your

understanding of the informational material presented in the course. Information is provided for reference and reflection to facilitate understanding and practice.

Data files for various activities as well as other supporting files for the course are available by download from the CHOICE Course screen. In addition to sample data for the course exercises, the course files may contain media components to enhance your learning and additional reference materials for use both during and after the course.

Checklists of procedures and guidelines can be used during class and as after-class references when you're back on the job and need to refresh your understanding.

At the back of the book, you will find a glossary of the definitions of the terms and concepts used throughout the course. You will also find an index to assist in locating information within the instructional components of the book. In many electronic versions of the book, you can click links on key words in the content to move to the associated glossary definition, and on page references in the index to move to that term in the content. To return to the previous location in the document after clicking a link, use the appropriate functionality in your PDF viewing software.

As You Review

Any method of instruction is only as effective as the time and effort you, the student, are willing to invest in it. In addition, some of the information that you learn in class may not be important to you immediately, but it may become important later. For this reason, we encourage you to spend some time reviewing the content of the course after your time in the classroom.

As a Reference

The organization and layout of this book make it an easy-to-use resource for future reference. Taking advantage of the glossary, index, and table of contents, you can use this book as a first source of definitions, background information, and summaries.

Course Icons

Watch throughout the material for the following visual cues.

Icon	*Description*
	A **Note** provides additional information, guidance, or hints about a topic or task.
	A **Caution** note makes you aware of places where you need to be particularly careful with your actions, settings, or decisions so that you can be sure to get the desired results of an activity or task.
	**Checklists** provide job aids you can use after class as a reference to perform skills back on the job. Access checklists from your CHOICE Course screen.
	**Social** notes remind you to check your CHOICE Course screen for opportunities to interact with the CHOICE community using social media.

1 Introduction to Ethics of Emerging Technologies

Lesson Time: 2 hours, 15 minutes

Lesson Introduction

The ability to apply ethical reasoning to technical projects requires a firm understanding of some basic concepts about both ethics and technology. This lesson introduces you to these concepts.

Lesson Objectives

In this lesson, you will:

- Identify real-world cases where circumstances raised grave concerns about the ethical use of technology.
- Describe what ethics of emerging technologies is and why it matters.
- Describe ethical reasoning in product development.
- Identify basic features of AI and associated ethical risks.
- Identify common sources of ethical failures.

TOPIC A

What's at Stake

Attention to ethics and emerging technologies is increasing. Technology contains incredible opportunities to improve lives and make the world better. Technology also comes with many risks. In recent years, numerous organizations have handled the power of emerging technology badly, resulting in significant detrimental impact to their businesses and the public at large. In this topic, you will explore some real-world examples of situations where emerging technologies confronted people with difficult ethical decisions.

Why Ethics Matters

We'll begin with some cases that highlight the importance of ethical reasoning. If the technology professionals involved in the cases described in this course had drawn upon ethical reasoning, they could have prevented adverse outcomes, including harm, conflict, financial loss, and embarrassment.

These cases also address common misconceptions about the ethics of emerging technologies; for example:

- That technological progress cannot be controlled.
- That ethical problems in technology have straightforward solutions.
- That technology is ethically neutral.
- That ethics is somebody else's job.

CRISPR Babies: Technology Raises New Ethical Challenges

Technologies often generate moral challenges and controversies by altering people's incentives, creating new opportunities, or redistributing resources. Consider this recent case:

Researcher He Jiankui announced in November 2018 that his team had used ***CRISPR*** technology to edit DNA in human embryos to make them less susceptible to contracting HIV. ***CRISPR technology*** is a powerful tool for editing genomes. It has many possible applications, including in correcting genetic defects, preventing the spread of diseases, and improving crops.

However, its promise also raises ethical concerns. Immediate questions include the safety and long-term consequences of editing the human genome. In addition to this are deeper questions about whether or when we should treat human beings as "customizable." Should we edit genomes to enhance human well-being beyond normal functioning, or only to cure diseases? Standards of health and disease vary widely across time and place. Consider that many people and groups regard homosexuality as a disease, while in some societies, schizophrenia can be considered a gift for communicating with the divine. How, then, should we define normal functioning, well-being, and disease?

Technologies create new possibilities of action. As this case shows, this can generate new moral challenges to which we do not yet have the answers. People working with emerging technologies—perhaps you are one of them—are the first to encounter these questions.

Technological determinism is the view that technological progress cannot be controlled. As a result, ethical judgment and regulation are futile. This view is controversial. Technological developments are the result of human choices, not forces of nature. When individuals face difficult choices, societies have many options for incentivizing (motivating) and disincentivizing (discouraging) different courses of action. Societies also have many options for deciding how to regulate and live with new technologies once they exist.

Additional Reading

For more information about the He case, visit **https://www.scientificamerican.com/article/what-crispr-baby-prison-sentences-mean-for-research/**.

Predpol: Technology Involves Ethical Tradeoffs

Technology can both promote and undermine values. It may even change which values guide us in our everyday lives.

Consider the case of Predpol, a software program using data from public authorities to predict crime rates in certain neighborhoods across the United States. By predicting where crime might strike, it aims to assist departments in allocating resources efficiently. However, this has led to disproportionate prosecution of petty crimes in neighborhoods that are predominantly poor and Black. Predictive policing thereby runs the risk of exacerbating racial inequalities.

Note: For additional information about predictive policing, visit **https://www.themarshallproject.org/2016/02/03/policing-the-future?ref=hp-2-111#.UyhBLnmlj**.

This case shows that ethical challenges in emerging technology are often not about bad intentions or foolish mistakes. They are often about making difficult choices between competing values.

Facebook: Context Matters

Sometimes the values that technology embodies depend on the context of its use. This means we cannot fully understand a technology's potential impact without understanding the different ways in which that technology might be used. This is an issue Facebook has confronted.

Facebook's mission is to give people the power to build community and bring the world closer together. But in 2017, members of the Myanmar military, using the social network, misused that power horrifically. Using Facebook, they carried out a systematic campaign to target the country's mostly Muslim Rohingya minority group, exploiting Facebook's wide reach in Myanmar. This is one of the first widely reported instances of an authoritarian government using the social network against this government's own people. While Facebook took down the official accounts of senior Myanmar military leaders in August 2017, human rights groups and the UN criticized Facebook's response for being too slow and ineffective. The violent campaign went undetected for long enough to allow the anti-Rohingya propaganda to incite mass murder, rape, and the largest forced human migration in recent history.

Note: For more information about the Myanmar incident, visit **https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html**.

Context matters. Creating technology to connect people can be a wonderful thing. But the same technology that brings people together can be used to incite genocide. One of the things this shows is that we cannot discuss technology in a vacuum, but must understand the different contexts in which it might be used, including unintended ones.

While context matters in evaluating technology, this does not mean that technology itself is ethically neutral. Researchers have argued, for instance, that usage of social media alters social norms. In one study, usage of social media was correlated with greater support for freedom of expression and reduced support for privacy.

Additional Reading

For more information about the use of social media changing social norms, see *The Online Citizen: Is Social Media Changing Citizens' Beliefs About Democratic Values?* at **https://link.springer.com**.

Credit Ratings: Democratization or Discrimination?

The impact of technology is often ambivalent, with positive and negative impacts going hand in hand. Yet once a technology is widely used, addressing its adverse impacts can be difficult.

Consider the case of credit scores based on data analytics, such as FICO in the United States. From the 1960s onwards, data-driven credit ratings helped to democratize access to credit, and provide a check against overt discrimination by loan officers. Today, consumer credit scores are widely used for making loan decisions, setting insurance prices, and for hiring decisions. Recently, AI and big data have opened up new data sources for assessing creditworthiness. Research showed that an AI algorithm based on five simple digital footprint variables such as borrower device type (e.g., PC or Mac) or email domain (e.g., Gmail or Hotmail) outperform the traditional credit score model in predicting who is more likely to pay back a loan.

Note: For more information about credit scores and AI, visit **https://www.theguardian.com/money/2020/jan/08/credit-score-default-agencies-report** and **https://www.fdic.gov/bank/analytical/cfr/2018/wp2018/cfr-wp2018-04.pdf**.

This raises several ethical questions: Should Mac users be able to get better interest rates, if they are in general less likely to default than PC users? Does your decision change if you know that Mac users are disproportionately white? More generally: should companies be allowed to make important decisions such as access to credit or jobs contingent on factors that have no inherent connection to what is at stake in the decision?

The impact of credit scoring is ambivalent. On the one hand, it has opened up access for large parts of the population that were previously excluded. On the other hand, within this now much enlarged pool of potentially eligible candidates, credit scoring has the potential to unfairly discriminate. The mechanisms by which credit scoring may discriminate have, however, become more complex and opaque. That generates new challenges for organizations and regulators in avoiding the adverse effects of credit scoring.

ACTIVITY 1-1
Discussing What's at Stake

Scenario
Consider the following questions as you discuss the contents of this topic.

1. Which of the cases discussed do you find most disturbing? Why?

2. In hindsight, sometimes it's easy to identify what went wrong and what should have been done instead. Other times, the complexity of the issues makes it harder to judge what's right or wrong. Which case do you find most difficult to evaluate, and why?

TOPIC B

Ethics and Why It Matters

Now that you have seen some example cases, it's time to define some basic concepts that you will use throughout the course.

What Is Ethics?

Ethics is concerned with the question: "What is the right thing to do?"

When justifying our opinions, decisions, and actions, we typically appeal to values like respect, well-being, or autonomy. Moral values determine how we think we should live, and how we relate to others. They're also at the heart of what we judge to be acceptable or unacceptable, admirable or contemptible, and forgivable or unforgivable.

What moral values are and where they come from are difficult questions. Different kinds of communities—social, cultural, political, and professional—establish conventions about how to act well. These conventions include laws, standards set by professional associations (like IEEE, ISO, and the American Bar Association), codes of conduct in organizations (like the Twitter Rules), informal norms (like tipping in restaurants), and commandments drawn from religious traditions (such as "love thy neighbor as thyself").

You may have noticed that talk of ethics and morality often go together. In fact, many people use these terms interchangeably.

Note: "Ethics" and "morality" are often used interchangeably. "Ethics" is derived from the Greek word "ethos" (character or customary conduct). "Morality" is derived from the Latin word "moralis" (manner, custom).

Is Ethics a Western Concept?

Some approaches to technology ethics draw exclusively on Western ethical theories. This may generate the impression that ethics is a Western concept. This is far from the truth. All human societies have developed sophisticated ethical codes, because, in order to flourish, societies rely on norms of good behavior and good ways of living together.

This course places less emphasis on particular ethical theories from any one tradition. Instead, we explore ethics by investigating ethical issues connected to emerging technologies as they arise today all around the world, and introduce techniques for ethical reasoning that are just as widely applicable.

What Ethics Is Not

While social conventions often provide helpful guidance for how to act, they are not the same as ethics. Following social conventions isn't necessarily the same as "doing the right thing," because conventions may be incomplete, inconsistent, or mistaken.

Determining the right thing to do in any given case requires that we engage in ***ethical reasoning***. Ethical reasoning involves assessing considerations that speak in favor and against different courses of action, and weighing these considerations to form a judgment about what to do. These considerations are what we typically call "reasons." Reasons are the building blocks of ethical decision-making and justification.

For example, the fact that a certain course of action would advance someone's well-being might supply one reason for acting in that way. The fact that the same course of action would violate a policy might supply a reason not to act in this way. A well-reasoned decision would weigh these and other relevant considerations to form a judgment about what to do in this situation.

Moral Philosophy

What makes certain actions or decisions right or wrong? For example, does it depend on what we think a "good" or "virtuous" person would do; on whether we act in accordance with moral duties; or on whether we bring about good consequences through our actions and creations? How do we *know* whether something is right or wrong, and how do we decide what we should do in a specific situation?

Moral philosophy is the study of ethics. It focuses on these different elements of what it means to "do the right thing." These are typically divided into three main areas:

- ***Metaethics*** is concerned with the nature of ethics itself. It focuses on questions like what it is that is "good" or "bad" when we say, "torture is morally bad," and whether moral statements (for example, "torture is wrong") are *true* or *false* just like other statements (for example, "Washington DC is the capital of the U.S.").
- ***Normative ethics*** is concerned with the basic elements of morality. Essentially, its focus is how we should determine what makes actions right or wrong in general. The most popular approaches focus on virtue, duties, and consequences.
- ***Applied ethics*** focuses on concrete issues that pose ethical challenges. Abortion, euthanasia, stem cell research, the environment, war, torture, the corporate world, and—you guessed it—emerging technologies all raise specific questions that require us to *apply* ethical reasoning. Common fields within applied ethics include bioethics, business ethics, engineering ethics, professional ethics, social ethics, and environmental ethics.

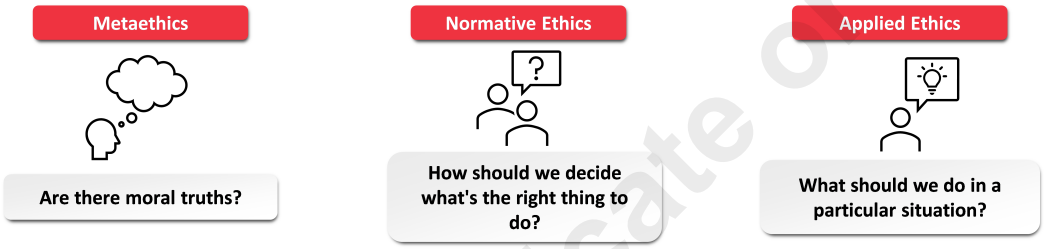


Figure 1-1: Three main elements of moral philosophy.

Research in moral philosophy has also developed several theories and methods to assist in the process of ethical reasoning. You will explore some of these resources in detail later in the course.

What Are Emerging Technologies?

Technology refers to skills, tools, and processes that assist in accomplishing objectives. ***Emerging technologies*** are recent innovations with the potential to transform how we live and interact. Many of today's emerging technologies are data-driven. They include, for example, big data, artificial intelligence, quantum computing, cloud computing, blockchain, Internet of Things (IoT), virtual reality, and gene editing. As you'll see in this course, emerging technologies raise important questions about fundamental ideas like autonomy, moral agency, personhood, and moral psychology, as well as questions about data ethics, which you will explore as it relates to privacy risks.

The current generation of emerging technologies are sometimes credited with bringing about a *Fourth Industrial Revolution*. On this view, the First Industrial Revolution turned on mechanization, the Second on electricity, the Third on computing and communications technology, and the Fourth Industrial Revolution will turn on the integration of digital and physical systems.

One significant feature of the current generation of emerging technologies is that they are mutually reinforcing. For example, big data helps to power AI, which in turn can be used to expand data collection and analysis. The impact of emerging technologies can also grow exponentially, radically transforming economic and social life.

ACTIVITY 1-2
Identifying Ethical Issues

Scenario
Consider the following questions as you discuss the contents of this topic.

1. Think of a hypothetical personal or business situation where someone faced a difficult ethical choice. What made it difficult?

2. Try to think of a time when your judgment about what's right clashed with a law, policy, or social norm. How did you decide what to do?

3. Why are ethical considerations so important when it comes to emerging technologies?

TOPIC C

Ethical Decision-Making in Practice

Now that you've reviewed what ethics is, it's time to look more closely at why it matters in the context of emerging technologies. For this, you will explore ethical reasoning in a real-life product development scenario.

Facebook Portal: A Case of Racial Bias in AI

We'll zoom in on the case of the AI-powered Portal Platform developed by Facebook.

Portal is a piece of software that provides users with a smart camera, using advanced computer vision to dynamically frame shots during video calls. Instead of having to move your device around yourself, the camera automatically follows you while you speak, using advanced facial recognition tools.

During a pre-launch test at Facebook, Lade Obamehinti—who was then in charge of technical strategy—noticed something strange. While Lade was talking, the camera focused on Lade's colleague instead of Lade. Lade is Black. Her colleague is white.

After Lade investigated the issue, she discovered that the software was trained on a non-representative dataset. As a result, Portal had a racial bias built into the product, which caused it to effectively prioritize white faces over the faces of people of color. The bias was not intentional. The problem was that the data that was convenient for the engineers to use had been non-representative.

Lade did not see this as a simple technical bug, but as a moral challenge. Facebook risked valuing convenience during the development process over basic fairness and inclusion, exacerbating racial inequality.

Lade not only sought to resolve this problem in this particular context, but to learn from this experience in designing future products. She did this by developing a framework for inclusive AI. Teams at Facebook now use this framework to reduce bias in their products.

Note: For more about inclusive AI and Facebook, visit **https://tech.fb.com/building-inclusive-ai-at-facebook/**.

Portal's Tech Specs: An Introduction to AI

The problem that Lade identified is not unique to Facebook Portal. To see why this is a more systemic phenomenon, consider the following details about the technology that drives Portal's smart camera:

- At the most general level, Portal works by using ***artificial intelligence (AI)***. This is the capacity of machines to exhibit human-like intelligence. ***Narrow AI*** is designed to solve particular problems and is the most common variety of AI currently in use. Google's AlphaGo, which defeated the world champion Go player, is an example of narrow AI. ***General AI*** is designed to be able to solve any problem. General AI does not currently exist but is actively under research and development by companies like DeepMind and OpenAI. More speculatively, there is the notion of superintelligence, or ***super AI***, which would outsmart human beings.
- AI often interacts with the physical environment; for instance, through devices and systems in our homes. The use of AI to create smart physical devices in our environment is also known as ***ambient intelligence***.
- The narrow AI used in Portal makes use of ***machine learning (ML)***. This is a process whereby large quantities of data are processed by algorithms to make intelligent decisions about the behavior of the software in a certain environment. If the environment is your kitchen, then Portal's ML guides the camera to follow you around in the room.

- Portal uses facial recognition technology, which is powered by a type of ML called ***deep learning***. This ML technique uses layers of information of increasing complexity to make decisions, which constitutes a training model. It processes first the simple aspects of images (shapes), then more complex aspects (eyes, noses), and finally the most complex aspects (entire faces). In order to learn how to recognize faces, deep learning processes use huge libraries with pictures of faces.

The problem that Lade Obamehinti had stumbled upon was that the huge library of faces that had been used by Portal to train its AI contained mostly faces of white people. This made the AI very good at spotting and following white faces, but much worse at doing the same with people of color. Algorithmic bias is not limited to Portal. It is a common problem in AI that affects many applications.

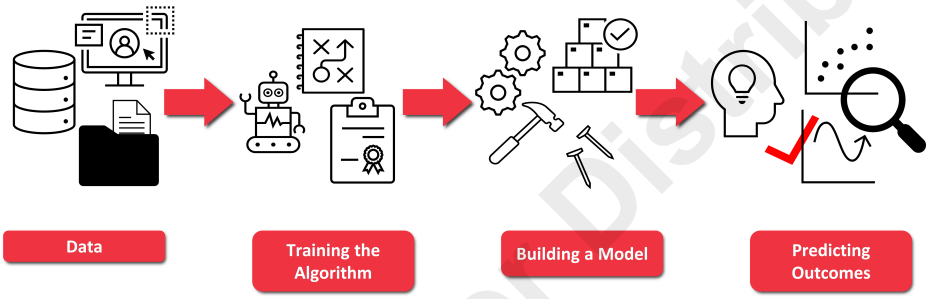


Figure 1-2: The machine learning process.

Ethical Risks in Product Development

Risk is a broad concept that refers to the likelihood of experiencing any negative outcome. An ***ethical risk*** is a risk of doing wrong, harming others, or failing to live up to certain values. One special ethical risk involves failing to take ethics into account at all.

As you have seen, some of these negative impacts that result from ethical risks are headline-grabbing. But it is important not to overlook the ethical dimension in day-to-day decisions. For instance, designing a software application in such a way that it uses energy inefficiently might have a negative impact on the climate.

This course focuses primarily on ethical risks in emerging technology applications, which we call ***products***. Products cover both commercial and noncommercial applications of a given technology.

One major source of ethical risk is that a decision may harm people, or that it might benefit some people at the expense of others. Thinking about these kinds of risks is complicated by the possibility that some harms might be more or less severe, or that some decisions might affect smaller or larger numbers of people.

In the case of Portal, Lade Obamehinti identified a product feature that would benefit white users while excluding users of color. What was the right thing to do in this case? Bringing to market a product that was effectively racist was, of course, not an option.

The Price of Ethical Failure

Facebook was lucky that Lade identified the issue in Portal before the product was released. If the product's faults had been exposed after release, this would have had unacceptable and troubling effects on many people, Facebook would have suffered business losses, and the company might have exposed itself to legal penalties.

But Facebook still paid a price. If organizational policies and individual employees had been more aware of the risks involved in using non-representative datasets, Facebook could easily have prevented the problem. This would have saved the company significant resources and prevented some reputational damage. Developing ethics safeguards is not always easy or cheap, but—as this case shows—it is better than paying for the failure to do so further down the line.

When ethical failure occurs, it also signals a lost opportunity for making the world a better place. Most people and organizations intend to do good for society. Tragically, when ethical failure occurs, resources that could have been spent on benefiting society must instead be diverted to correct mistakes.

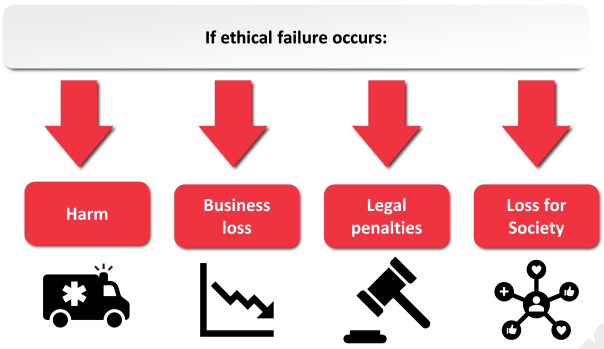


Figure 1-3: The price of ethical failure can be enormous.

ACTIVITY 1-3
Discussing Ethical Decision-Making

Scenario
Consider the following questions as you discuss the contents of this topic.

1. In the Portal case, what everyday design decision posed a serious ethical risk?

2. Have you been involved in any projects where the use of technology introduced ethical risks? How can you guard against these risks?

TOPIC D

Causes of Ethical Failures

As you have seen, ethical dilemmas can affect many different types of companies in big and small ways. It is safe to say that every company has or will face ethical challenges while developing or using emerging technologies. The ability to recognize when the ethical breakdowns can happen and why they happen is an essential tool for anticipating potential pitfalls and avoiding their effects on the business. In this topic, you will identify common sources of ethical failures.

Ethical Failures Due to Bad Intent

News coverage tends to attribute failings to bad intent or gross neglect for societal interest. We can understand bad intent in light of the model of ethical decision-making. For instance, the British consulting firm **Cambridge Analytica** collected personal Facebook data without people's consent. The data was collected through an app running on Facebook. It consisted of questions used to build psychological profiles. The app also collected the personal data of users' Facebook friends. The U.K. Information Commissioner's Office found that Cambridge Analytica's abuse of Facebook user data was intentional. In fact, the abuse of the data of American voters was at the heart of the analytics service Cambridge Analytica provided to political campaigns, including the 2016 presidential campaigns of several U.S. political candidates.

Note: For more information, visit **https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html**.

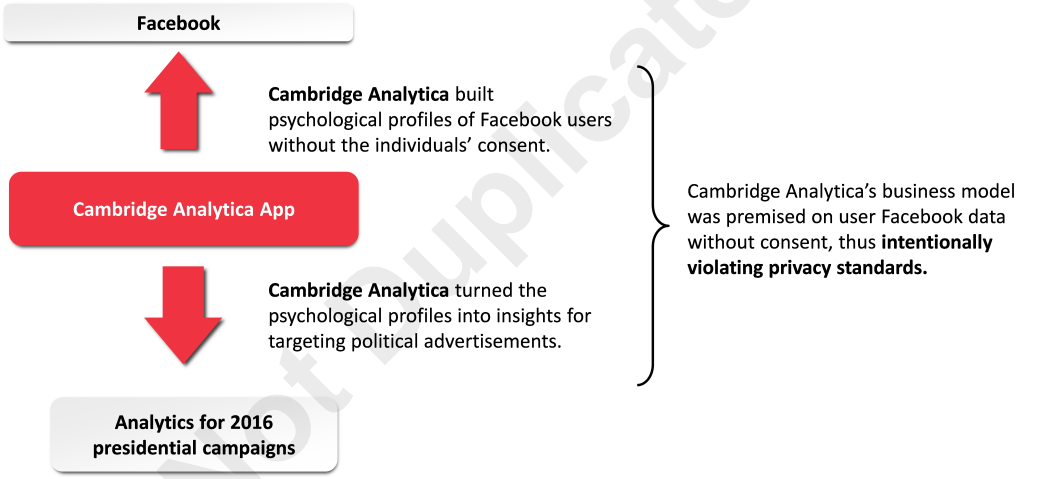


Figure 1-4: Ethical failure due to bad intent.

Beware of Ultimate Attribution Error

Psychologists have suggested that we tend to attribute others' negative behavior to bad intentions while excusing our own bad behavior because of the circumstances (*ultimate attribution error*). As a result, we may overestimate the extent to which malevolence drives ethical failures. On the flipside, this means that being well-intentioned ourselves is not sufficient to prevent ethical failure. While actors with bad intentions are sometimes responsible for ethical failures, other causes are more common.

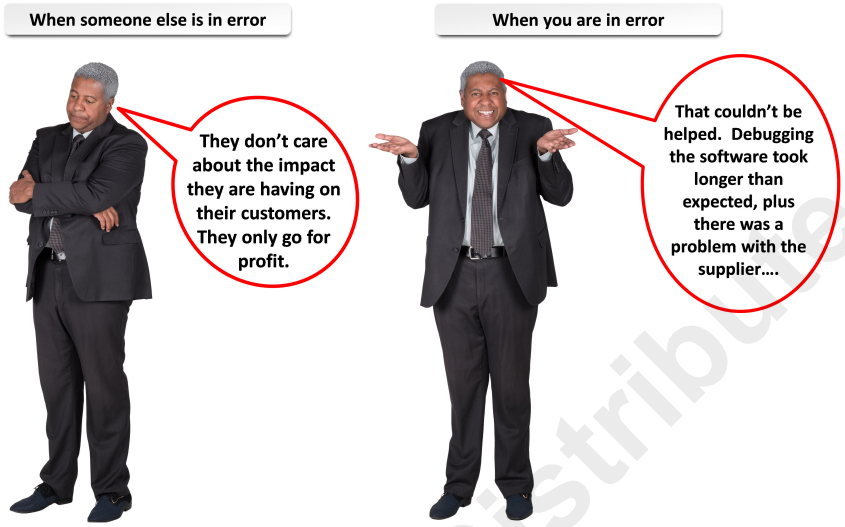


Figure 1-5: Ultimate attribution error.

Other Causes of Ethical Failure

In addition to intent, there are at least three other root causes of ethical failure:

- **Lack of awareness**: Potential negative impacts of emerging technologies are never considered.
- **Poor analysis**: Conflicting ethical considerations are poorly resolved.
- **Poor governance**: The development or use of emerging technology is poorly governed.

Each of these root causes is discussed in the next few sections.

Awareness

Some ethical issues are like birds: They can be difficult to spot, even when hidden in plain sight. This applies particularly for the ethical impacts of emerging technology, which are not yet well understood. Consider the case of **Facebook Portal**. Through the intervention of one employee, the company noticed that the smart camera was less able to recognize people of color than white people. While Facebook narrowly avoided the consequences of this ethical failure, the fact that the problem went undetected until just before shipment is troubling. The case illustrates that sometimes the biggest challenge in preventing ethical issues is to spot them in the first place.

Note: For more information, visit **https://www.scu.edu/ethics-in-technology-practice/overview-of-ethics-in-tech-practice/**. Shannon Vallor, one of the authors on this site, has coined the phrase that ethical issues are like birds.

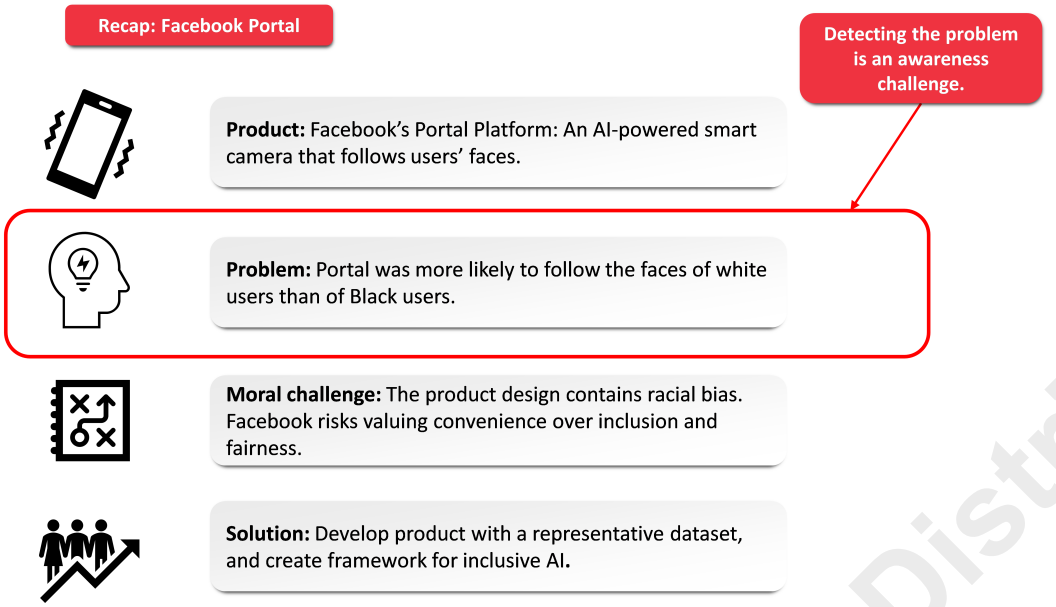


Figure 1-6: Awareness can be a root cause of ethical failure.

Analysis

Some ethical issues are complex. Addressing one dilemma often creates another. And a solution to an ethical issue may conflict with other ethical principles. Consider the case of **Apple vs. the FBI**. In the wake of the December 2015 terrorist attack in San Bernardino, CA, a federal judge asked Apple to provide technical assistance to the FBI in accessing the information on the suspect's iPhone, hoping to discover additional threats to national security. To decide whether to comply with the court order, Apple had to decide how to weigh considerations of safety and security against its users' privacy. Apple decided to send engineers to advise the FBI, but refused to comply with the court order to bypass the phone's security measures. A number of major tech firms filed amicus briefs in support of Apple, while the White House and Bill Gates stood behind the FBI. The tradeoff Apple faced was difficult, and the case remains as controversial today as it was at the time.

Note: For more information, visit **https://www.scu.edu/ethics/focus-areas/business-ethics/resources/apple-vs-fbi-case-study**.

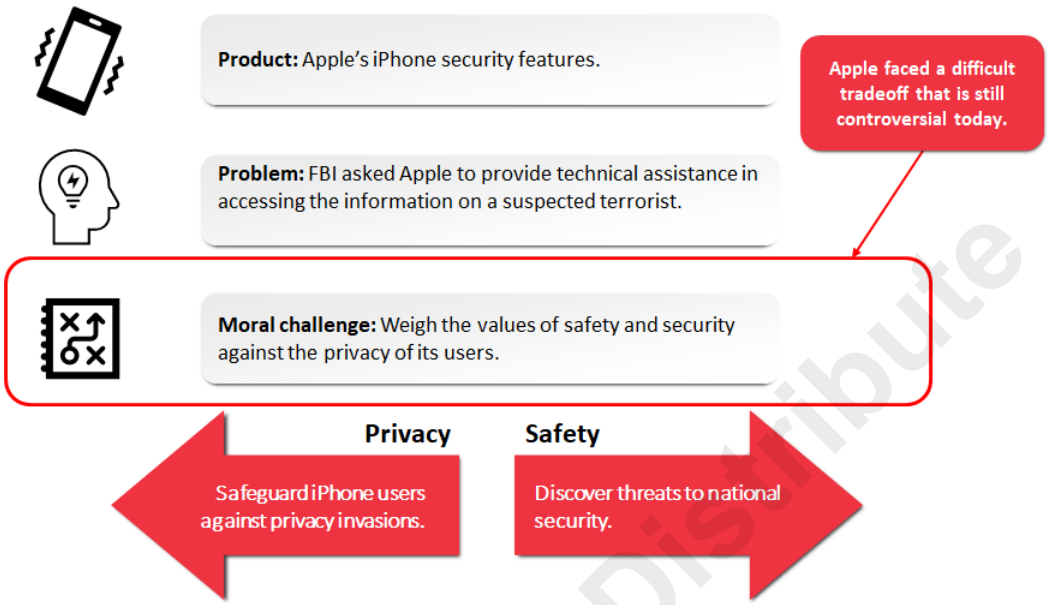


Figure 1-7: Analysis (or lack of analysis) can be a root cause of ethical failure.

Governance

Governance is the system of rules, practices, and processes by which a firm is directed and controlled. It answers the question: Who decides what, and how?

Many technology firms rely on metrics to govern the development and use of emerging technology. While metrics can be a powerful lever of innovation, over-reliance on metrics can lead to problematic outcomes. **YouTube's recommendation algorithm** aims to direct users to content they will like based on their previous viewing habits and searches. This seemingly innocuous feature has given rise to numerous scandals. YouTube has been found to promote terrorist content, extreme hatred, and conspiracy theories. Research by DeepMind has shown that feedback loops in recommender systems can give rise to "echo chambers," which can shift a user's worldview. Guillaume Chaslot, who worked on YouTube's artificial intelligence recommendation engine, suggests that the root cause of these failures is that the algorithm's developers were driven by bad incentives. Software engineers were given a single metric: to increase the time that people spend on YouTube. Misinformation, violent content, and divisive content all drive engagement. Since the product teams working on the recommendation engine have been incentivized to single-mindedly pursue maximize engagement, other values got predictably side-lined.

Note: For more information, visit **https://www.wired.com/story/the-toxic-potential-of-youtubes-feedback-loop/** and **https://deepai.org/publication/degenerate-feedback-loops-in-recommender-systems**.

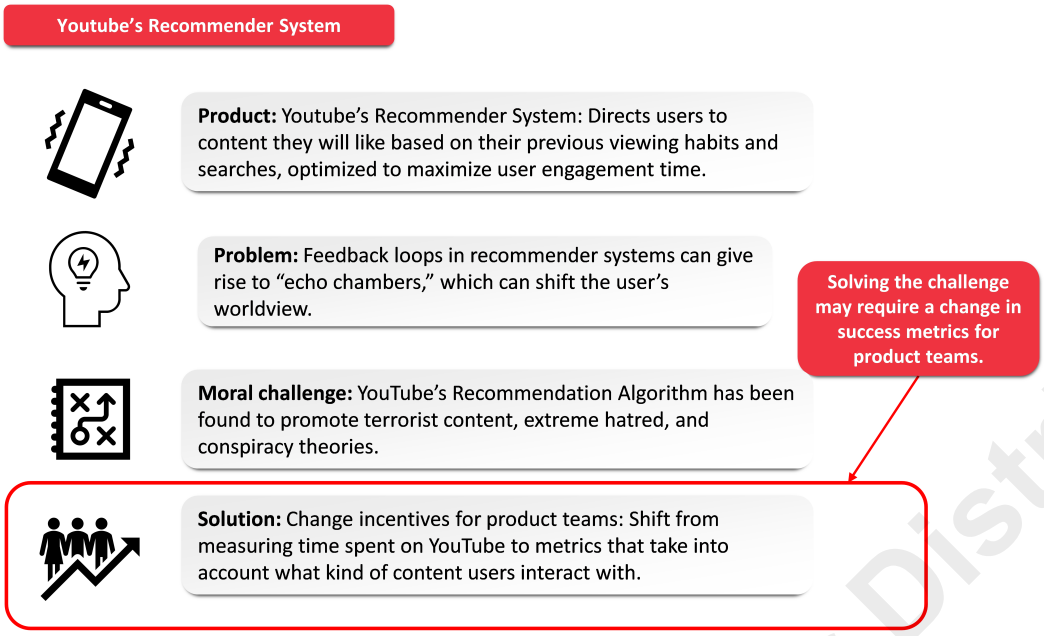


Figure 1-8: Governance (or lack of governance) can be a root cause of ethical failure.

Tech for Good

So far, we have focused on the risk of ethical failure. But "doing the right thing" does not simply mean avoiding bad outcomes; sometimes it can mean trying to do more good. Each of the root causes of ethical failure can equally be leveraged for doing good. Some companies are even formed around the intent to realize a social purpose. For instance, **GoodSAM** is an app that enables users to alert both the emergency services and any nearby registered responders in GoodSAMs network of qualified early responders to provide first aid in case of emergency such as cardiac arrest.

	Note: For more information, visit **https://www.goodsamapp.org/**.

Lumkani uses IoT technology for detecting fire in slums before they cost lives. Shack fires are a serious threat in South Africa. Since most of cooking in slums is done on the open fire, smoke detectors do not work. Instead, Lumkani has built a device that uses sensors to detect heat increase. When a fire is detected, the device notifies everybody within a 60-meter radius via SMS audio alarms.

	Note: For more information, visit **https://lumkani.com/**.

Engineering activism is closely related to the tech for good concept. This is the rising trend of engineers focusing more effort on social justice issues such as environmental ethics and sustainability.

ACTIVITY 1-4
Identifying Causes of Ethical Failures

Scenario
You have recently joined Rudison Technologies, a tech company that specializes in offering consulting and technical services to other companies. A big part of your job responsibilities will be to assist client companies in ensuring their AI and IoT products are ethically designed and distributed. Prior to working with any customers individually, your manager has asked for your participation in a seminar that discusses ethical failures in technology, as well as the Tech for Good movement.

1. Research recent examples of ethical failures.
   a) Find a recent case of an ethical failure involving an emerging technology in the news.

   Note: To find examples of ethical failures, you might search popular news sites for terms like "tech ethics scandal," "social media ethics," "algorithmic bias," or "ethical issues emerging technologies."

   b) Read about the case, and identify the likely root cause(s) of the ethical failure.
2. Present your findings to the rest of the class.
3. **Can you think of an example of a product using emerging technology that uses tech for good?**

Note: Some suggested phrases to search include "tech for good" or "tech startup benefit company."

Summary

In this lesson, you identified basic concepts, theories, and challenges related to ethics and emerging technologies. This information provides a solid foundation for the frameworks, processes, and strategies presented in the rest of the course.

Can you provide an example of ways that you employ ethical reasoning in your day-to-day life?

You looked at quite a few different cases where ethical issues arose surrounding the use of emerging technologies. If you have witnessed any similar situations at your workplace, what do you think were the underlying causes of the ethical issues?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

2 Identifying Ethical Risks

Lesson Time: 2 hour, 15 minutes

Lesson Introduction

Emerging technologies raise questions not only about what we are able to do, but also about what we *should* do. The track record of companies and other organizations identifying and addressing ethical issues before they have a negative impact on society is not impressive, as we have seen. Technologists play a critical role in addressing ethical issues. They are often the first who are in a position to identify ethical issues during product development, and they can play a key part in reasoning through and resolving these issues.

This lesson begins by distinguishing different types of ethical reasons. Distinguishing rights, values, and interests will help you address ethical issues appropriately. Next, this lesson addresses stumbling blocks for ethical reasoning, including myths about ethics and reasoning fallacies. Also covered is an explanation of how ethical risks arise in the product development lifecycle, as well as ways of identifying ethical risks. You will also focus on practical tools you can use with your team and explore the critical roles of regulation, standards, and human rights for identifying ethical risks.

Lesson Objectives

In this lesson, you will:

- Describe different types of ethical reasons.
- Describe common sources of skepticism about ethics and possible responses, as well as common fallacies and mistakes in ethical reasoning, and how to avoid them.
- Describe how to anticipate and identify ethical issues that can occur during product development.
- Describe common tools used to identify ethical risk.
- Describe how regulations, standards, and human rights documents can help in identifying ethical risks.

TOPIC A

Ethical Reasons

This lesson focuses on identifying ethical risks so that you can develop a plan to mitigate the risks. To begin, you will examine some of the basic reasons that play a role in identification and resolution of these types of risks.

What Is Ethical Reasoning?

The most important part of identifying ethical risks is to ensure that you employ sound ethical reasoning as you examine products and processes. ***Ethical reasoning*** is the process of identifying and evaluating reasons for action. ***Ethical reasons*** are considerations that count in favor of, or against, a certain course of action.

Ethical reasoning helps you to:

- Make decisions consciously and reflectively.
- Justify actions in terms that others can understand and accept.
- Avoid bad decisions due to bias or fallacies.

Types of Reasons

There are three main categories of reasons that play a common role in identifying and resolving ethical risks: rights and their corresponding duties, certain kinds of interests, and values:

- ***Rights*** are entitlements to act or be treated in certain ways. They generate weighty reasons, often amounting to obligations, for ethical decision-making.
- ***Interests***, particularly economic interests, typically generate reasons that reflect concern for tangible benefits, such as jobs, money, or economic opportunity.
- ***Values*** generally present reasons for respecting what people find desirable, such as autonomy, spirituality, or fairness.

It is important to bear in mind that the lines between these categories are blurry. For example, certain interests—like an interest in not being assaulted—are also rights; and certain values, like fairness, can generate important obligations much like rights do. The hierarchical categorization presented here merely serves to illustrate that different kinds of reasons often have different roles in moral reasoning. For example, reasons based on rights and duties are generally significantly weightier than reasons based on general values or economic interests.

Rights and Duties

Reasoning about what to do often includes thinking about rights and duties. One view of rights is that they describe what we can reasonably expect other people, and society, to do to protect our interests. Many rights protect basic needs and interests, like our right to life, bodily integrity, free speech, and privacy. The U.N. Universal Declaration on Human Rights contains a list of rights that are commonly regarded as fundamental. Rights often correspond to ***duties***. For example, your *right* to bodily integrity corresponds with your doctor's *duty* not to operate on you without your consent.

Rights and duties function as guardrails in situations that require trading off competing reasons. Rights are sometimes grounded in the notion of human dignity, to explain why there are certain things we must never do to others, or make others do, even if this could benefit a larger number of people. For example, your right to life and bodily integrity protects you from having your organs harvested by your doctor, even if your organs could save several other people's lives. Correspondingly, your doctor has a *duty* to respect your right to life and bodily integrity.

Rights and corresponding duties provide strong reasons against actions that might transgress those rights. For example, if a particular product would violate people's rights, this provides a strong reason—perhaps even amounting to a duty—for rethinking how the product might be developed differently. For example, the right to privacy imposes a duty on healthcare providers to protect our health data, even when making this data available to researchers could lead to faster advances in drug development.

Sometimes, different people's rights will clash, and moral reasons will pull in different directions. In these cases, the aim of moral reasoning is to work out which rights, and which duties, are strongest.

Interests

Interests are possessed by people and groups and by other sentient beings such as animals. To have an economic interest in something is to be concerned for a tangible benefit, such as jobs, money, or economic opportunity. Often, several people have an interest in the same economic benefit, and their interests need to be traded off against each other. For instance, platforms like Google Play or Apple Store distribute apps to users for a share of the revenue. App developers have an interest in decreasing the share the platforms take, while the platforms have an interest in keeping their share high.

Interests also come into play in discussions of broad values like privacy, dignity, and liberty. For instance, the interests that social media users have in the liberty to express themselves can sometimes clash with the interests of other users in maintaining their dignity, which can be harmed by hateful speech.

Values

Values reflect conditions that are desirable or believed to be desirable. There is broad consensus about many values at a general level, such as fairness, knowledge, or autonomy. But agreement at such a general level masks disagreement about the specific meaning of each value, and the weight of each value when values conflict. In addition, there are some values that are not universally shared. Spirituality and curiosity are examples of values that some people find extremely desirable, while others do not.

Once rights and basic interests have been accounted for, it is useful to engage with groups potentially affected by a product, to understand which values matter to them in a given context. For example, two app users may possess the same interests in affordability, reliability, and functionality, while holding different viewpoints about the kind of privacy protections that are important. This kind of engagement is always a work-in-progress, because our values evolve, not least in reaction to new technologies.

ACTIVITY 2-1
Discussing Ethical Reasons

Scenario
Consider these questions as you discuss the content presented in this topic.

1. Can you describe a situation when different ethical reasons came into play and conflicted with one another?

2. In your opinion, what are the values that are widely agreed on, and which are less so?

TOPIC B

Stumbling Blocks for Ethical Reasoning

This topic covers common stumbling blocks for identifying and addressing ethical issues. You will first focus on myths about ethics, and then consider some common reasoning fallacies.

Myths About Ethics

It is worth being aware of some of the assumptions that can hamper identifying and addressing ethical issues. These include:

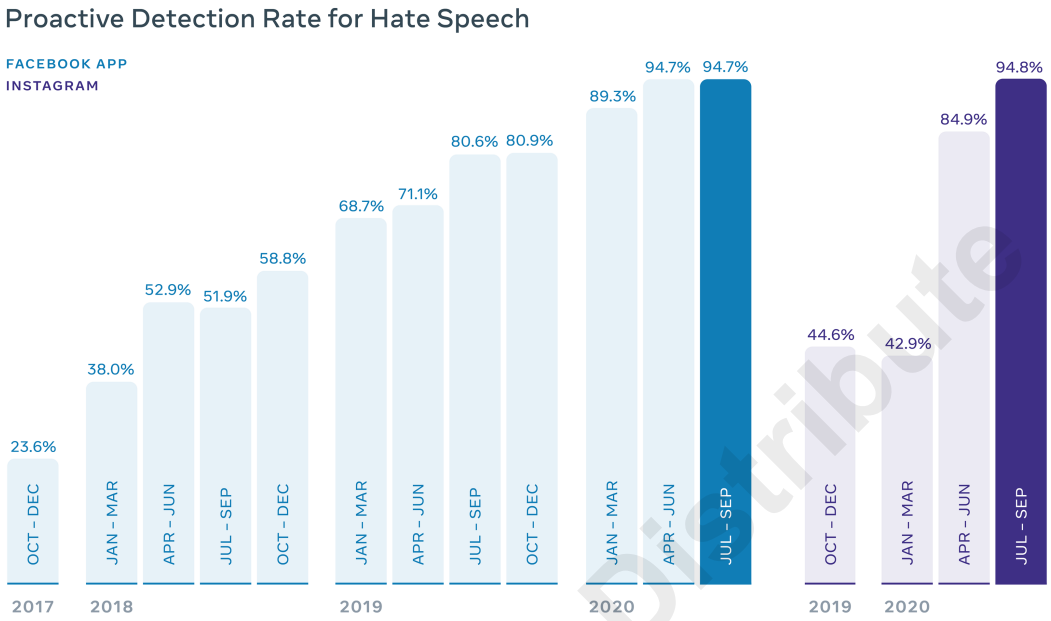
- **Technical solutionism**: The assumption that technology alone suffices to solve any ethical issue.
- **Moral relativism**: The view that ethics is completely subjective, and that trying to identify the right thing to do is therefore pointless.
- **Moral righteousness**: The attitude that all who disagree with one's own ethical position are wrong.

Let's look at each of these in more detail.

Technical Solutionism

Technical solutionism assumes that technology is the best way to solve any ethical issue. Emerging technologies do not only create ethical challenges, but they can also be an important part of the solution. For instance, Facebook has made significant progress in building algorithms that detect hate speech. Facebook started reporting the accuracy of these algorithms in Q4 of 2017. Back then, the algorithms were able to detect 25% of hate speech that was eventually removed from the platform. By the summer of 2020, improved algorithms were able to detect 95% of speech that was eventually removed from the platform.

Note: For more information, visit **https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/**.



Source: https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/

Figure 2-1: Tracking Facebook's fight against hate speech.

Such advances in putting AI to use to *solve* an ethical issue are impressive. But it would be a mistake to assume that technology can solve *any* ethical issue. In fact, the Facebook example shows why technical solutionism is mistaken: In order to train algorithms, Facebook has to first define standards for hate speech, and it must continuously revise and refine them. Effectively operationalizing the notion of hate speech in a way that captures the nuance, history, and changing cultural norms requires a good deal of ethical reasoning. Defining these standards, as Facebook well knows, requires input from many global experts and stakeholders.

Technical solutionism is right that technology can play a significant role in implementing solutions to ethical challenges. But, for the time being, humans cannot pass on the task of reasoning through ethical challenges.

Additional Reading

To dive deeper into this subject, consider reviewing the following:

- The Facebook Community Standards Enforcement Report for Q3 2020: **https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/**
- A Politico article on the dangers of automated content moderation: **https://www.politico.eu/article/facebook-content-moderation-automation/**
- A Verge article on the hardships of human content moderators for Facebook: **https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona**

Moral Relativism

Moral relativism assumes that what it means to do the right thing varies from person to person, or from community to community. Ethics, on this view, is hopelessly subjective—There is no objective value framework to determine whether decisions are right or wrong.

Moral relativism makes it seemingly easy to handle disagreements about ethical choices. Many ethical risks are controversial. For example, should we prioritize privacy through end-to-end

encryption or provide safety by screening chats for attempts to spread misinformation, which requires unencrypted communication? Such tradeoffs are difficult to make, and people passionately disagree about the right course of action.

But this does not mean that there can never be right answers, and that ethical reasoning about emerging technologies is futile. Some arguments about what we ought to do are supported by stronger reasons than others. While it may not always be obvious what "the right thing to do" is, or whether there is only one right answer in any given case, it is often possible to differentiate better and worse courses of action, and to identify the path of action for which there are the strongest reasons. Contrary to what some moral relativists might suggest, moral disagreements should make us take ethical reasoning *more*, not less, seriously.

Relativism is right that we should be tolerant of difference. After all, often there are multiple valuable perspectives on any difficult ethical issue. But, taken seriously, relativism commits us to the view that there is no right or wrong, because different people hold different views, all of which are equally valid. It is therefore important to draw a line between *relativism* and ***pluralism***, which is the view that *anything* goes and the view that there may be *multiple* valid perspectives. To tolerate different views, we need not take the relativist position that there simply is no one right thing to do.

Moral Righteousness

The flipside of the relativist position is moral absolutism, which can take the form of ***moral righteousness***, or the view that everyone who does not agree with the ethical position you hold to be true is wrong, and perhaps therefore also a bad person. The danger of moral righteousness is that it might make you unwilling to listen to people who take a different view. The morally righteous tend to stick to like-minded people and despise those who disagree.

Because ethical reasoning depends on considering different reasons and arguments, it requires a certain degree of open-mindedness and intellectual humility; that is, to recognize that we might have been wrong about our moral judgements when others present us with better reasons for different judgements.

Reasoning Fallacies

Fallacies are a common source for failures of moral reasoning. Fallacies are general patterns in which bad arguments fall. Here are a few of the most common types of fallacies:

- **Overgeneralization and stereotypes**: Stereotypes about people are an example of ***overgeneralization***, which are assumptions about a group of individuals, or a broad range of cases, based on an inadequate sample. Examples include statements like:
  - "Computer scientists are shy and nerdy."
  - "Men are bad at learning languages."
- **Slippery slope**. A ***slippery slope*** is a type of argument that claims an otherwise permissible action would lead to a chain reaction with ultimately catastrophic consequences. It is important to scrutinize slippery slope arguments carefully, because while some slippery slope arguments are good arguments, often there is little evidence that an alleged slippery slope will indeed occur. ("If we allow preimplantation genetic diagnosis, we'll end up designing babies according to parents' wishes.")
- **Appeal to authority**: Sometimes, we try to support our reasons by appealing to respected authorities. But even authorities can be wrong. And people who are authorities on one subject are not authorities on another. Being an authority doesn't automatically make one right, and sometimes following authorities can be wrong.

Additional Reading

To dive deeper into this subject, consider reviewing the following article by Elena Sgarbossa: "Heads I Win, Tails You Lose": Logical Fallacies and Ethics in Everyday Language at **https://translationjournal.net/journal/38fallacies.htm**.

ACTIVITY 2-2
Discussing Ethical Reasoning Stumbling Blocks

Scenario
Consider the following questions as you discuss stumbling blocks to ethical reasoning.

1. How would you respond to a colleague who insists that ethics is "just a matter of personal opinion"? Try to think of this in the context of a business need.

2. How would you respond to a colleague who wonders whether more effective technology should make ethical reasoning obsolete?

TOPIC C

Identify Ethical Risks in Product Development

Ethical reasoning begins with anticipating and identifying ethical risks. This is more challenging than you might think. Ethical risks arise in many different domains, from privacy to sustainability. This part of the lesson explores ethical risks in the product development lifecycle. It also introduces a particular product and product team, RudiBrace, a wearable device that is used in the workplace.

Ethical Risks in the Product Development Lifecycle

Ethical risks take different forms depending on where teams are in the product development lifecycle.

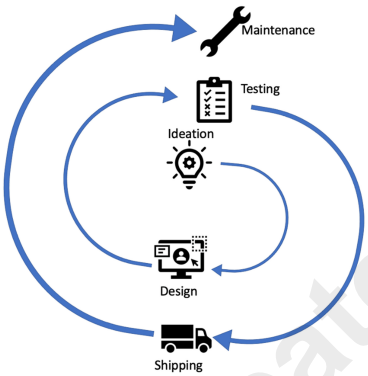


Figure 2-2: The product development lifecycle.

During the **ideation and the prototyping phase**, ethical risks are difficult to anticipate, because the product is in an early stage. Product teams benefit most from engaging with stakeholders and conducting ethical foresight exercises to shape the product in accordance with high ethical standards. The ideation and prototyping phase presents excellent opportunities for considering proactive ways of resolving social problems or promoting human values.

During product **development and testing**, product teams can begin to identify ethical risks on stakeholders but lack data about impacts of the product if deployed at scale. Product teams can leverage user research to screen for potential ethical risks, with a focus on whether or not any particular groups are adversely affected by the product.

When **getting ready to ship**, product teams can focus on how to shape communication about the product as well as permissible use policies to address potential ethical risks that do not have a technical solution but require to educate users, restrict access, or limit functions of certain applications.

During **maintenance**, product teams can collect data about the actual impacts of their product at scale. Ethical foresight exercises and ethical impact assessments conducted in earlier phases can form the basis for designing meaningful metrics. In addition, organizations should conduct grounded research to discover impacts they have not anticipated.

Why Are Ethical Risks Often Overlooked?

There are many reasons why ethical risks raised by emerging technologies are missed. Thinking through the ethical implications of your work isn't easy, and often involves a great deal of

uncertainty about potential outcomes. That said, there are a few common reasons worth bearing in mind:

- **Failure to consider key stakeholders' rights, interests, and values**. Recall the case of Facebook Portal. One issue was that the smart camera that Facebook had developed tracked people of color less reliably than white people because the dataset included mostly white people, failing to represent people of color.
- **The assumption that ethical risks are someone else's problem**. Very few (if any) people working in organizations have ethics as their primary responsibility. Rather, people have responsibility for a certain aspect of a product or a certain technical area of expertise. Ethical risks often arise between and across areas of responsibility. Since ethics is nobody's explicit responsibility, you can easily fall into the trap of assuming that someone else is going to take care of it.
- **The lack of clear and consistent practices of anticipating and identifying ethical risks**. Teams working on emerging technologies often work in a high-pressure environment, having to meet demanding production targets. Considering the ethical implications of this work can seem like a time-consuming distraction. As a result, many organizations lack regular practices for systematically anticipating and identifying ethical risks and monitoring the impact their products have on stakeholders.

How Can We Anticipate and Identify Ethical Risks?

The three reasons why ethical risks easily get missed suggest three corresponding strategies, and guiding questions, for becoming more sensitive to ethical risks:

1. Consider affected stakeholders' rights, interests, and values.
   - Who are all the stakeholders who will be affected by our product?
   - Are their rights, interests, and values adequately protected?
   - How do we know what their interests and values are—*have we asked*?
2. Give everyone responsibility for identifying ethical risks.
   - How do we treat team members who raise potential ethical risks? Is this something our culture encourages?
   - Do we have ways for rewarding and celebrating team members that bring ethical risks to our attention?
   - Does everyone on the team feel in charge of looking for cross-cutting ethical risks?
3. Build and conduct regular exercises for anticipating and identifying ethical risks.
   - Do we have regularly scheduled ethical risk-sweeping exercises?
   - Do we collect and analyze data to identify potentially harmful impacts on stakeholders?
   - Are there channels and forums for people on the team to raise ethical risks, and do people feel empowered to use them?

Meet RudiBrace

For the remainder of this lesson and throughout the course, you'll follow one product team as it grapples with ethical risks and ethical decision-making. The team is part of Rudison Technologies, a firm that provides business-to-business IT services.

The team's mission is to help organizations become more productive by fostering team interaction and communication among colleagues. The problem it wants to address is a lack of social interactions and a sense of community among co-workers because working schedules are too individualized. If workers would know when their co-workers are available for social interactions; for instance, a little chat at the coffee table, social interactions would increase.

To achieve its mission, the team is developing a product called RudiBrace, a digital bracelet that offers a new way to track employee activity in pursuit of the team's mission. The bracelet is supposed to connect to an organization's internal social network, such as Microsoft® Teams® or

Slack. The main use case is organizations with large office spaces, where it is difficult for employees to bump into each other. The tool displays status messages when colleagues arrive at the office or take coffee breaks.

The RudiBrace Team

The RudiBrace team consists of five people: a project manager, a hardware engineer, a data scientist specialized in AI, a user-experience (UX) designer, and a software engineer. The team members had been working at Rudison Technologies previously and came together because they believed in the RudiBrace project and mission.

From the start of the project, the team set out to engage with ethics throughout the product development lifecycle. It is aware of the promises of wearable technologies, such as promoting more seamless interactions, gaining valuable insights about our activities, and so forth. However, it is also aware that there may be unintended ethical risks with the different decisions made in product development.

Once the product passes the ideation stage, the team plans to build a box that contains all the sensors that will eventually be in the bracelet. Employees can carry the box around with them, gathering the data that the bracelet would. A mock-up management dashboard shows statistics based on data from the box. The team also builds a mock-up of the Slack plug-in, displaying when users arrive at the office or take breaks. During these user tests, valuable insights about ethical risks can be gathered.

Once the prototype is validated, the product team also has a unique chance to test the box in the office of a big software company, with 10,000 employees scattered over a large office complex in Silicon Valley. Based on this test, management will decide either to commit to a full build-out of the project and take it to market, or to discard the project. The tests will reveal ethical risks among a large and varied range of stakeholders and will guide the team and the company in deploying an ethical product.

Ethical Risks of RudiBrace

To take ethics seriously, the RudiBrace team builds in a variety of reflection moments, workshops, stakeholder consultations, and ethics reviews throughout the product development lifecycle. In the remainder of this course, we will encounter many of these exercises. By means of these exercises, the team will anticipate ethical risks and be able to mitigate them early on.

There are many types of ethical risks that the RudiBrace team will have to consider. The RudiBrace will gather large quantities of sensitive user data, which needs to be securely stored and transferred. This data is produced by workers, whose privacy will have to be respected. Moreover, the algorithms developed by the team might affect the daily routine of workers, which should not lead to unfair practices. Finally, employees have an interest to understand what data the device collects for which purposes, and to have ways to hold their employer accountable for how it uses the data.

ACTIVITY 2-3

Identifying Potential Risks the RudiBrace Team Might Face

Scenario

As part of project planning, the RudiBrace product team wants to consider possible ethical risks they could encounter.

1. Are there any stakeholders whose interests, rights, or values the RudiBrace team might fail to consider? If so, whom?

2. Is there a risk that ethical issues might fall through the cracks due to the way the team is structured? If so, what might you suggest to reduce that risk?

TOPIC D

Tools for Identifying Ethical Risks

Ethical risks often become painfully apparent once it is already too late, when some harm has been caused or data has been compromised. For example, think of the Cambridge Analytica Scandal, which exposed how user data gathered from Facebook was illegitimately used to manipulate voters through political campaigns. This became a prominent topic of discussion only after a significant amount of data had already been compromised. But *anticipating* ethical risks *before* your work generates negative impacts can be challenging. In this part of the lesson, you will explore practical tools and practices that can help you anticipate and identify ethical risks.

Commonly Used Tools

Three tools that help teams identify ethical risks are:

- **Consequence scanning** to prompt key ethical considerations. This tool is especially helpful towards the prototyping stage of a product, when potential uses can be foreseen. Its major benefit is that it helps identify a broad range of ethical risks; its major drawback is its lack of depth.
- **Scenario analysis** to improve foresight. This tool is most helpful close to the ideation stage, when there is still uncertainty about possible uses of a product. Its major benefit is its capacity to anticipate ethical risks before they materialize; its major drawback is the difficulty to assess the probability of ethical risks to occur.
- **Stakeholder prompts** to structure stakeholder engagement. This tool is helpful in the testing phase of a product, when different uses for different stakeholders are known. Its major benefit is that it helps identify ethical risks from a broad range of perspectives; its major drawback is its subjective nature.

We will discuss these tools in detail in the following sections.

Consequence Scanning

Key ethical considerations are concepts that tend to capture the most important areas of ethical concern in emerging technology. Product teams can use key ethical considerations to help identify common ethical risks by means of ***consequence scanning***. This method makes use of key ethical considerations to trigger people's intuitions about a wide range of possible impacts of a product.

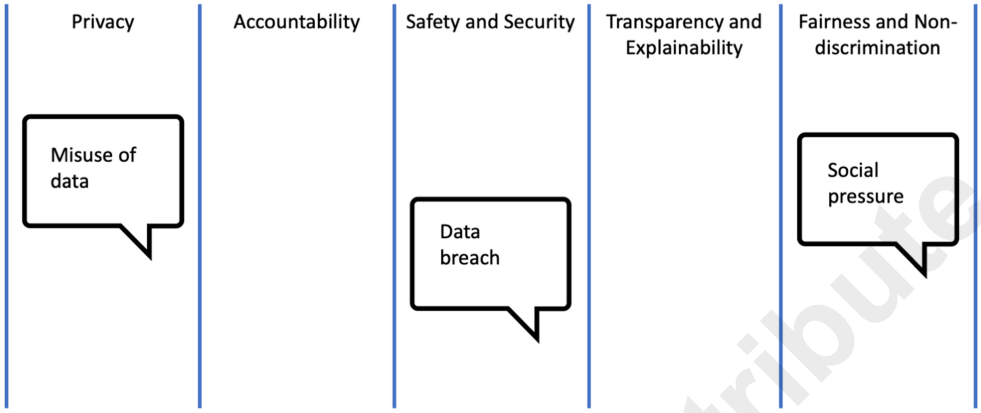


Figure 2-3: Consequence scanning.

The RudiBrace team conducted a consequence scanning workshop in the following way:

- The project manager put the key ethical considerations on a large piece of paper and invited the rest of the team to gather around.
- She handed out sticky notes and asked the team members to brainstorm about ethical risks belonging to one particular ethical consideration for 10 minutes.
- The team members wrote risks down on sticky notes and put them on the wall. The team found 20 ethical risks in total.
- One important ethical risk belonged to the privacy consideration: Employers may use the productivity data of employees for decisions about promotion or firing. This would create negative impacts on employees who are classified as unproductive by the algorithm. Unless the algorithm is 100% accurate, some employees may be incorrectly classified.

Additional Reading

Doteveryone has developed a useful toolkit to conduct a consequence scanning session. For more information, visit **https://www.doteveryone.org.uk/project/consequence-scanning**.

Scenario Analysis

A good way to structure a foresight exercise is to brainstorm and discuss possible future scenarios: little storylines that explore the "what if" of a product's use and misuse.

Scenario analysis can be structured in accordance with well-known ethical risks posed by emerging technologies. For instance, product teams can ask themselves:

- What if our product contributed to the emergence of a surveillance state?
- How could our product contribute to this?
- What other ways might our product be used to violate privacy?

According to the Ethical OS Toolkit, there are eight primary risk zones, as shown in this image from **ethicalos.org**:

Figure 2-4: Scenario analysis.

The RudiBrace team conducted an ethical risk analysis in the following way:

- The project manager put the Ethical OS risks analysis tool on a large interactive screen and pre-selected the most relevant risk zones, which are zones 3 up to 7.
- She divided the team into two groups and asked each group to create a problematic scenario for a particular risk zone, assigning two rounds of 15 minutes.
- The scenarios were drafted on pieces of paper and consist of words, figures, and diagrams.
- After the drafting period, the project manager invited the two groups to present the most problematic scenario.
- A particularly important ethical issue came from a scenario of a future of work, in which tech firms are all monopolists and worker rights have eroded. In such a scenario, the team imagined that the RudiBrace might worsen economic equalities by setting up workers against each other in having better performance metrics, which might generate a race to the bottom.

Additional Reading

Visit the following websites for useful tools for conducting scenario exercises: **https://ethicalos.org/** and **https://ethicalexplorer.org/**.

Stakeholder Prompts

Stakeholder prompts are a useful tool for identifying ethical risks through stakeholder engagement. One way to organize stakeholder prompts is with stakeholder cards. A stakeholder card represents the particular perspective of a person who relates to a technology product. It lists relevant information about this person and gives an impression of the person's character; for instance by means of a quote.

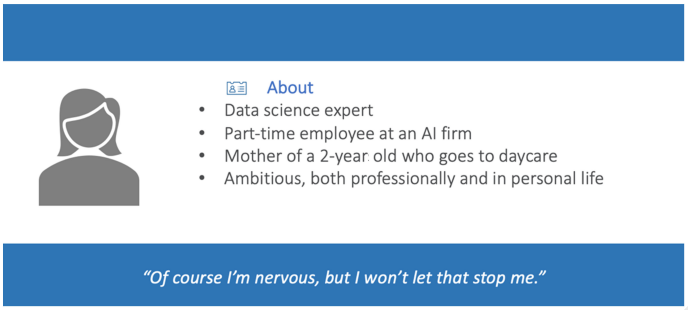


Figure 2-5: A stakeholder card.

The RudiBrace team conducted a stakeholder card exercise in the following way:

- The project manager put together stakeholder cards that she customized based on a generic set that is managed by the board of Rudison Technologies, representing some typical potential users of the RudiBrace.
- The project manager prompted three types of stakeholders in particular: a mother working part-time, a male employee with a high career ambition, and a trainee who is struggling to get by.
- She gave one card to each team member and asked them to step into the shoes of their stakeholder and identify ethical risks from that perspective.
- When considering the product from the perspective of the first stakeholder, Sarah, the team found the ethical risk that the RudiBrace might induce pressure for people to be in the office at particular times, which might cause difficulties for a young mother with caregiving commitments. From her perspective, the RudiBrace might pose a tradeoff between a healthy family life and success at work—with potentially discriminatory implications for women in particular.

Additional Reading

Examples of stakeholder engagement tools include:

- **Judgment Call**, a tool for cultivating stakeholder empathy (**https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/judgmentcall**).
- **Community jury**, a tool to engage directly with stakeholders (**https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/**).

ACTIVITY 2-4
Using Consequence Scanning

Scenario
The RudiBrace team has identified an ethical risk within the privacy theme, in that there is the chance the data could be misused. You will assist them in finding additional potential risks. You can use this image to help complete the activity.

Privacy	Accountability	Safety/Security	Transparency/ Explainability	Fairness/Non-Discrimination
• Misuse of data	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•

Figure 2-6: RudiBrace consequence scanning.

1. Identify additional privacy risks that might arise from the RudiBrace product.
2. Identify accountability risks that might arise from the RudiBrace product.
3. Identify safety and security risks that might arise from the RudiBrace product.
4. Identify transparency and explainability risks that might arise from the RudiBrace product.
5. Identify fairness and non-discrimination risks that might arise from the RudiBrace product.

	Note: To view a sample completed table, open **C:\095029Data\Identifying Ethical Risks\RudiBrace Consequence Scanning.rtf**.

.

TOPIC E

Use Regulations, Standards, and Human Rights to Identify Ethical Risks

Since many ethical risks created by emerging technologies are new or ill-understood, it can seem daunting to ensure to identify and address them all in practice. It is one thing to understand that your product poses a privacy risk, and another to pin down exactly what the risk consists of and what can be reasonably required to mitigate it. Applicable regulations, standards, and human rights documents can all help in concretizing ethical risks. Compliance with regulations and standards is different from ethics in that they capture only what you are legally required or expected to do, while ethics is about what you ought to do. Nonetheless, regulations, standards, and human rights are all critical tools for identifying ethical risks. This topic introduces how regulations, standards and human rights can assist you in identifying ethical risks.

The Relationship Between Standards, Regulation, and Ethics

To understand what regulation and standards have to do with ethics, consider the example of medical ethics. Much reflection in medical ethics starts from four guiding principles:

- **Beneficence**: Help others to further their legitimate interests.
- **Autonomy**: Enable others to make informed and voluntary decisions.
- **Non-Malevolence**: Do not intentionally harm others.
- **Justice**: Treat others equally.

These principles are helpful starting points to organize our thoughts. But they are too abstract to make all relevant ethical risks in medical scenarios salient. That is why there is detailed regulation, guidelines, and standards to guide medical professionals in day-to-day decision-making. The latter have been developed with the four guiding principles in mind and can be evaluated and revised on the basis of these principles. But guidelines and regulation concretize these principles, by for instance prescribing what is practically required to obtain informed consent from a patient in a critical condition.

Regulations

A ***regulation*** is a set of rules made by a sovereign legislative body, often in consultation with subject matter experts. In contrast to ethical frameworks, regulations have legal standing and are enforceable. Regulation has several purposes and often results from compromises struck by policymakers with various different motives. However, one common function of regulation is to concretize ethical minimum standards. Regulation is often a useful way of screening out options for action that fail to meet ethical requirements.

Because not all regulatory authorities are equally legitimate, however, it is important to view regulations critically. Regulations promulgated by a regime with a poor human-rights record may be less legitimate than regulations enacted by a government with a strong commitment to human rights.

GDPR

One of the regulations you might have already heard about is the ***General Data Protection Regulation (GDPR)*** adopted by the European Union (EU) in 2016. This regulation is intended to protect individual privacy by holding data collection and data processing entities accountable for the information of EU citizens. The regulation applies to all entities that collect or process the personal data of EU citizens, even if the entity is not based in the EU, and includes provisions concerning the export of personal data outside the EU.

The GDPR ultimately provides the following:

- It upholds the privacy rights of individuals (e.g., the right to correct inaccurate personal data).
- It enforces restrictions and security obligations for organizations (e.g., report data breaches within 72 hours).
- It issues penalties for noncompliance (e.g., fines up to €20 million or 4 percent of global turnover).

Data Privacy Regulations and Ethical Risks

During consequence scanning, the RudiBrace team has identified privacy as an important area for potential ethical risks. For instance, managers might use the data to make promotion and firing decisions. This has negative impacts on employees classified as unproductive by the algorithm.

Consulting even the broad principles underlying privacy regulation can help identify specific ethical risks. The RudiBrace team gets advice on the constraints that data privacy regulation imposes on them. By considering some of the principles underlying GDPR, currently one of the most stringent frameworks for privacy regulation, the team can better understand potential solutions to the privacy concerns that their product raises.

The following table describes some selected GDPR principles and implications for the RudiBrace team.

Theme	*Principle*	*Implication for RudiBrace*
Fairness and transparency	Data collection must be fair and transparent.	• Need to explain to users which data is collected and what happens with it. • Need to consider whether fairness requires an opt-out option for employees
Purpose limitation	Organizations should collect personal data only for a specific purpose, clearly state what that purpose is, and collect data only for as long as necessary to complete that purpose.	The data gathered with the tool cannot be used beyond its stated purpose.
Accuracy	Every reasonable step must be taken to erase or rectify data that is inaccurate or incomplete.	Need to ensure that the metrics computed really capture interaction and communication and are appropriately linked to productivity.
Integrity and confidentiality	Data processing needs to ensure appropriate security of the personal data.	Given the sensitivity of the data collected by RudiBrace, the tool needs to meet exceptionally high security standards.

Standards

Standards are rules or guidelines generally created by industry and civil society organizations, such as the International Standards Organization (ISO), the Institute of Electrical and Electronics Engineers (IEEE), and the Forest Stewardship Council (FSC). They serve to establish common norms for interoperability, product quality, and professional conduct, within the boundaries established by regulation. Standards are also an important mechanism for establishing and interpreting ethical requirements for emerging technology.

Standards set by these organizations are not legally binding and are therefore not legally enforceable. However, standard-setting bodies can penalize compliance failures by refusing to certify products, revoking membership in professional associations, and public shaming.

Typically, standards are not developed democratically but represent the opinions of experts. Like legal regulations, therefore, their legitimacy is not guaranteed. Standards may not always take due account of different interests and viewpoints of the people affected by them. But they can be a useful starting point for determining ethical requirements.

Standards and Ethical Risks

Standards often focus on implementing legal regulations and ethical best practices. They are well-suited to provide concrete technical guidance and policy details that can help make ethical risks concrete. Standards aim to be in sync with the views of experts on how best to operationalize an ethical requirement.

For instance, RudiBrace could get advice on principles and best practices contained in the ***ISO Standard 27701***, which relates to the way an organization collects personal data and prevents unauthorized use or disclosure. The standard contains detailed guidance that the RudiBrace team should take into account when building the product, including guidance on incident management, access to systems and services that process personal identifiable information, cryptographic protection, and information transfer policies.

Human Rights

Human rights are rights we have as human beings. They create a protective zone around persons and allow them the opportunity to further their valued personal projects without interference from others. Examples of human rights include:

- Security of the person
- Due process and a fair trial
- A right to own property
- Freedom of movement
- Freedom of speech and political participation
- Freedom from discrimination
- Freedom to marry
- The right to work
- Religious freedom

Human rights complement standards and regulation when filtering permissible options. Regulation and standards are detailed and often apply to specific emerging technologies. Since the development of regulation and standards takes time, they tend to lag behind recent advances in emerging technology. By contrast, human rights lack the specificity of law and regulation. Yet human rights apply generally to all organizational activity and are widely accepted as ethical minimum standards. Organizations can use human rights to explore whether a given product may cause unacceptable negative impacts, even in areas where regulation and standards do not provide sufficient guidance.

Important Human Rights Documents

An authoritative list of the core internationally recognized human rights is contained in the **International Bill of Human Rights**, consisting of the **Universal Declaration of Human Rights** and the main instruments through which it has been codified: the **International Covenant on Civil and Political Rights** and the **International Covenant on Economic, Social and Cultural Rights**. Coupled with the principles concerning fundamental rights in the eight ILO core conventions as set out in the **Declaration on Fundamental Principles and Rights at Work**, these documents provide a benchmark against which social actors assess the human rights impacts of business enterprises.

Document Links

This table lists the web links for the documents discussed in this section:

Document	*Link*
International Bill of Human Rights	**https://www.ohchr.org/documents/publications/factsheet2rev.1en.pdf**
Universal Declaration of Human Rights	**https://www.un.org/en/universal-declaration-human-rights/**
International Covenant on Civil and Political Rights	**https://www.ohchr.org/EN/ProfessionalInterest/Pages/CCPR.aspx**
International Covenant on Economic, Social and Cultural Rights	**https://www.ohchr.org/EN/ProfessionalInterest/Pages/CESCR.aspx**
Declaration on Fundamental Principles and Rights at Work	**https://www.ilo.org/declaration/**

The Role of Organizations with Respect to Human Rights

The UN Guiding Principles on Business and Human Rights assign to organizations the role with respect to human rights. This means organizations should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved.

In many jurisdictions, human rights have been incorporated into national law. However, the responsibility of business enterprises to respect human rights is distinct from risks of legal liability and enforcement. Organizations should respect human rights as an ethical red line, regardless of whether human rights are legally binding in the jurisdiction they operate in or are likely to be enforced.

Additional Reading

For more information:

- UN Guiding Principles on Business and Human Rights: **https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf**
- "On Artificial Intelligence: A European approach to excellence and trust": **https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en**

Human Rights Due Diligence

In order to prevent and address adverse human rights impacts, organizations should conduct ***human rights due diligence***. The process should include assessing actual and potential human rights impacts, integrating and acting upon the findings, tracking responses, and communicating how impacts are addressed.

Additional Reading

For more information:

- The UN Global Compact provides guidance on how to conduct human rights impact assessments in the report "5 Steps towards Managing the Human Rights Impacts of your Business": **https://www.unglobalcompact.org/library/4921**.
- The Danish Institute for Human Rights has developed a toolbox and assessment guidance for human rights impact assessments: **https://www.humanrights.dk/business/tools/human-rights-impact-assessment-guidance-toolbox**.

Human Rights and Ethical Risks

Human rights can be useful in making ethical risks concrete, particularly in areas where regulation and standards have not yet been developed or are lagging behind recent advances in emerging technology. The process for taking human rights into consideration includes:

1. Identify human rights relevant to your product.
2. Ask: Are there risks that our product will fail to respect human rights?

For instance, the RudiBrace team can consult the ***Toronto Declaration***, a statement on human rights standards for machine learning by Human Rights Watch and a coalition of rights and technology groups. The declaration aims to protect the right to equality and non-discrimination in machine learning systems. It sets out the elements required in a human rights impact assessment focused on discrimination:

- Identify potential discriminatory outcomes.
- Take effective action to prevent and mitigate discrimination and track responses.
- Be transparent about efforts to identify, prevent, and mitigate against discrimination in machine learning systems.

The RudiBrace team can learn that it needs to pay attention to how their product may have discriminatory impacts on employees.

Additional Reading

For more information, visit **https://www.torontodeclaration.org/declaration-text/english/#due_diligence**.

ACTIVITY 2-5

Using Regulations, Standards, and Human Rights to Identify Ethical Risks

Scenario

As a member of the RudiBrace team, you meet to review some regulations, standards, and other material designed to help you identify ethical risks that might be present in the RudiBrace product. You have received a suggestion to review the EU assessment list for trustworthy AI, which is developed by the High-Level Expert Group on AI, appointed by the European Commission (the executive branch of the European Union)

1. In a new browser window or tab, navigate to the report titled **Ethics guidelines for trustworthy AI** stored in the European Commission's document library.
   a) If necessary, open a new browser window or tab.
   b) Navigate to **https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai**.
2. Read the summary information on this page.

   Do you think the assessment list mentioned in the final paragraph is a regulation or a standard?

   ○ Regulation
   ○ Standard
3. **Why did you answer as you did?**
4. Read the first two or three bullets that describe the "7 key requirements that AI systems should meet." in the Ethics guidelines for trustworthy AI.

   What are some practical learnings the RudiBrace team can take from this?

Summary

In this lesson, you identified ethical risks. You investigated ethical reasoning and some of the events and mindsets that can hamper ethical reasoning. You also identified and used tools, regulatory documentation, standards, and guidelines to assist in finding ethical risks inherent in product development. The ability to identify and address ethical issues before they have a negative impact on society is critical to your success as an ethical technologist.

At your workplace, what types of situations have you experienced that required you to use ethical reasoning?

Of the tools and practices discussed in this lesson, which do you think will be the most effective at your workplace?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

3 Ethical Reasoning in Practice

Lesson Time: 2 hours

Lesson Introduction

In this lesson, you will apply a method for ethical decision-making used by some of the world's largest tech firms, governments, and non-governmental organizations (NGOs). For a decision to be ethical, it should not only be in line with applicable regulation, standards, and human rights. Rather, it should also appropriately balance the interests and values of everyone affected.

This lesson starts by discussing important ethical theories and how to apply them in ethical decision-making. Next, it covers ethical decision-making frameworks and demonstrates how to put them to practical use. You will also identify strategies for addressing ethical issues at different stages of the product development cycle. To finish, you will discuss how to avoid common pitfalls in ethical decision-making, including biases and ethics washing.

Lesson Objectives

In this lesson, you will:

- Describe ethical theories and how to apply them in practice.
- Use ethical frameworks to guide decisions concerning emerging technologies.
- Select options for action to address ethical risks.
- Avoid common pitfalls in ethical decision-making.

TOPIC A

Ethical Theories

One way to implement ethical reasoning is to consider ethical issues from the perspective of moral theories. Ethical theories capture various approaches to determining what makes actions right or wrong.

What Is Ethical Decision-Making?

Ethical decision-making is the process of identifying ethical risks, developing options for action, and selecting an option that is supported by the best available reasons. The ethical theories discussed in this topic can go a long way toward helping you with the process of ethical decision-making.

Overview of Ethical Theories

Ethical theories are systematic efforts to understand moral concepts and justify moral principles. Ethical theories fall into three main categories: the first focuses on a person's moral character, the second on fulfilling one's moral duties, and the third on bringing about good consequences. These approaches are commonly referred to as virtue ethics, deontology, and consequentialism, each of which we will discuss in this topic.

Figure 3-1: Ethical theories.

Virtue Ethics

Virtue-based approaches assume that ethics is not primarily about identifying right actions—rather, it is about leading a virtuous life. ***Virtue ethics*** goes back to Confucius and Aristotle, and also has influential modern proponents. Its main assumption is that to lead a virtuous life we need to cultivate certain virtues—excellences of character that promote human flourishing. These include qualities such as wisdom, justice, beneficence, temperance, courage, and honesty.

Virtue ethics can help us identify the right thing to do by focusing on what a virtuous person would do in a given situation. But, generally speaking, the focus of virtue ethics is much broader: namely, on developing the capabilities we need to live well in harmony with others.

When applying insights from virtue ethics, we might consider the precedent that our actions might set for others. We might consider what our behavior says about our own character, and how our behavior might affect that of others. Virtue ethics often identifies virtuous conduct as finding a balance between extreme opposites. The virtue of courage, for instance, requires cultivating habits for navigating between recklessness and timidity. In designing technology, this might mean seeking a middle path between the impulse to disrupt the status quo and the fear of disastrous consequences.

Additional Reading

For more information about ethics and virtue, visit: **https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/ethics-and-virtue/**.

Deontology

Deontology prioritizes acting in accordance with moral duties. According to deontology, whether an action is right or wrong depends on whether one complies with one's moral duties, not on the outcome of one's actions. One key deontological principle is that we must never treat other people as means, or as tools to achieve some goal, but must always treat others as what the philosopher Immanuel Kant described as "ends in themselves." This means that there are certain things we simply must never do to others, because this would violate their status as beings with intrinsic moral worth. For example, killing a person to distribute their organs to several other patients who would otherwise die would treat that one person as a means—as a mere tool for the benefit of others. A deontologist would find this troubling because, for deontologists, the rightness or wrongness of our actions is determined primarily by the intrinsic nature of our actions, not by whether they bring about good consequences.

Applying insights from deontology in practice often means focusing primarily on our duties to respect the rights of others. For example, we might think carefully about how our decision could affect the dignity and autonomy of others. Applying insights from deontology in technology fields might involve acting transparently, securing the consent of affected stakeholders, and establishing guardrails for permissible and impermissible conduct—for the chief reason of conforming with moral duties.

Additional reading

For more information on duty-based or deontological ethics, visit: **http://www.bbc.co.uk**, search the site for ***duty-based ethics***, and select the link **BBC - Ethics - Introduction to ethics: Duty-based ethics** dated 9 September 2009.

Consequentialism

Consequentialist approaches assume that the rightness or wrongness of actions depends exclusively on the goodness or badness of consequences they bring about. According to ***consequentialism***, an action is right insofar as it promotes good consequences and wrong insofar as it brings about bad consequences.

The most famous consequentialist theory is **classical utilitarianism**, typically associated with English philosophers Jeremy Bentham and John Stuart Mill. The best-known utilitarian slogan is "the greatest happiness for the greatest number"—roughly the idea that our actions should aim to maximize happiness for as many people as possible. There are various versions of utilitarianism. They differ, for example, over what the ultimate value is that should be promoted (for example, happiness, well-being, utility, or certain human capabilities), how it should be promoted, and whether our focus should be on evaluating the consequences of our individual acts or the rules we follow.

The appeal of consequentialist theories like utilitarianism is easy to see: often, they provide a straightforward way of determining what we should do. For example, when faced with the choice of killing one person to save five others, a utilitarian would see a very strong reason for killing the one person to save the five. Since consequences are paramount, consequentialists are generally more willing to sacrifice other considerations, like rights, to bring about the best outcomes.

To apply consequentialism in practice, we might first consider the potential benefits and burdens of different options, and look for the options with the greatest net benefits. Consequentialist approaches that seek to maximize welfare, for example, might direct us not to neglect the interests of distant others, such as animals and future generations, which also stand to benefit or suffer from our actions. In the context of technological development, consequentialism can be especially helpful

for directing our attention to missed opportunities to do good in the world, rather than focusing only on avoiding harm.

Additional Reading

For more information on the utilitarian approach, visit: **https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/calculating-consequences-the-utilitarian-approach/**.

Triangulation Between Ethical Theories

Each theory has strengths and weaknesses. In some cases, different courses of action are supported by different reasons, and different moral theories may seem to pull us in different directions. In such cases of moral conflicts, determining the course of action for which we have the strongest reasons can be especially difficult. For example, if your self-driving car must either crash into a wall, which would kill you, or run over a child who has run onto the road to catch a ball, the algorithm in the car's crash-avoidance software has both a reason to make the car crash into the wall (to protect the child) *and* a reason to keep going (to save you, the passenger).

In cases like this, and in many other, less dramatic ones, what we think is the right thing to do will come down to how we weigh different reasons. There is no simple formula for resolving these cases.

Guidelines for Using Ethical Theories in Decision-Making

Note: All Guidelines for this lesson are available as checklists from the **Checklist** tile on the CHOICE Course screen.

Follow these guidelines when you are using ethical theories in your decision-making.

Using Ethical Theories in Decision-Making

The previous note of caution notwithstanding, considering guiding questions based on the three approaches outlined earlier can help in determining what types of reasons might play a role in our decision-making. These include:

- Virtue ethics:
  - Would we want future generations of technologists to use our practice as the example to follow?
  - What will this product say about us as people in the eyes of those who receive it?
  - What habits of character will this project foster in users and other stakeholders?
  - Do our choices strike the appropriate mean between deficit and excess, and do we appropriately balance all the virtues?
- Deontology:
  - What rights of and duties to others must we respect?
  - Can our choice be universalized?
  - Are we following a rule that could be generalized?
  - How might stakeholders' dignity and autonomy be affected?
  - Does our product treat people in ways that are transparent?
- Consequentialism:
  - What are the benefits and burdens created by this project, and how are they distributed among various stakeholders?
  - Will the effects likely create more overall good than harm?
  - How might future generations be affected by this product?
  - Are there missed opportunities to generate even greater benefits?

ACTIVITY 3-1
Applying Ethical Theories in Decision-Making

Scenario
You are continuing your work with the RudiBrace team. To analyze potential ethical risks, you would like to draw some guidance from the three main ethical theories discussed in this topic.

1. What is one ethical concern that consequentialist reasoning might raise about RudiBrace?

2. What is one ethical concern that deontological reasoning might raise about RudiBrace?

3. What is one ethical concern that virtue ethics might raise about RudiBrace?

TOPIC B

Use Ethical Decision-Making Frameworks

For ethical reasoning to be most effective, a systematic approach is recommended. One way to ensure such an approach is to base your decision-making on established ethical frameworks.

Phases of Ethical Decision-Making

You can break down ethical decision-making into three phases:

1. Product teams *identify* ethical risks. To do this in a structured way, teams can employ ethical foresight methods and engage with stakeholders.
2. Product teams *understand* ethical requirements. This requires identifying which options are permissible in the light of regulation, standards, and human rights.
3. Product teams *select* options for action. This involves generating options for addressing the ethical risk. Among permissible options, organizations need to weigh competing considerations to arrive at a choice that is supported by good reasons.

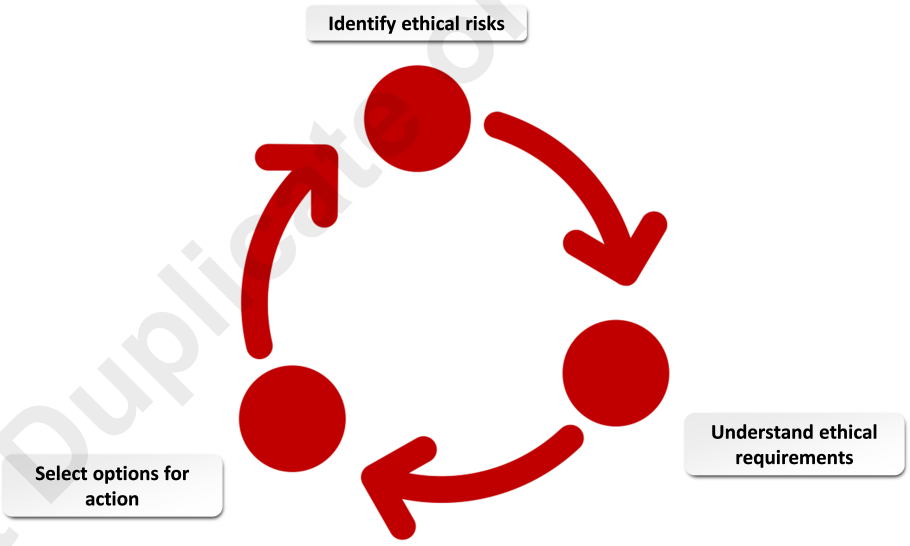


Figure 3-2: Phases of ethical decision-making.

Ethical Decision-Making Frameworks

Over the past few years, some important global actors, including governments, large tech firms, and professional organizations, have introduced frameworks for ethical decision-making. These frameworks incorporate different aspects of the ethical decision-making process, all of which will be covered in this lesson.

Aspect	*Description and Examples*
Common standards for defining and responding to ethical reasons	Frameworks cover the most salient areas of ethical concern in emerging technology, such as privacy, fairness, and accountability. Typical examples of frameworks are: • **Montreal Declaration** • **Toronto Declaration** • **Universal Guidelines for AI** • **UNI Global Union Top 10 Principles of AI**
Ethical challenges	Some frameworks cover ethical challenges in emerging technology in great technical detail. These include: • **Asilomar AI Principles** • **IEEE Ethically Aligned Design** framework
Methods for ethical decision-making	Frameworks often present methods that can be followed to successfully engage in ethical decision-making. These methods may spell out how ethical reasons can be addressed in product design or how to respond to ethical failure. Typical examples are: • **EU Ethics Guidelines for Trustworthy AI** • **IEEE Ethically Aligned Design** framework
Context of ethical decision-making	Frameworks can also discuss particular aspects of the context of ethical decision-making. A typical example is the **AI Government Readiness index**, and the related **Responsible AI Map**, which map the state of AI technology in different countries and the extent to which regulations are in place to guarantee ethical data-driven technologies.
Policy for supporting ethical decision-making	Frameworks may present guidelines for policy making to support ethical decision-making. Typical examples are: • **Beijing AI Principles** • **OECD Principles on AI** • **G20 AI principles**

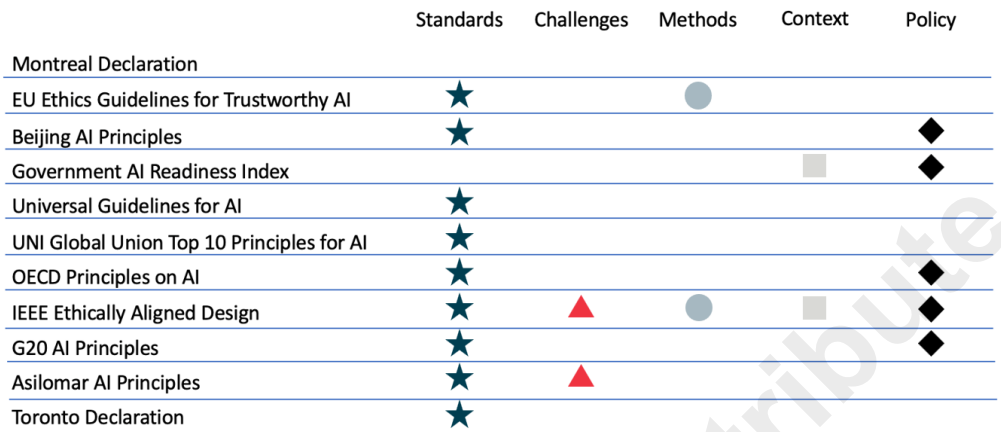


Figure 3-3: Aspects of ethical decision-making frameworks.

Additional Reading

For more information, visit:

- Montreal Declaration: **https://www.canasean.com/the-montreal-declaration-for-the-responsible-development-of-artificial-intelligence-launched/**
- Toronto Declaration: **https://www.torontodeclaration.org/**
- Universal Guidelines for AI: **https://thepublicvoice.org/ai-universal-guidelines/**
- UNI Global Union Top 10 Principles of AI: **http://www.thefutureworldofwork.org/opinions/10-principles-for-ethical-ai/**
- Asilomar AI Principles: **https://futureoflife.org/ai-principles/**
- IEEE Ethically Aligned Design: **https://ethicsinaction.ieee.org/**
- AI Government Readiness index: **https://www.oxfordinsights.com/government-ai-readiness-index-2020**
- Beijing AI Principles: **https://www.baai.ac.cn/news/beijing-ai-principles-en.html**
- OECD Principles on AI: **https://www.oecd.org/going-digital/ai/principles/**
- G20 AI principles: **https://www.g20-insights.org/related_literature/g20-japan-ai-principles/**

Key Ethical Considerations

Ethical decision-making is structured around ***key ethical considerations***. These are concepts that tend to capture the most important areas of ethical concern in emerging technology. Product teams can use key ethical considerations as a tool to identify common ethical risks.

The great benefit of key ethical considerations is that they are general and can be used throughout the product development lifecycle. The downside of key ethical considerations is that they are less sensitive to particular contexts, because they do not help with thinking about novel problems or different stakeholder perspectives.

Many key ethical considerations are common in global frameworks for ethical decision-making. The Berkman Klein Centre has surveyed key ethical considerations across the global frameworks in their paper entitled **Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI**.

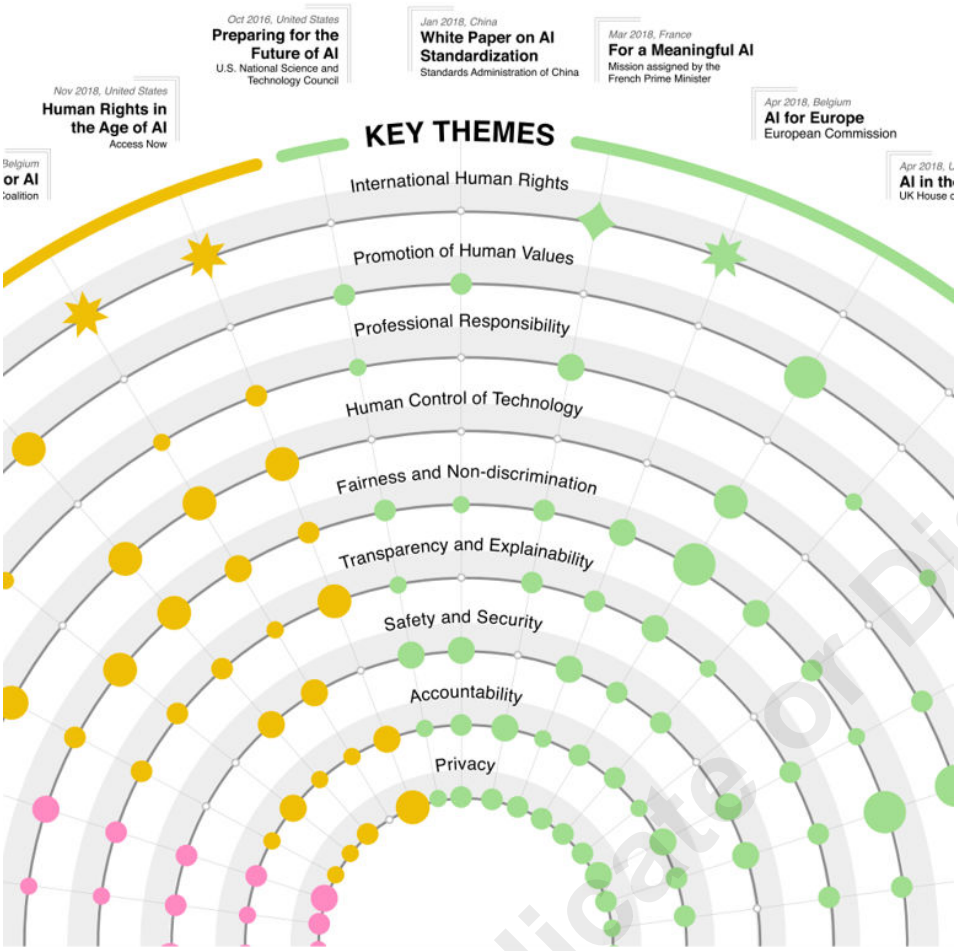


Figure 3-4: Key ethical considerations.

Note: This image shows just a fraction of the documents that were analyzed in the study. For the complete study, including the entire image, visit **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482**.

Additional Reading

You might want to explore these tools for a high-level overview of potential ethical risks with your product:

- Data Ethics Canvas by the Open Data Institute: **https://theodi.org/article/data-ethics-canvas/**
- RAI Design Assistant by the Responsible AI Institute (formerly AI Global): **https://designassistant.responsible.ai//**
- Ethics Self-Assessment Tool by the UK Statistics Authority: **https://uksa.statisticsauthority.gov.uk/about-the-authority/committees/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/**
- The Ethics Canvas by the ADAPT Centre: **https://www.ethicscanvas.org/**

Stakeholder Engagement Strategies

Bias and the limitations of our own perspectives can easily make us overlook important ethical risks. Therefore, the ability to engage stakeholders and experts is a crucial part of ethical decision-making. Product teams can consider three strategies for stakeholder engagement:

- **Perspective taking**: Product teams look for people or groups affected by their product and engage in imaginative perspective-taking.
- **Expert consultation**: Product teams consult representatives of a given stakeholder group or experts on a stakeholder group or type of ethical risk.
- **Focus group**: Product teams conduct focus groups with members of an affected group to hear about their perspective first-hand.

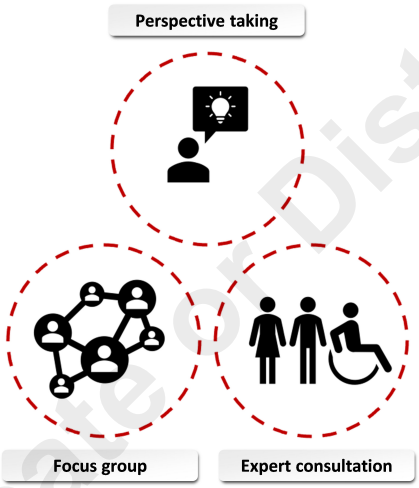


Figure 3-5: Three strategies for stakeholder engagement.

Challenges to Ethical Decision-Making

There are limits to the approach to ethical decision-making described in this lesson. Two factors that have far-reaching effects are complexity and disagreement.

Some ethical risks are too serious or complex for organizations to tackle successfully by themselves. This raises difficult questions about the responsibilities of individual organizations. Minimizing environmental damage is a case in point. For instance, some tech firms have taken advantage of green energy subsidies and shifted all their data centers to run on renewable power. This might seem like an environmentally responsible choice. In some countries, however, the result has been that fewer subsidies are available to households, even though the environmental gains of households switching may be greater. This shows that addressing ethical risks sometimes has surprising adverse effects unless undertaken systematically.

While the amount of disagreement in ethics should not be overstated, there *is* persistent disagreement on how to resolve certain ethical tradeoffs. Take the ethical issue of algorithmic bias. Some maintain that non-discrimination requires that algorithms remain blind to membership of individuals in protected groups. Others maintain that eliminating bias requires empowering marginalized groups by using positive discrimination. There are good arguments on both sides of this debate, and it is unlikely to be resolved soon. This shows that in addressing ethical risks, organizations need to navigate a certain amount of persistent disagreement.

The complexity and disagreement of some ethical risks show that ethical decision-making is not a box-ticking exercise, but a process requiring skill, reflection, and sometimes coordination with other

actors. Once an ethical decision has been made, organizations should measure the impacts of their decision and stay open to revisiting how they made their tradeoffs.

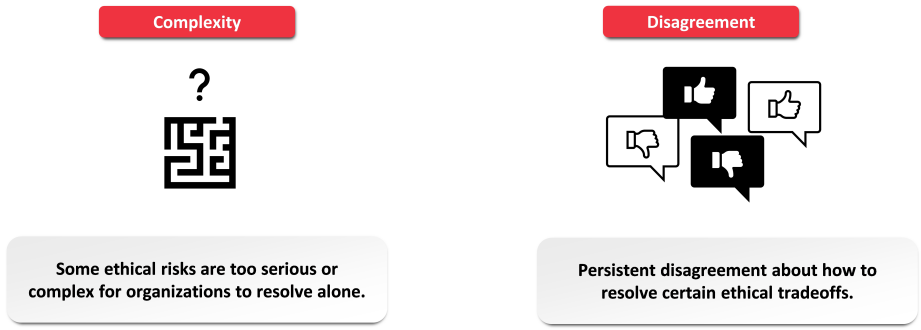


Figure 3-6: Challenges for ethical decision-makers.

ACTIVITY 3-2
Identifying the Best Framework for a Situation

Scenario
Decide which framework for ethical decision-making would be best to consult in the following situations.

1. You want to introduce a company-wide method for ethical decision-making.

   Which framework or frameworks are most suited to this situation?

2. You want to apply local guidelines for ethical decision-making in a subsidiary company in China.

   Which framework or frameworks are most suited to this situation?

3. You want to expand your business to several African countries, but you are unsure about the level of regulation for ethical decision-making in local contexts.

   Which framework or frameworks are most suited to this situation?

4. You want to recommend to your colleagues a document that has comprehensive coverage of information about ethical decision-making.

   Which framework or frameworks are most suited to this situation?

TOPIC C

Select Options for Action

After you identify ethical risks and have a greater understanding of the ethical requirements of a project, you can select options for action to mitigate or remove the risks. In this part of the lesson, you will explore strategies and tools for selecting actions aimed at mitigating ethical risks.

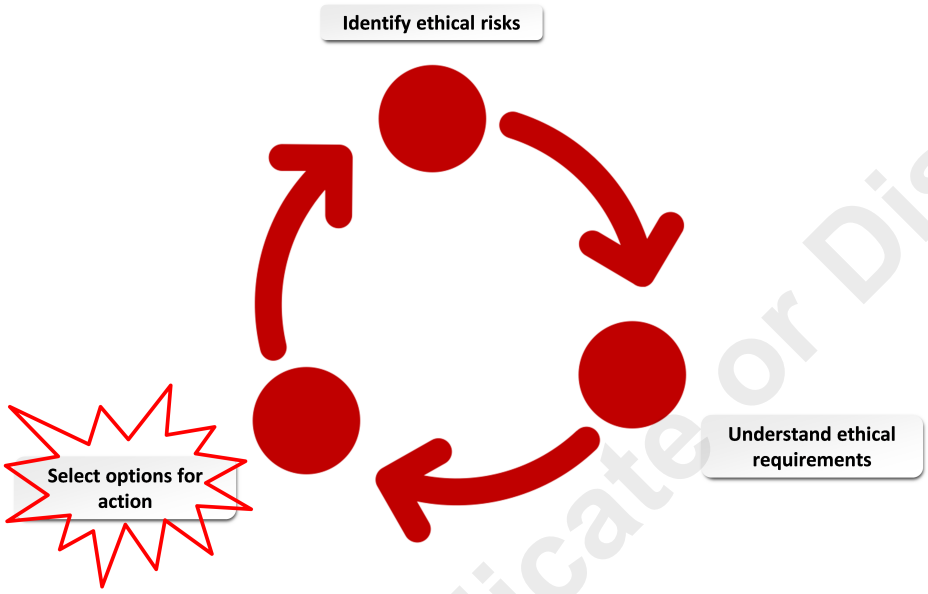


Figure 3-7: Completing the ethical decision-making cycle.

Strategies to Resolve Ethical Risks

Once ethical risks have been identified, teams take time to consider ways forward. What could they do differently? How could they bring about positive change? This topic introduces tools to help product teams with resolving ethical risks:

- Generate options for action
- Select an option supported by the best available reasons

Overall, there are four broad strategies for addressing ethical risks:

- Intervening in product design
- Intervening in the business case
- Proposing internal policies
- Influencing external standards and regulations

Figure 3-8: Strategies for addressing ethical risks.

Addressing Ethical Risks in Product Development

The type of reflection required to address ethical risks depends on the product development lifecycle. The closer a product team finds itself to the ideation phase, the better it will be able to intervene directly in the product design process. The closer it finds itself towards shipping and maintaining the product, the more it will reflect on options like regulating access and usage.

We will consider two types of reflection to address ethical risks:

- Ethics by design, which includes value-sensitive design. Ethics by design is especially helpful when the product is still close to the ideation stage. It enables the promotion of human values in technology design.
- Impact assessment, which includes cost-benefits assessment. Ethical impact assessment is particularly helpful when ethical risks are more tangible, closer towards shipping and maintenance of a product.

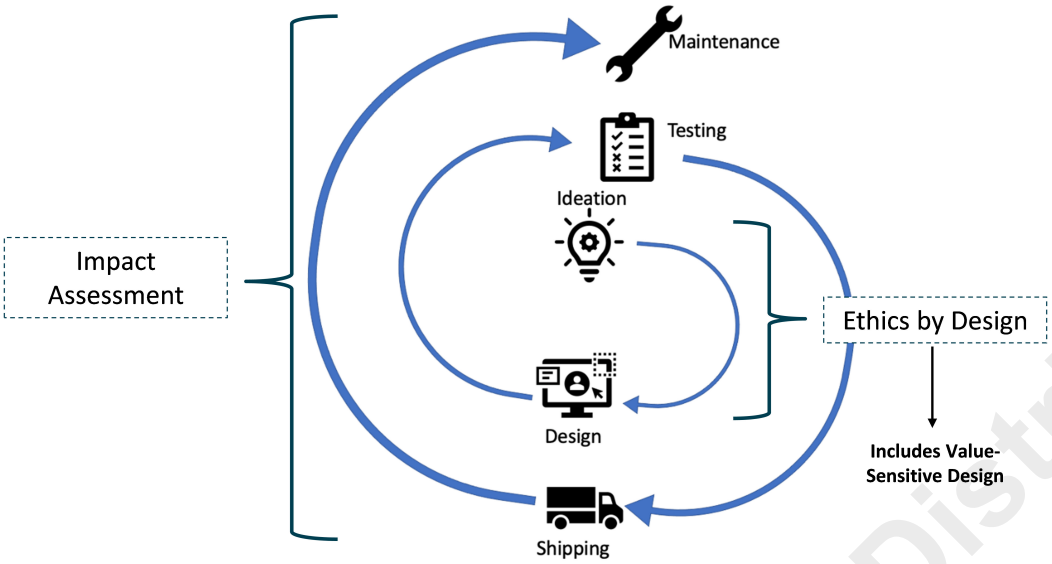


Figure 3-9: Addressing ethical risks in product development.

Ethics by Design

Ethics by design focuses on addressing ethical risks early on, between the ideation and design phases. It is known under different headings, such as value-sensitive design, design for values, or, more specifically, privacy by design. It focuses on ethical risks in general or on a particular ethical theme like privacy or inclusivity.

To engage in ethics by design, product teams take the following steps:

- They establish ethical aims, which can be to design the product to address a social problem, promote a positive human value, or respond to a consumer demand in an ethically sensitive way. They might ask themselves:
  - What problems should we try to solve?
  - What values should we try to promote?
- They consider potential design ideas for realizing their ethical aims. They ask themselves:
  - What viable products could we create to achieve our ethical aims?
  - What are the tradeoffs involved?
- They think of setting design requirements that help them achieve their aims. They might spell out technical requirements for privacy settings of the product.

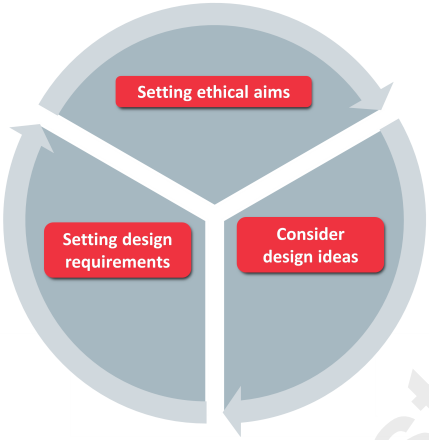


Figure 3-10: Ethics by design.

Value-Sensitive Design

One approach to ethics by design is ***value-sensitive design***. In this approach, product teams first set themselves values that they want the product to promote. Then, for each value, they stipulate norms that make explicit what the product should or should not do. Finally, they come up with design requirements that implement these norms in the product design.

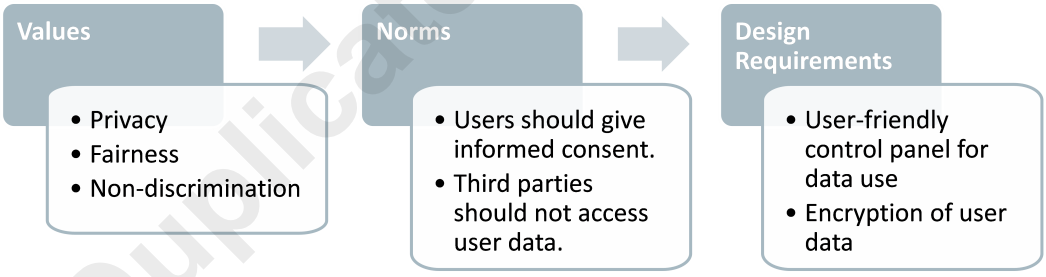


Figure 3-11: Value-sensitive design.

The RudiBrace team conducted a value-sensitive design workshop in the following way:

- The project manager put the most pressing ethical risks of RudiBrace that were identified in the consequence scanning session on the wall in a room. She asked the team members to translate these risks into design values. The team members agreed that RudiBrace should promote the values of privacy and fairness.
- In this session, the team focused on privacy and brainstormed about several important norms that the product should follow. These include that users should provide informed consent about the use of their data and that data should not be accessible to anyone outside the company.
- Based on these norms, the product team developed concrete design requirements that will help achieve the privacy value. An important one was to design a user-friendly control panel in which users can opt-in and opt-out of certain ways in which their data might be used.

Additional Reading

The book *Value Sensitive Design* by Batya Friedman and David G. Hendry (Cambridge, MA: MIT Press, 2019) provides a great introduction to the principles underlying value-sensitive design.

Impact Assessment

Impact assessment focuses on assessing different options for action to resolve ethical risks. It involves making tradeoffs. ***Tradeoffs*** occur when alternative options for action are each supported by good reasons, and one must be chosen over another. For instance, to protect user data from being stolen, the RudiBrace team may rely on end-to-end encryption. But end-to-end encryption makes it harder to investigate cyber attacks on the system. This generates a tradeoff between security and user privacy.

There are different types of ethical tradeoffs:

- There is only one value at stake, but impacts are different for different stakeholders. For instance, economic benefit of social media engagement might flow to either the platform owner or to the users.
- An ethical risk affects only one stakeholder group, but different values are at stake. For instance, a new product function might increase user safety, but might also negatively impact user privacy.
- An ethical risk affects different stakeholders, and multiple values are at stake. For instance, restaurant closures during a pandemic might negatively impact the economic opportunities of people working there, but at the same time alleviate the health risks of vulnerable groups.

Common Strategies for Making Tradeoffs

There are three primary strategies for making tradeoffs:

- Taking into account considerations of distributive justice.
- Cost-benefit reasoning.
- The Delphi Method for stakeholder engagement.

Let's look at each of these in more detail.

Difficulties in Making Tradeoffs

Making tradeoffs involves analyzing and comparing the strength of competing interests of different stakeholders. This is often difficult. Tough cases may call for the assistance of experts in applied ethics. Like medical ethicists in hospitals who advise doctors and families on difficult treatment choices, applied ethicists are trained in advanced methods for identifying and resolving conflicts of interest.

However, since people may reasonably disagree about values, certain conflicts cannot be adequately resolved. This is so especially in cases where tradeoffs must be made between different values for the same group of people. For example, users of RudiBrace may reasonably assign different weights to the values of privacy and security. Some cases of disagreement might call for stakeholder engagement.

Distributive Justice

Competing interests are not limited to the ethics of emerging technologies. Interests conflict in government, law, and business all the time, to name but a few examples. To address conflicting interests, it can be helpful to take into account considerations of ***distributive justice***. This might include considering:

- Whether a product decision will affect different stakeholders differently
- Whether a product decision will disproportionately affect already disadvantaged people, whether it might exacerbate or mitigate existing disadvantages
- Whether a product decision will make certain people better off than others
- Whether a product decision will make certain people better off at the cost of others

The RudiBrace team considered whether to give people with care duties at home the possibility to opt out of using the product in the workplace. Not implementing this option would increase overall

social engagement. However, this would put people with care duties under social pressure. Considerations of distributive justice highlight that allowing employees to opt out of using RudiBrace would benefit care givers, who are under particular pressure, and so would help mitigate existing disadvantages.

Cost-Benefit Reasoning

Often, we might be able to resolve conflicts between competing interests by analyzing the strengths of these competing interests. This is where ***cost-benefit reasoning*** is helpful. This basically assumes that an action should provide a greater benefit than the cost required to implement the decision or action.

For example, if everything else is equal and you face a choice between benefiting two people with $1 each and benefiting a single person with $1, you ought to benefit the greater number of people. If you face a choice where one outcome would benefit one group by $5 and another choice would benefit a different group of similar people by $1, you ought generally to benefit the first group. Few real-life situations are as simple as these, though.

Often, the tradeoffs involved cannot be meaningfully reduced to money and the affected parties are heterogeneous in their characteristics. Nonetheless, you can often narrow down the options by breaking down complex scenarios into simpler comparisons. Other, tentative measures that might indicate the strength of a course of action might consider the number of people that might be affected and the relative intensity of the impact.

The RudiBrace team considered whether to give management indiscriminate and full access to all the data and analytics of the RudiBrace. They listed costs and benefits of this action in a table:

Costs	*Benefits*
Privacy infringement of all employees	Optimization of RudiBrace use
Overall reluctance to use RudiBrace	Greater social engagement for some employees

Based on this analysis, the RudiBrace team concluded that the potential benefits would not outweigh the potential costs, for on balance the negative impact would be more severe for a greater number of people.

The Delphi Method for Stakeholder Engagement

In Topic A of this lesson, Strategies to Resolve Ethical Risks, we discussed the importance of engaging stakeholders. The ***Delphi method*** is a particular methodology for engaging stakeholders that is particularly useful for resolving tradeoffs. The Delphi method uses iterative questionnaires to gain information from a panel of experts until a group decision or opinion is reached. Engaging stakeholders for making tradeoffs is an important complement to conducting an ethical analysis. Relying on ethical analysis alone risks ignoring how affected groups themselves view an option for action.

While preferences of stakeholders cannot determine the best option, stakeholder views provide an important input to decision-making. In particular, organizations should seek to identify a (potentially hidden) consensus among stakeholders. When it comes to difficult decisions, debate between stakeholders can become divisive, hiding common ground. Yet certain options for action may be acceptable for a broad range of stakeholders whose interests or values are normally considered to clash.

- Invite stakeholders to react to preferred options for action.
- Ask: Do stakeholders perceive options in line with ethical analysis, or do they prefer different options?
- Ask: Is there a "hidden consensus" among stakeholders about the best way to resolve an ethical issue?

Of course, consensus is not always possible on every issue. In some cases, the next-best alternative may be to favor the opinion of the majority, or to think carefully about whose interests should take priority in any given case.

Additional Reading

For more information on how to use the Delphi method for remote stakeholder engagement, visit **https://ahha.asn.au/news/using-delphi-method-engage-stakeholders-these-covid-19-times**.

ACTIVITY 3-3
Selecting Strategies to Resolve Ethical Risks

Scenario
As a member of the RudiBrace team, you want to identify actions that will help resolve the ethical risks posed by the product and its use. Earlier, the product team identified, among others, the following ethical risks:

- Security risk of data breach and leaking of personal data
- Privacy risk from misuse of personal data
- Transparency risk of having insufficient insight into collected data
- Efficiency risk from lower social engagement due to lack of data collected

1. **Can you suggest an action to remedy one of the ethical risks the team identified for RudiBrace?**

2. **Can you suggest a possible tradeoff for the action you identified?**

3. **Can you suggest a strategy for resolving the tradeoff?**

4. Investigate human-rights issues that are relevant to your industry.
   a) In a new browser window or tab, navigate to **https://www.business-humanrights.org/en/**.

b) Select **Companies** and scroll down the page until the Company Pages are displayed.

c) Use the **Sector** drop-down list to filter the list to show companies within your industry.
d) Select a company name, and review the information provided.
e) In the **Indicators** section, select **Human Rights Policy** to view the company's policy, if it exists.
f) If the **Human Rights Policy** was displayed in another tab, close the tab to return to the Company Page.
g) In the **Stories with associated response requests** section, select the most recent story and briefly review its contents.
h) In the **Company Responses** section, locate the company name, and if a response was provided, select **View Response**.
i) Review the response, and then select the browser's **Back** button twice to return to the Company Page.
j) If time permits, review at least one item in the **Associated top issues** and the **Top associated countries** sections.
k) Close the browser windows or tabs when you have completed your review.

TOPIC D

Avoid Problems in Ethical Decision-Making

Implementing ethical decision-making can be a complex process. In addition to each of the actions and decisions that need to be made throughout the process, there are some common pitfalls that can severely hamper the success of your ethical decision-making efforts.

Biases and Prejudices

Biases and prejudices can undermine the ethical decision-making of both individuals and groups. Seeing things as they really *are* isn't the same as seeing things as we want them to be, or as we have always supposed them to be. For this reason, it is important to remember that it is generally easier to believe things that already fit within our worldview, and that this doesn't necessarily mean that they are true.

Biases, which are prejudice in favor of or against a person, group, or thing as compared with another, usually in an unfair way, come in different forms. What they all have in common is that they skew our view of how important something is. Being biased towards or against something essentially means thinking that something is more or less important or relevant than it really is. For example, gender bias often means that a person's gender is treated as relevant in some way that typically disadvantages them.

Biases are a huge problem for moral reasoning because they skew our view both of what should count as a reason and how we should weigh different kinds of reasons. In this sense, checking our biases essentially means making sure that we don't mistakenly perceive things as reasons that aren't really reasons.

Groupthink

One important bias that is specific to the dynamics of reasoning in groups is ***groupthink***. Groupthink occurs when group members fail to voice opposing views. While agreement and harmony are positive features of groups in many contexts, groupthink is a danger to moral reasoning. Recall that good moral decision-making means going through different reasons, and ending up with the strongest. Good moral decision-making therefore involves challenging the reasons we come up with, to test how strong they really are. Since this cannot happen if everyone agrees, groupthink is a danger to moral reasoning.

Here are some strategies to avoid groupthink:

- Assign at least one group member the role of ***devil's advocate***, rotating at each meeting. A devil's advocate expresses a contrary opinion to provoke debate or to test the strength of an opposing argument.
- Discourage leaders from expressing an opinion early in the discussion, and even encourage them to absent themselves from group meetings to avoid excessively influencing the outcome.
- Set up several independent groups to work on the same problem.
- Invite outside experts into meetings.

Additional Reading

For more information, visit **https://www.verywellmind.com/what-is-groupthink-2795213**.

Establishment of a Legitimate Decision-Making Process

Ethical reasoning is an essential skill for those developing emerging technologies. But this is often not sufficient. As we have seen, different courses of action are sometimes supported by good

reasons. Being in a position to trade off competing reasons is to exercise power. Who should exercise this power? For example, should the passenger be the one who decides whether a self-driving car should sacrifice a pedestrian to save the passenger's life? Or should it be the government? Or the company that produced the car? Sometimes, what it means to do the right thing isn't just dependent on *what* different reasons we have. Sometimes, it's also a question of *who decides*.

Therefore, ensuring that the decision is legitimate is crucially important. To ensure that the decision is legitimate and is also perceived as such by stakeholders, organizations need a credible decision-making process. Markers of a sound process are:

- Decisions are grounded in ethical principles that the organization consistently applies to difficult cases.
- Affected groups had fair opportunities for input in the decision-making process.
- The process is informed by relevant information and expertise.
- The reasons for the selected option are clearly articulated.
- The decision-making process is transparent.
- Decisions can be criticized and appealed.
- Future decisions adapt to feedback and changing contexts.

Additional Reading

For an example of one approach to legitimate decision-making in organizations, see **https://www.managementexchange.com/hack/deliberative-corporation**.

Strategies for Legitimate Decision-Making

To ensure that organizations are sensitive to all relevant ethical reasons, all affected parties should have input into decision-making. Ways of doing this include consulting representatives and experts, deferring questions to government bodies, and working from ethical principles that all affected stakeholders can be reasonably expected to endorse.

Stakeholders will often have their own opinions on what is the right thing to do. A failure to give adequate consideration to these opinions can provide grounds for objection. This issue arises frequently with emerging technologies, as there are often wide disparities in power between those who design technologies and those affected by them.

Paternalism and Technocracy

Including affected parties safeguards against the risks of paternalism and technocracy:

Paternalism is often used to characterize situations where decisions about a person's own interests are made by someone else. This is often appropriate when the subject is a child, an animal, or a severely mentally impaired adult. But paternalism is a violation of respect—a failure to respect someone's autonomy—when such a person is capable of looking after her own interests.

Technocracy refers to rule by experts. It often carries a pejorative connotation and is used in contrast with *democracy*, rule by the many. A common complaint about emerging technologies is that they have enormous influence over contemporary life, but that influence is directed inequitably. Decisions are dominated by designers, engineers, and executives, at the expense of users, third parties, and governments. Complaints about the distribution of power are also common within technology companies, where opportunities to voice views may be—or may be perceived to be—unfairly distributed.

Additional Reading

For more information, visit:

- The risks of paternalism in medicine: **https://medicine.missouri.edu/centers-institutes-labs/health-ethics/faq/provider-patient-relationship**.

- The challenges of technocracy: **https://www.radicalxchange.org/media/blog/2019-08-19-bv61r6/**.

Ethics Washing

Ethics washing occurs when an organization engages in ethical marketing or communication without being motivated primarily by a commitment to doing the right thing. Instead of engaging in genuine reflection, making real commitments, and taking decisive action, organizations engaging in ethics washing merely make superficial statements, gestures, and empty promises. Sometimes companies put significantly more effort into virtue signaling, i.e., portraying themselves as being highly sensitive to ethical issues; for example, by setting up ethics committees without any real power, than into effective regulation, practices, and policies for which they could meaningfully be held accountable.

A good way to avoid the charge of ethics washing is to share or give up decision-making power about ethical issues. Organizations benefiting from developing an emerging technology will rarely be able to avoid the charge of conflict of interest when making important ethical decisions by themselves. Genuinely sharing or giving up power about ethical decision-making is one way of being seen as genuinely interested in reaching good ethical decisions.

ACTIVITY 3-4

Establishing an Ethical Decision-Making Process

Scenario

Companies like Ancestry or 23andme provide direct-to-consumer genetic testing based on a saliva sample. Genetic data is analyzed to generate reports relating to the customer's ancestry and genetic predispositions to health-related topics. In addition, these companies use the databases of genetic information to conduct medical research. The model has the potential to promote good outcomes, such as discovering new drugs. Yet the model also raises serious ethical issues about user privacy and data protection. For instance, in what ways are these companies allowed to use genomic data of their users? Is any kind of medical research permissible? Can the data also be sold to insurance providers to set insurance premiums?

The steps in this activity invite you to consider the processes that companies should use to resolve these questions.

Note: For more information on research ethics, visit **https://blog.23andme.com/23andme-research/protecting-people-in-people-powered-research/**.

1. **Which people or groups should be involved in setting company policy about these questions?**

2. **What are some ways that these companies could share or give up power to improve the legitimacy of these decisions?**

Summary

In this lesson, you performed practical applications of ethical decision-making. This skill is essential for every ethical technologist, as it is the cornerstone of incorporating ethics into emerging technologies.

At your workplace, which phase of ethical decision-making do you think will be the most challenging to implement?

Do you feel that any of the ethical pitfalls discussed are prevalent in your workplace? If so, which ones, and why?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

4 Identifying and Mitigating Security Risks

Lesson Time: 2 hours

Lesson Introduction

One of the most basic requirements of emerging technologies is that they are secure. Core to the idea of security is protecting users and other stakeholders from harm. Keeping data safe and protecting users and infrastructure from cyber attacks is an important element of cybersecurity.

But building safe and secure products with emerging technology cannot be reduced to the security of technical systems. Rather, products should protect all aspects of people's health and well-being. In this lesson, you will explore what it means to promote secure emerging technologies, how to identify security risks, what central tradeoffs can be expected, and how security risks can be mitigated.

Lesson Objectives

In this lesson, you will:

- Describe basic concepts of secure technologies.
- Identify security risks that affect emerging technologies.
- Identify the most prevalent tradeoffs concerning security.
- Apply methods and techniques to mitigate security risks.

TOPIC A

What Is Security?

Before you can fully grasp the relationships between security and emerging technologies, you need a solid base of understanding about each subject. In this topic, you will examine some basic security concepts.

Security and Emerging Technologies

Building secure products requires their creators to strive to avoid causing harm. Emerging technologies bring new ways of making our lives better and more comfortable. At the same time, they create new risks for harm. For instance, the more our lives depend on the sharing of personal data, the more a data breach exposing personal data can harm us. It might give criminals access to our personal banking environment or personal communications.

Harms can be intentional or non-intentional. Building a secure bridge, for instance, means among other things that it doesn't collapse when someone crosses it. When security fails, we consider the resulting harm to be unintended, in the sense that the collapsing bridge did not mean to harm the people crossing it, and neither did its designer.

Security also appeals to the idea of protection against intended harm. Malicious agents attempt to harm others through blackmailing, IP theft, identity theft, and so forth. Security against intended harm is often more difficult to achieve because it needs to take into account motives and strategies of potential malicious actors.

Is Security a Good Thing?

Security is not always a good thing. A malicious government might use very secure information systems to oppress its population. Criminals might also make use of secure communication channels to facilitate their crimes. Hence, the security of data and infrastructure is valuable only to the extent that it does not facilitate harm.

The security of data-driven technologies rests on three aspects:

- ***Confidentiality***: refusing unauthorized actors access to systems.
- ***Integrity***: preventing attackers from modifying data.
- ***Availability***: making sure that resources are available to authorized actors when they need to access the resources.

For individuals, being secure is intrinsically valuable. Personal security, which is an experience of being free from dangers or threats, is a basic condition of a good life. Security of the person entails physical security, including safety from injury; economic security, including safety from material deprivation; and psychological security, including safety from stress and disrespect.

Security of technical systems, by contrast, is often instrumentally valuable. It is not valuable for its own sake, but rather because it protects other things like personal security. The security of your banking system gives you the personal security of financial stability, enabling you to do things without fearing the loss of your money. This instrumental security can thus be a condition for personal, intrinsic security.

At the same time, there is also a risk for ***security fetishism***. People working with data-driven technologies often fixate on security as the only and paramount value. This sometimes diverts them from paying attention to other important values.

The Principle of No Harm

The principle of no harm asserts that the liberty of people should not be limited unless it results in the harm of others. Preventing harm to others is what security is about. Security, therefore, often creates a tension with liberty, because it is a ground for limiting people's freedom. The no-harm principle makes security stand out among other key ethical considerations as the primary ground that governments can invoke to limit the use of new technology.

The no-harm principle is a moral ideal that cannot always be reached in practice. Car drivers constantly expose each other, as well as bystanders, to small risks of deadly harm. Unfortunately, yet predictably, these risks sometimes materialize. In the US, there are more than 30,000 fatal car accidents every year. However, this risk is balanced somewhat by the increased mobility and freedom provided by the widespread use of automobiles in modern society.

At a minimum, the no-harm principle requires that we take all reasonable steps to minimize risks of harm. When a course of action imposes risks on others that surpass a certain threshold, the no-harm principle prohibits this course of action.

Bad Actors

Because of the open nature of many data-driven technologies, there is a growing variety of actors who aim to gain advantage by attacking these technologies. Knowing who these bad actors are and what their motives might be can help in addressing security risks.

- **Motives**: The following are typical motives of bad actors:
  - **Financial gain**. Bad actors might compromise systems or blackmail users to gain money.
  - **Political gain**. Bad actors might attack systems to cripple a political group or even a nation, for instance in the case of cyberwarfare attacks.
  - **Espionage**. Bad actors might compromise systems to gain access to confidential information such as technological specs or government documents.
  - **Recognition or revenge**. Bad actors might attack systems to gain reputation among their peers or take revenge on the basis of perceived personal harm.
  - **Entertainment**. Some bad actors attack systems as a pastime, simply to see whether they can breach security measures.
- **Bad Actor Types**: The following are the most common types of bad actors:
  - **Script kiddies**: low-skilled or inexperienced hackers who use existing scripts created by other hackers to attack systems.
  - **Professional hackers and cyber criminals**: higher-skilled hackers who attack systems for a living, often specialized in particular domains such as finance.
  - **Cyber terrorists**: attackers who compromise systems to cause fear.
  - **State-sponsored hackers**: attackers who act on behalf of governmental cyber strategy, sometimes involved in cyberwarfare.
  - **Hacktivists**. Hackers driven by a purpose other than personal gain, such as social change.

Some people who aim to compromise systems see themselves as ethical hackers. Ethical hackers attack systems with the purpose of disclosing vulnerabilities to the system stakeholders.

ACTIVITY 4-1
Discussing the No Harm Principle

Scenario
In this activity, you will discuss how much risk products using new technologies can impose on others without violating the no-harm principle.

1. How safe do driverless cars need to be to satisfy the no-harm principle?

2. Consider the RudiBrace example. Imagine that bad actors are trying to access the RudiBrace to harm the users. Discuss what types of bad actors might want to attack RudiBrace, and what reasons would motivate them.

TOPIC B

Identify Security Risks

We all face some security risks as we use modern technology, like identity theft and credit-card fraud. Emerging technologies are also subject to security risks. In this topic, you will identify security risks that commonly affect emerging technologies.

Sources of Security Risks

Not every security risk is an attack. Security risks have different sources, of which there are four main types:

- **Human mistakes**. Designers and engineers are fallible like all human beings and sometimes make mistakes. For instance, a mistake can be a software bug that causes a system to crash. Especially for vital infrastructure, like banking software, human mistakes can bring serious security risks. Human mistakes are the most common source of security risks.
- **Technical failures**. Sometimes a technology might fail without an attributable human mistake. A power outage, for instance, can trigger technical system failure that cripples its functionality.
- **Natural disasters**. Natural events like floods, storms, or earthquakes can cripple technical infrastructures and cause security risks. Consider, for instance, the Fukushima nuclear disaster, in which a tsunami caused security risks in a nuclear reactor.
- **Malicious attacks**. Criminals, malicious governments, and bad actors might have motives to attack a data-driven technology. They will use certain strategies or techniques to cripple a system's functionality and to harm its users.

Cyber Attacks

For data-driven technologies, malicious cyber attacks are a particularly important source of security risks. The following types of attack are most common:

- **Malware**: Any type of executable computer code used for malicious purposes. Some of the most common include viruses, which replicate as they spread between files or computers; worms, which—unlike viruses—don't attach themselves to files but also replicate across systems; trojan horses, which include malicious payloads disguised as legitimate software; spyware, which spies on users' behavior without their knowledge or consent; and ransomware, which restricts access to data or systems and demands payment for access to be returned.
- **Denial of Service (DoS attack)**: An attack that prevents people from using a service. This can come in many different forms. For example, a website or server might be taken down, so people cannot use it as intended. Or some flaw in application code might be exploited to block important services and make them inaccessible.
- **Zero-day exploits**: Attacks that exploit unknown system vulnerabilities. Zero-day exploit attacks pose some of the greatest security risks, because they are difficult to protect against, and often go unnoticed.
- **War driving**: Searching for Wi-Fi signals and hijacking them. Since Wi-Fi signals often extend beyond, say, a particular building, and anyone within a certain perimeter of the premises might be able to pick up the signal.
- **Passive wiretapping**: An attacker intercepts information on your organization's network. This can give attackers access to all (unencrypted) data transmitted via your network.
- **SQL injection**: SQL is a popular database format. SQL injection is an attack where the requests or queries to one of your organization's SQL-databases are interfered with. This allows attackers to view data from the database they are not normally able to see.

AI-Specific Cyber Attacks

Some cyber attacks specifically target the functioning of models used in AI systems. The following two are most common:

- **Evasion attacks**: In evasion attacks, an input to a machine learning algorithm in deployment is modified to avoid correct classification. For instance, spammers may craft messages that go undetected by spam filters. Evasion attacks exploit blind spots in an algorithm that is already deployed.
- **Poisoning attacks**: Poisoning attacks target algorithms during training. A poisoning attack attempts to contaminate training data to influence the performance of a model. The case of making a self-driving car that makes mistakes reading road signs is an example. In poisoning attacks, training data is manipulated to teach the AI system the wrong things and compromise its decision-making process.

It is often surprisingly easy for attackers to introduce malicious input into AI training models. This is because many modern AI systems are trained by taking input from public sources on the Internet. By connecting to servers on the Internet, and manipulating network traffic data, attackers can influence the training data and ultimately cause AI systems to fail to work properly.

This kind of introduction of malicious data into training can be difficult to detect and difficult to mitigate. Since this can create serious risks to safety and security, it's all the more important to be aware of how adversarial attacks might compromise AI systems. It is also useful to **preserve the provenance of data**, i.e., documenting the data origin, any manipulations applied to the data, and where the data moves over time.

Regulations and Standards

There are established standards that can help organizations understand what is required of them in meeting security concerns. The most important family of standards to know is ***ISO/IEC 27000***. This family of standards provides best practice recommendations on information security management.

The core standard in this family to be aware of is **ISO/IEC 27001**. The standard contains requirements for establishing, implementing, maintaining, and continually improving an information security management system (ISMS). The aim of an ISMS is to help organizations make their data and IT infrastructure more secure. Organizations meeting the standard's requirements can get certified by an accredited body following an audit.

You should also be aware of **ISO/IEC 27002**, a code of practice for information security controls, and **ISO/IEC 27017**, a security standard developed for cloud service providers and users to make a safer cloud-based environment and reduce the risk of security problems.

Note: ISO is the International Standards Organization, and IEC is the International Electrotechnical Commission.

Key Standards

An overview of other important standards is listed below.

Standard	*Domain*
Payment Card Industry Data Security Standard (PCI DSS)	Security and privacy standards for payment card holder data.
National Institute for Standards and Technology (NIST) Special Publication (SP) 800-53 Rev. 5: **Considerations for Managing IoT Cybersecurity and Privacy Risks**	Security of information systems used by the US Federal government; any organization that works with the federal government needs to comply.
NIST **Cybersecurity Framework** (CSF)	Cybersecurity in private sector organizations.

Standard	*Domain*
NIST Internal/Interagency Report (NISTIR) 8228: **Considerations for Managing IoT Cybersecurity and Privacy Risks**	Cybersecurity and privacy in IoT devices.
Center for Internet Security Critical Security Controls (CIS CSC)	Cybersecurity and customer privacy for all organizations using network software infrastructure to process customer data.
ISO/IEC 27001: **Information Security Management**	Information security in offline networks and cloud-based systems.
ISO/IEC 27017: **Code of practice for information security controls based on ISO/IEC 27002 for cloud services**	
ISO/IEC 27018: **Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors**	
MITRE ATT&CK® (Adversarial Tactics, Techniques, and Common Knowledge)	Adversarial testing of cybersecurity and current database of attack types.
IEEE P7000 series standards	Processes to include consideration of human ethical values, bias, fairness, privacy, and security in the use of data-driven technology.

Additional Reading

For more information, visit:

- NIST SP 800-53 Rev. 5: **https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf**
- NIST CSF: **https://www.nist.gov/cyberframework**
- NIST NISTIR 8228: **https://csrc.nist.gov/publications/detail/nistir/8228/final**
- CIS CSC: **https://www.cisecurity.org/controls/**
- ISO/IEC 27001: **https://www.iso.org/isoiec-27001-information-security.html**
- ISO/IEC 27017: **https://www.iso.org/standard/43757.html**
- ISO/IEC 27018: **https://www.iso.org/standard/76559.html**
- MITRE ATT&CK: **https://attack.mitre.org/**
- IEEE P7000 series: **https://ethicsinaction.ieee.org/p7000/**

Risk Libraries

Dealing with security risks is a constant process of learning for organizations. New threats and adversarial activities arise constantly, which means that systems have to stay up to date. To promote internal learning, organizations can maintain threat and ***risk pattern libraries***. These are databases with reusable use cases that outline particular threats, weaknesses, and countermeasures, which in turn are the basic building blocks of threat models.

Organizations can internally collect cases as well as draw from available external threat and risk pattern libraries. For instance, the Swiss National Centre for Cybersecurity maintains a monthly updated catalog of cybersecurity risks.

Security Risk Identification

The most common way to identify security risks is conducting a risk assessment. At their core, quantitative risk models can be broken down in two components:

- An estimate of the probability that the risk will materialize.
- An estimate of the impact of a risk in case it materializes, usually expressed as a monetary value.

Multiplying the estimated impact with the probability yields the **expected impact** of the risk.

For example: The probability of user data leaking has been estimated at 5%. The impact has been quantified at $500,000. As a result, the expected impact of the risk is $25,000.

Quantitative risk analysis is useful for making risks comparable with each other and for making decisions about appropriate steps to take in mitigating the risk.

The main ethical challenge with quantitative risk analysis is not to lose sight of the nuances of ethical decision-making. Some risks threaten to violate basic human rights. But the special gravity of rights violations cannot readily be translated into a dollar figure. Therefore, distinctions between risks violating rights and other negative impacts are not readily apparent in quantitative risk analysis.

The significance of risks may manifest differently for different stakeholders, or risk roles. Yet, the distribution of risks across different stakeholders is not always taken into account in standard quantitative risk analysis.

A connected challenge consists in translating impacts into dollar values. For instance, the costs of a data leak might be different for your organization than for users themselves. Some impacts might be challenging to translate into monetary value, such as risks to reputation and risks of physical harm.

Risk Roles

To address the challenges with quantitative risk analysis, you should analyze the risk positions of different stakeholders for each risk you take.

Risk roles are ways that people can be involved in a risky situation, such as:

- Being exposed to a risk.
- Being someone who benefits from the risk being taken.
- Being a decision maker.

Consider the risk of someone breaking into your laptop by guessing your password. In this case, you have all three risk roles: You are exposed to the risk, you make the decision about whether or not to use a safe password or two-factor authentication, and you benefit from taking a greater risk by using a more convenient password that is also easier to guess. Hence you are in the center of the Venn diagram below.

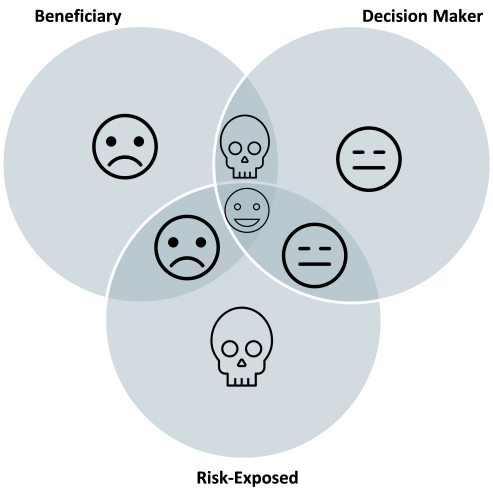


Figure 4-1: Risk roles.

Four types of risk positions are especially problematic:

- **Beneficiary and Decision Maker**. From the viewpoint of risk management, the role combination is highly problematic since it gives rise to risk-inducing incentives. Example: Social media companies benefit from collecting user data, and they decide which data to collect, but it is not their own data that gets leaked in case the data gets stolen—it is their users' data.
- **Only Risk-Exposed**. People are exposed to a risk that is decided by others and from which only others gain advantages. For instance, future generations are at risk of climate change, but it is current generations who benefit most from climate change and make decisions about how to manage it.
- **Beneficiary and Risk-Exposed**. Users often find themselves in the situation where organizations make decisions that benefit them as well as expose them to risk. For instance, governments make decisions about whether to use nuclear energy, which benefits citizens but also puts them at risk.
- **Beneficiary Only**. For the individual, this is a favorable position, but it can mean free-riding on the risks borne only by others. For instance, those who use vaccines free-ride on the volunteers who participate in clinical trials and risk suffering unknown side effects.

Security Risk Analysis and Ethical Rights

Risk roles do not make a difference to the gravity or magnitude of the risks imposed. But if we find ourselves in a favorable risk position, we should nonetheless apply a higher bar. For instance, higher consumption today may expose future generations to risks due to climate change. This puts future generations in a very unfavorable risk position, creating a higher bar for justifying high consumption today.

ACTIVITY 4-2
Identifying Security Risks

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. List at least one security risk associated with the product and the risk roles that your organization has with respect to these risks.

2. Are there any groups with respect to these risks that are in problematic risk roles?

TOPIC C

Security Tradeoffs

Security is often thought of as a balancing act between control and freedom. For example, consider the benefits and drawbacks of using a weak or strong password. It's much easier to remember a weak password, but it's also much easier to guess or crack it. Finding the balance between security and ethical considerations is a fundamental skill required by ethical technologists.

Potential Security Conflicts

Security should be a basic feature of emerging technologies. However, greater levels of security may come into conflict with other ethical considerations, such as fairness, privacy, and freedom. In particular, there are four different tradeoffs—with privacy, accountability, fairness, and environment—that your organization might face when considering security measures. These challenges do not have simple answers, and they require teams to both think carefully about the tradeoffs involved in matters of security and be prepared to defend their choices.

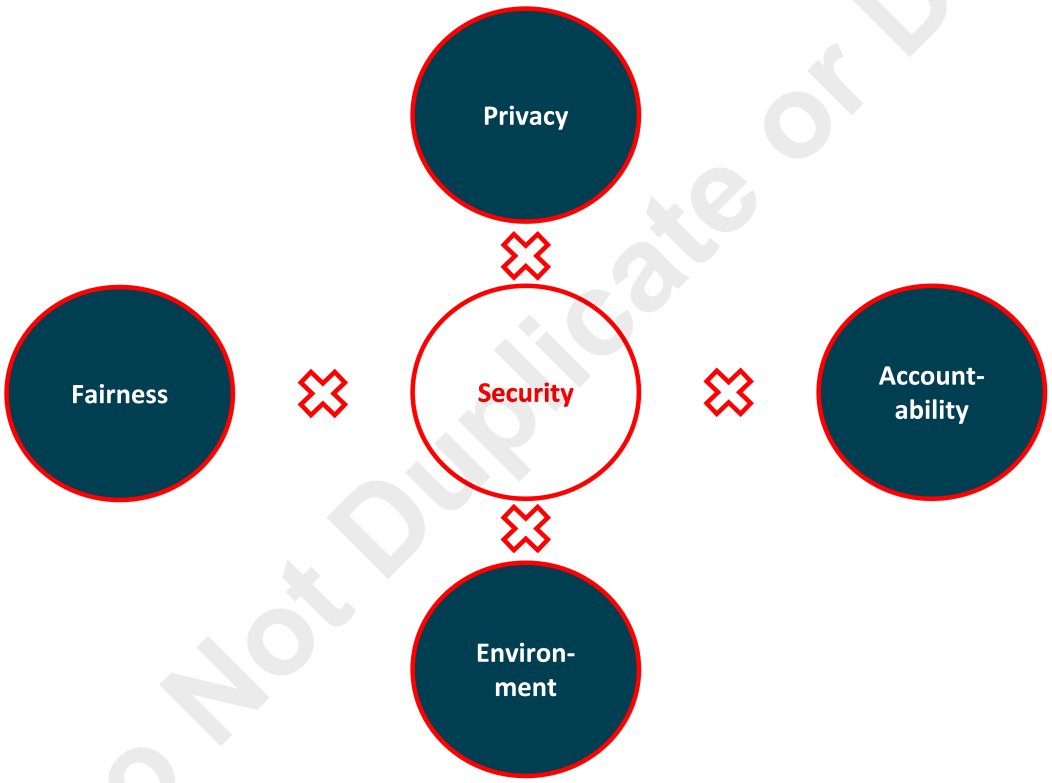


Figure 4–2: Potential security tradeoffs.

Privacy Tradeoffs

Security often strengthens the privacy of data-driven technologies. For instance, data encryption leads to greater security of services and provides privacy to users. However, security and privacy can conflict; for example, when national-security aims of protecting the population against criminals conflict with privacy-preserving technologies that enable criminals to communicate.

Consider the Apple vs. FBI case. After a terrorist attack in San Bernardino, California, left 15 people dead and 22 wounded in December 2015, the FBI asked Apple for assistance in unlocking the suspect's iPhone. To prevent unauthorized access, Apple had programmed this iPhone version to automatically delete all user data once 10 unsuccessful attempts had been made to guess its PIN code. The FBI demanded that Apple create and provide software to bypass these security protocols. Apple objected to this request, claiming that creating a device backdoor would threaten the security of all users and give governments unchecked power to invade user privacy. Although a federal judge sided with the FBI and ordered Apple to comply, the FBI ultimately dropped the case when a third party volunteered to create the software needed to unlock the iPhone.

The case raises the important issue of ethical tradeoffs when it comes to security. There is, in principle, no limit to the degree of security one can apply to a technological artifact. Any kind or level of security we set can always be increased in some way. For instance, a smart phone can be programmed to automatically delete data after 10 unlock attempts; it can also be set to begin this process at 5 attempts, or 3 attempts—or to self-destruct entirely. It can require two-factor authentication, three-factor authentication, or authentication with any number of factors. These possibilities force us to ask, how secure is secure enough?

In the case of Apple versus the FBI, the security of user devices came into conflict with national security and public safety. Increasing the security of the iPhone afforded great protections for user privacy, but those protections also made it easier for malicious actors to avoid detection and accountability.

Accountability Tradeoffs

The democratization of AI (for example, through open-source AI technologies) may also entail the proliferation of AI to unscrupulous actors. Security measures are often targeted at bad actors and should therefore remain unknown to these actors. This means that security often goes hand in hand with secrecy—the need to keep certain information confidential. Yet, secrecy might conflict with accountability, because it makes the basis for security-driven decisions non-transparent.

Consider the case of the Panama papers. In 2016, a collective of investigative journalists published information on offshore financial assets, which were created by and managed by the Panamanian law firm and corporate service provider Mossack Fonseca. The publication was based on a giant leak of financial information coming from an anonymous whistle blower or hacktivist.

The publication exposed a large-scale global system of money laundering, political corruption, and tax avoidance. The leak of the Panama papers was made possible by vulnerabilities in the security measures of the Mossack Fonseca firm. Yet, it also helped hold powerful individuals and organizations accountable for dubious or illegal practices.

Fairness Tradeoffs

Increasing security of data-driven technologies often means putting additional burdens on people. Yet, these burdens can exclude people from opportunities they would have enjoyed with a lower security level in place.

Consider new regulations after the financial crisis of 2008-2009 to make mortgages more secure. To make banks more resilient, jurisdictions around the globe introduced new borrowing requirements. As a result, borrowers need to contribute more capital, show a higher and more consistent income, and have better credit scores to obtain credit. This excludes less affluent people from obtaining the credit needed for a home mortgage.

Environmental Tradeoffs

Security measures can lead to higher use of energy and therefore lead to harm to the environment. Because security protocols can be computationally intensive, they sometimes require disproportionate amounts of energy.

Consider the case of Bitcoin. Security is one of the key features of the technology of Bitcoin. The underlying blockchain uses a cryptographically secured system of incentives that allows for the secure transfer of funds without risks like the possibility of double spending, meaning that an actor is not allowed to spend the same amount of cryptocurrency twice. The secure incentive system is also known as the proof of work system. This system relies on network nodes (miners) to compute the solution for an increasingly complex mathematical problem.

Hence, the proof of work system greatly increases the security of the system against possible attacks. However, it also uses enormous computational power. In 2019, it was claimed that Bitcoin used as much energy as the entire country of Switzerland.

Additional Reading

For more information, visit **https://www.bbc.com/news/technology-48853230**.

ACTIVITY 4-3
Identifying Security Tradeoffs

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. **Consider at least one security risk associated with the product and ways of addressing the security risk.**

2. **Which tradeoffs with privacy, accountability, fairness, or the environment do these strategies create?**

TOPIC D

Mitigate Security Risks

After you have identified the security risks for an emerging technology and considered the tradeoffs required to balance security with other values, you can identify methods to mitigate the risks.

Methods for Mitigating Security Risks

Security risks can be mitigated in different ways:

- By making your own organization more secure. This means establishing a baseline systems behavior and creating possibilities for rapid response.
- By securing data in storage and in transit. This means applying techniques such as encryption and secure network protocols.
- By analyzing and moderating security risks. This means continuously running threat models and testing and analyzing potential attack strategies.

Establishment of Baselines for System Behavior

Identifying abnormal system behavior and defending against attacks requires knowing how a system is normally supposed to operate. For instance, if you observe that a system is suddenly drawing twice as much electricity as it normally needs, that may be a sign something is amiss. What counts as normal operation will often be different for each system. For instance, some systems may have specific inputs and outputs, and an unexpected change in input or output will be cause for concern. Other systems rely on variable inputs or outputs, and these kinds of variations may not necessarily signal abnormal behavior. Thus, it is important for a team to reflect carefully on what counts as normal behavior, what symptoms are reliable indications of abnormal behavior, and how these symptoms can be effectively monitored.

Common baseline metrics include:

- Bandwidth consumption
- Software versions
- Upload and download times
- Task completion times
- User access and behavior
- Key performance indicators

Baseline metrics and monitoring systems can be considered successful when they can properly diagnose abnormal behavior and alert the relevant staff. More advanced baseline behavior monitoring systems can also isolate and automatically correct certain problems without human intervention.

Designation of Rapid Response Teams

Preparing for safety and security incidents requires designating personnel to perform particular tasks in the event of a breach or malfunction. For instance, a Cyber Security Incident Response Team (CSIRT) includes:

- Investigators to identify causes of abnormal behavior.
- Security specialists trained to fix systems or install new protections.
- Help desk staff to assist users who may be affected by an incident.
- Crisis communications experts to provide information to stakeholders.
- Managers to coordinate the entire response.

A response team can be a full-time business unit, a unit assembled from existing staff with different primary roles, or outsourced to specialists.

Protection of Stored Data

As mentioned earlier, data protection generally has three main objectives: ***confidentiality, integrity, and availability (CIA)***.

Ways of achieving CIA objectives include encryption, access control, physical barriers, and destruction.

- **Encryption**: Data is stored in a format that cannot be understood without a decryption key.
- **Access control**: Only certain individuals are permitted to access or modify the data.
- **Physical barriers**: Data is stored in a secure physical location, where unauthorized users or intruders will face difficulties in attempting entry.
- **Destruction**: Data that is no longer needed is automatically deleted or overwritten, especially temporary databases containing aggregated data.

Protection of Data in Transit

Data is mobile by nature, and securing data as it is transmitted from place to place and person to person is even more challenging than securing data in storage. It is important for a team to consider all the ways in which its data might be moved, how each move might represent a security risk, and how to mitigate these risks. Fortunately, the CIA framework of confidentiality, integrity, and availability can also help with securing data in transit.

Network protocols that protect data in transit include ***Secure Sockets Layer, Transport Layer Security (SSL TLS)***, which is often used for web services, and ***Secure Shell (SSH)***, which is often used for remote access. SSL TLS is a form of ***link encryption***, meaning that data is decrypted at each intermediary point. SSH is a form of ***end-to-end encryption***, meaning that data is decrypted only at its endpoints. End-to-end encryption is generally more secure than link encryption.

Both forms of encryption rely on ***digital signatures***, which use cryptography to identify the sender and receiver of the data as well as its contents. Digital signatures help to ensure both data confidentiality and data integrity, as they can verify whether data has been altered in transit.

Ensuring the availability of the transmitted data mainly requires maintaining the medium of transmission. This can be challenging when the number of users or bandwidth is variable, or when attackers are attempting to interfere with transmission. Techniques for ensuring availability include:

- Load balancing across multiple servers.
- Creating redundancies and failovers.
- Purchasing DDOS mitigation protections from specialist firms.

.

Threat Modeling and Analysis

Threat and risk pattern libraries feed into threat intelligence and forensic analysis. These techniques consist of at least three components:

- Profiles of potential attackers, including their potential aims and strategies.
- A library of threats and vulnerabilities.
- An abstraction (simulation) of a system.

One can compare these techniques with techniques of defense in a game of chess, in which each player tries to model the potential next moves of his or her opponent and mitigate these with counter moves.

Some frequently used techniques are:

- ***Vulnerability scoring***, which provides a metric to assess the severity of a potential threat or vulnerability. It consists of a base score (based on technical findings), a temporal score (based on external fluctuations that might worsen the vulnerability), and an environmental score (which takes into account the particular impact of a vulnerability on an organization).
- ***Cognitive security***, which is the use of AI to detect security threats. It implements big data analytics to find connections and vulnerabilities in systems that are very difficult for humans to detect.
- Attack trees, which map out the different routes to compromising a system or asset.
- Visual, Agile, and Simple Threat (VAST) modeling, which models operational and application threats.
- Challenger models, which refers to developing alternative models—such as with a different method or data—and comparing performance with the original model to identify security differences.
- Security information and event management (SIEM) systems, which provide an overview of all security issues and events in an environment.
- Comparing system attributes with provisions in API agreements and negligence law to assess potential liabilities.
- Reviewing the potential limitations of training data or models in AI systems.

Breach and Attack Simulations

To identify and mitigate unknown safety and security risks, organizations can perform breach and attack simulations on a regular basis. These are automated solutions for manual exercises of penetration testing, in which one internal team tries to defend the security of a system while another tries to attack it.

Simulations are based on known malware attacks and cybersecurity breaching strategies. A program automatically uses these to attempt to breach the security of a system. Once breaches are found, the simulation automatically proposes mitigating actions and recommendations for follow-up. These methods are sometimes called black-box multi-vector testing, as they seek to penetrate the system from different angles without reading the underlying code.

Organizations should also elicit user feedback on security issues and revisit their security practices on a regular basis.

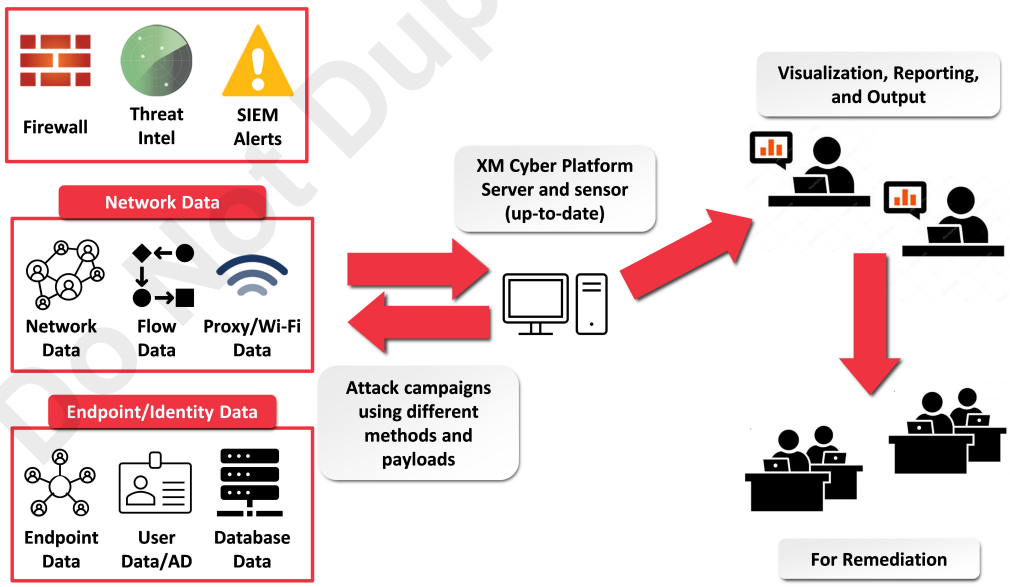


Figure 4-3: Breach and attack simulations.

ACTIVITY 4-4
Identifying Security Risk Mitigation Techniques

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. **For the RudiBrace or your own product, can you describe some baselines for normal system behavior that can be used to detect security problems?**

2. **For a security risk associated with the product, list some techniques that can help to mitigate the security risk.**

Summary

In this lesson, you examined the relationship between security and emerging technologies, paying particular attention to identifying and mitigating security risks while remaining cognizant of the tradeoffs that might be required to balance the need for security and the needs of the organization and society itself.

What do you think is the primary security risk for emerging technology at your workplace?

What types of tactics do you think are best suited to mitigating this risk?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

5 Identifying and Mitigating Privacy Risks

Lesson Time: 1 hour, 30 minutes

Lesson Introduction

Many emerging technologies are driven by data. Machine learning relies on rapid access to large amounts of data, the Internet of Things (IoT) has brought us interconnected devices, and the collection and use of vast amounts of consumer data has become the norm for many types of companies. Companies have to take responsibility for how they handle the personal data they collect. At the center of these developments are concerns about privacy. In this lesson, you will identify and mitigate privacy risks.

Lesson Objectives

In this lesson, you will:

- Describe the basic requirements of privacy.
- Identify sources of privacy risks.
- Describe the most prevalent tradeoffs concerning privacy.
- Apply methods and techniques to mitigate privacy risks.

TOPIC A

What Is Privacy?

As emerging technologies increase connections between people, and as tech companies, governments, and private actors gain access and control over increasing amounts of data, privacy has become an enormous concern. For example, online communication and the increased connectivity through the Internet of Things (IoT) have raised a host of privacy and data security issues. Think of home assistants that record everything that goes on in our personal spaces—our conversations with friends and family and our daily activities—and it soon becomes clear why people are concerned with maintaining their privacy. In this topic, you will investigate the basic concepts of privacy.

Privacy and Personal Information

Privacy means being free from observation or intrusion of our personal lives by others. It is a multifaceted concept. There is privacy of behavior, privacy of thoughts, privacy of the body, local privacy (not having our location revealed), decisional privacy (not having our decisions and actions interfered with), and ***informational privacy***, which is the ability to control who has access to personal information, and to what extent.

One way to think about privacy is as limiting what we want others to know, and perhaps even control, about us. We might say that, as individuals, privacy is instrumental to our ability to be ourselves. It allows each of us to develop our individual identity as an autonomous person. Being in control of our personal information is part being in control of our own lives and agency. The "others" might be other people, law enforcement, governments, and corporations. At the individual level, privacy is intertwined with our ability to control who has information about us, and what they do with it.

Personal information is any information that can directly or indirectly be linked to someone's identity. Personal information may be abused. For example, bad actors may use sensitive personal information for malicious purposes, like blackmail or extortion.

The need for privacy always depends on the context within which information is shared. For instance, we do not want our health data to be accessible to criminals, but are happy for the same information to be easily accessible to doctors in emergency situations. This aspect of informational privacy is also called ***contextual integrity***.

Privacy and Emerging Technologies

We have reasons to care about privacy, not only as private individuals but also as creators of technologies that might impact privacy. At the organizational level, companies often deal with data privacy. This means collecting, storing, and using data and information about people responsibly. This not only means handling data in compliance with laws and regulations, but also in accordance with people's interests and expectations. In this sense, data privacy is connected to data ethics. This means 'doing the right thing' with people's data, and handling it responsibly—for example, by carefully considering how using people's data might affect their rights and interests, and society more broadly.

Data privacy protection also faces critical challenges: the potential to abuse the technology used to store and process data may pose severe risk, to which we'll turn next.

Privacy and Power

Privacy is also closely connected to **power**. You might think of George Orwell's *1984*, in which Big Brother's constant surveillance controls people's behavior. If people know that they are being watched and observed, this will affect their behavior. Having one's privacy protected is therefore essential to the ability to be oneself, without having one's behavior influenced by knowing that one is being watched. This can also go the other way—think of how the privacy of Internet anonymity has emboldened online trolls.

States or large corporations sometimes violate the privacy of individuals. In 2013, the former NSA employee Edward Snowden revealed that the U.S. government was using extensive surveillance capabilities to spy on its citizens. This opened people's eyes to the fact that our private communications can be intercepted and used for purposes we might not intend or even be aware of.

More recently, the Chinese government has initiated a national program known as the *social credit system*. This system aims to collect data about individuals on a massive scale, and to use this data to give people a score as trustworthy or untrustworthy. On this basis, certain infractions have put Chinese citizens on a blacklist, which in turn barred these citizens from purchasing train tickets, renting hotel rooms, or taking out credit. The system relies on private-public partnerships, including commercial technologies such as Alibaba's Sesame credit. Such collusions of state and private power can create situations in which private communications are constantly subject to potential privacy violations. Another tension to bear in mind in this context is that between fair competition and corporate hegemony; for instance, through data assets.

Privacy in Different Cultures

Note that the value of privacy also varies across different cultures. For example, in Western cultures, individual privacy is often seen as an intrinsic good. Eastern traditions, such as Confucianism, by contrast, generally prioritize the collective over the individual good. In these contexts, the notion of collective privacy—for example, of a family—might therefore carry more weight than the notion of individual privacy.

ACTIVITY 5-1
Discussing Privacy Basics

Scenario
Use the following questions to drive your discussion of the basic concepts surrounding privacy.

1. Have you ever experienced an invasion of privacy through an emerging technology? Did you take any measures in response?

2. Do you think privacy matters for its own sake, or because it protects other important interests?

TOPIC B

Identify Privacy Risks

Before you can act to mitigate privacy risks, you need to know what risks you are up against. In this topic, you will identify privacy risks and their sources.

Sources of Privacy Risks

Privacy is at risk if user data is collected, shared, or processed in ways to which users did not meaningfully consent. To unpack the different types of privacy risks, it is useful to differentiate between risks arising with first versus third-party data and risks arising with primary versus secondary uses of data.

- **First-party data** is collected directly from users or any other audience with whom you work.
- **Third-party data** is collected from audiences with whom you do not have a direct relationship.
- **Primary use of data** is using data according to the stated purpose for which it was collected.
- **Secondary use of data** is using data beyond the original stated purpose.

Risks Arising with Data Collection and Use

First-party data is data an organization collects directly. The main risks you should be aware of are:

- **Collecting data without the users' knowledge**. A common example of this is web browser cookies.
- **Processing data in ways to which users did not meaningfully consent**. For example, a payday loan provider may use information like the type of device a customer is using or their email provider to set interest rates.

Third-party data is data collected from audiences with whom you do not have a direct relationship. It may often be received from other organizations who do have direct relationships with the data subjects. Risks of collecting and using third-party data include:

- **Adherence to privacy standards is unverified**. When you do not collect the data yourself, it can be difficult to verify whether the data was collected according to applicable privacy standards.
- **Data quality is unverified**. When you do not collect the data yourself, it can be difficult to determine whether the data is accurate, complete, or representative.

Secondary use of data is use of data beyond the original intent with which it was collected. The most common sources or risks of secondary use are:

- **Monetizing data**. For instance, introducing advertisements on a previously ad-free website that crowd-sources content, or selling user health data that users revealed about themselves to connect to other users with similar health conditions.
- **Cross-correlating datasets**. For example, linking a dataset with personal contact information of your customers to a dataset with purchase information. Cross-correlation can reveal facts about users that they do not want to divulge, such as inferring whether they are pregnant, perhaps before users know themselves.

Privacy Regulations Around the World

Regulation and standards can help organizations minimize the risks involved in collecting and using personal data and protect users from having their personal data exposed.

Data privacy regulation applies to almost every product using an emerging technology. In addition, many jurisdictions have developed specific regulation applicable to more narrow domains.

The map below shows important data privacy regulation for emerging technologies. Note that most of these regulations apply not only to organizations headquartered in the respective countries, but also to all companies doing business in the respective jurisdiction. Therefore, organizations making their product available to customers globally do well to comply with the most stringent rules and regulation.

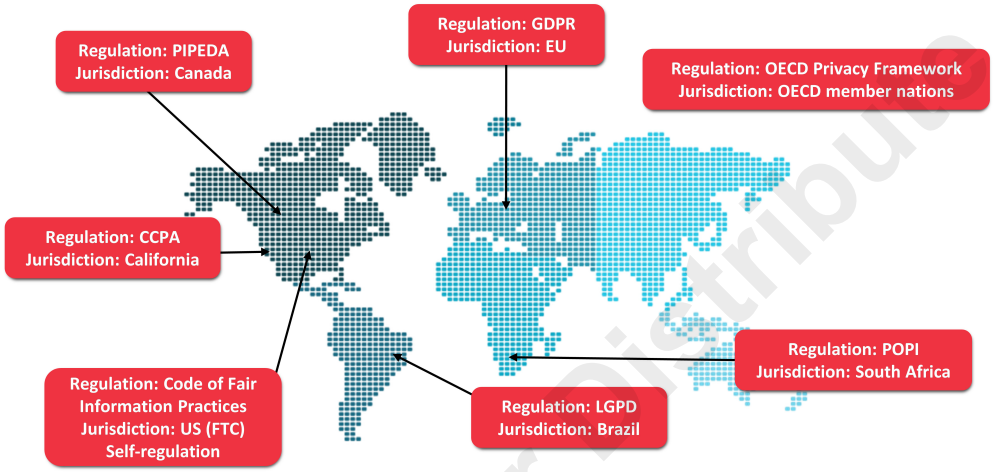


Figure 5-1: Privacy regulations around the world.

Additional Reading

For information about these regulations, visit the following web pages:

- General Data Protection Regulation (GDPR): **https://gdpr-info.eu**
- US FTC Code of Fair Information Practices: **https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission**
- OECD Privacy Framework: **https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf**
- California Consumer Privacy Act (CCPA): **https://oag.ca.gov/privacy/ccpa**
- Personal Information Protection and Electronic Documents Act (PIPEDA): **https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/**
- Protection of Personal Information (POPI): **https://popia.co.za/**
- Brazilian General Data Protection Act (LGPD): **https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/**
- World Legal Information Institute Database for International Privacy Law: **http://www.worldlii.org/int/special/privacy/**

The EU GDPR

The EU's ***General Data Protection Regulation (GDPR)*** is a pan-European data protection law and is widely adhered to by organizations around the world. The GDPR expands the rights of individuals to control how their personal data is collected and processed and places a range of new obligations on organizations to be more accountable for data protection. We cannot cover all of this complex piece of regulation in this course. However, it is worthwhile to understand some general principles underlying the regulation to inform your privacy strategy:

- **Transparency**: State what data you collect and the reason you are collecting it.
- **Purpose limitation**: Only collect the data that you need to achieve your stated purposes.

- **Data minimization**: Pursue the way to achieve your stated purpose that requires the least collection and processing of personal data.
- **Accuracy**: Take all reasonable steps to erase or rectify data that is inaccurate.
- **Storage limitation**: Delete personal data when it is no longer required to achieve your stated purposes.
- **Confidentiality**: Process data in a way that ensures security of personal data.

Privacy Regulations in the United States

Unlike other jurisdictions, the United States relies mainly on industry-specific regulation to protect privacy. The following table shows some noteworthy privacy regulations in the United States. Consult with a legal expert to confirm specific regulation applicable to your product.

Name of Regulation	*Domain and Jurisdiction*
Health Insurance Portability and Accountability Act (HIPAA)	• Domain: Healthcare information • Jurisdiction: United States
Children's Online Privacy Protection Act (COPPA)	• Domain: Data privacy for children • Jurisdiction: United States
Algorithmic Accountability Act (AAA) (proposed)	• Domain: Fairness in computer models • Jurisdiction: United States
Family Educational Rights and Privacy Act (FERPA)	• Domain: Access to educational information • Jurisdiction: United States
FTC Fair Information Practice Principles (FIPPS)	• Domain: E-commerce • Jurisdiction: United States
Biometric Information Privacy Act (BIPA)	• Domain: Data privacy of biometric information • Jurisdiction: Illinois

Note: The Algorithmic Accountability Act (AAA) of 2019 is proposed legislation that seeks to regulate fairness in computer models throughout the U.S. and has been in committee for two years. It is unlikely to be passed without being reintroduced as a new piece of legislation. According to its creator, it will be revised and reintroduced some time in 2021.

Additional Reading

The IAPP maintains a table of privacy laws that are introduced by each US state. You can find this table at **https://iapp.org/resources/article/state-comparison-table/**.

For more information about these regulations, visit the following web pages:

- HIPAA: **https://www.hhs.gov/hipaa/index.html**
- COPPA: **https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule**
- AAA: **https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info**
- FERPA: **https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html**
- FIPPS: **https://ethics.berkeley.edu/privacy/fipps**
- BIPA: **https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57**

ISO Privacy Standards

The most important standard to be aware of is the **ISO/IEC 27701** standard. It provides guidance for implementing a privacy information management system. It builds on the requirements in ISO/IEC 27001, the information security management system standard, and the code of practice for information security controls in ISO/IEC 27002.

Implementing a management system compliant with these standards will enable you to meet the privacy and information security requirements set forth in GDPR and other data protection regulations. It is therefore a good place to start to embed privacy practices in your organization.

Privacy Risk Identification Techniques

Three basic techniques for identifying privacy risks are mapping the presence of private user data in the organization, tracking personal user data from collection to use, and modeling customer personas to identify non-obvious privacy risks.

- **Map the presence of personally identifiable information (PII).** Identify all points in your workflow where PII is connected, and conduct a risk assessment for each piece of information.
- **Track customer data**. Track each piece of personal user information. Watch out particularly for situations where you receive data from third parties, share data with third parties, and cross-correlate data.
- **Model customer personas**. Sometimes, it is not obvious what information users consider most important and want kept private. Identify user groups and engage them in interviews or focus groups on privacy. This can help understand users' goals, concerns, and level of technical expertise. Use the stakeholder-mapping methodology to map the privacy expectations and technical expertise of users.

ACTIVITY 5-2
Identifying Privacy Risks

Scenario
Consider either the RudiBrace example or a product you are working on in your own company.

1. List two privacy risks associated with the product and the risk role that Rudison Technologies (or your organization) has with respect to these risks.

2. Are there any groups with respect to these risks that are in problematic risk roles?

TOPIC C

Privacy Tradeoffs

Now that we've looked at relevant regulations, we'll turn to some practical challenges. As with other ethical risks, you will often have to consider other values as you determine how best to mitigate privacy risks. In this topic, you will describe commonly encountered privacy tradeoffs.

Finding the Balance between Privacy and Other Values

Handling privacy risks is more than a matter of complying with regulations. Sometimes, protecting privacy requires difficult tradeoffs with other values, particularly:

- Security
- Public health
- Convenience
- Efficiency

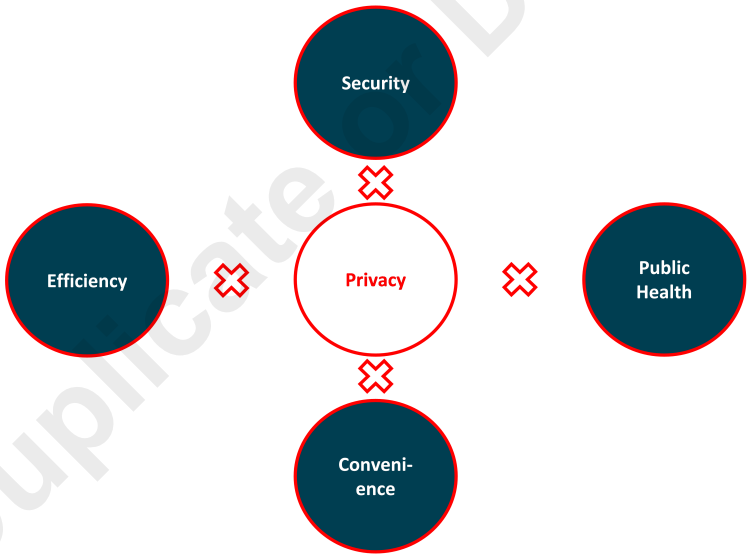


Figure 5-2: Potential privacy tradeoffs.

Increased privacy can also conflict with a sense of personal safety online. For instance, the privacy that comes with online anonymity has significantly increased online bullying. For people to feel safe in online interactions, it is sometimes necessary to access private data about their interactions; for instance, to stop someone from bullying other users.

Effect on Public Interest

When different fundamental interests are at stake—like privacy and public health—we have to make tradeoffs. In these cases, it can be helpful to account for the following three considerations.

- **Context**: When intrusions on privacy in the public interest (for example, in a public health crisis) are warranted, these exceptions should be explicitly treated as exceptions.
- **Justification**: When intrusions on privacy are warranted in the public interest, and privacy is traded off against another good like public health, the justification for this tradeoff must be clear.

For example, the justification for contact tracing is that it will significantly decrease the spread of infection and save a significant number of lives.

- **Minimizing future risk**: The use of data and information collected through exceptional invasions of privacy should be clearly restricted. For example, people's private information collected during the pandemic should be used only for the specific purpose of protecting people from the pandemic.

Tradeoffs with Security

Measures that strengthen data protection to increase privacy can at times conflict with different forms of security. Recall the case where Apple refused to build a backdoor to the iPhone for the FBI, because, in the wrong hands, it had the potential to undermine the security of hundreds of millions of people who used Apple products. The FBI was working to protect people from a terrorist threat. Apple was working to protect people's security in a different sense. But protecting user privacy, in this case, had the flip side of appearing to enable terrorism. Similar concerns are often raised about technologies such as blockchain, which can help users in gaining more privacy, but that are also often used by cybercriminals precisely for this reason.

Now that we've spent some time on privacy, are you thinking any differently about whether or not Apple did the right thing?

Tradeoffs with Public Health

To contain a public health crisis, it may be necessary to track and trace people, which may infringe on their privacy.

Public health crises confronted governments and companies with the challenge of determining when, if ever, it's appropriate to lift data privacy protections. With scientists rushing to develop medicines, rapid access to data anywhere was of the essence. In addition, some effective responses to containing a public health crisis rely on extensive government surveillance tools to track and isolate infected persons and those with whom they had been in contact. Quick access to large amounts of data can be vital for machine-learning forecasters in predicting the trajectory of a crisis. Intrusions on people's privacy can save lives.

But privacy advocates are often concerned about the precedent this might set. In early 2020, the World Economic Forum released a statement, urging companies to maintain proper AI oversight. WEF's head of AI and machine learning, Kay Firth-Butterfield, warned that "We need to keep in mind that the big ethical challenges around privacy, accountability, bias, and transparency of artificial intelligence remain."

Tradeoffs with Convenience and Efficiency

Protecting privacy often comes with measures for data-minimization, which means that service providers collect only the data necessary for the basic functioning of a service. However, data-minimization can affect the potential effectiveness of a service and the convenience with which it is used.

Tailoring services to individuals based on their personal characteristics might make services like targeted advertising significantly more accurate and efficient. But this also uses information about people's tastes and preferences in a way that compromises their privacy and autonomy interests.

Sacrificing some of our privacy might save us time, and give us access to things we wouldn't otherwise be able to access. The majority of people who use the Internet have little understanding of who has access to their data. Even those of us who have some idea typically treat our personal information like a currency - we're willing to give up some of it for convenience. In fact, in November 2019, Pew Research reported that "roughly six-in-ten U.S. adults say they do not think it is possible to go through daily life *without having data collected about them* by companies or the government."

Tradeoffs between privacy and convenience also occur when we use wearables to track our health and fitness. Wearables like Fitbit collect vast amounts of health-related data from their users. For machine learning algorithms that are used for data-analysis, it is often useful to have access to a great range of datasets to produce the best results for the user. Yet, greater access to these datasets can come with privacy risks.

Another context in which tradeoffs with privacy occur is home convenience, and the IoT. For example, Ring and Nest doorbell systems capture video of both your front door and your neighborhood. Ring has been actively collaborating with local law enforcement, which has raised many eyebrows about privacy protection at the local level.

Additional Reading

For more information, visit the following web pages:

- Pew Research: **https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/**
- IoT devices and law enforcement: **https://www.washingtonpost.com/technology/2019/01/31/doorbells-have-eyes-privacy-battle-brewing-over-home-security-cameras/**

ACTIVITY 5-3
Discussing Privacy Tradeoffs

Scenario
Consider the following questions as you discuss privacy tradeoffs.

1. What limits to privacy should we accept for the sake of national security, public health, and the quality and efficiency of services?

2. In what contexts do you think privacy overrides other values?

TOPIC D

Mitigate Privacy Risks

Now that you know more about privacy standards and tradeoffs, you should be better equipped to make informed decisions about mitigating privacy risks in emerging technology projects. In this topic, you will mitigate privacy risks.

Framework for Selecting Mitigation Strategies

There is a dazzling number of tools and techniques for protecting user privacy, from discussions about how to obtain meaningful consent from users to sophisticated encryption tools. Often, there will be multiple ways of addressing a privacy risk. Yet there is a hierarchy between options for improving privacy. Here is a three-step framework to cut through the noise.

- **Minimize data collection and sharing**. This is the foundation of managing privacy risk. The minimization requirement is aimed at avoiding gratuitous privacy risks.
- **Protect data**. For all user data that needs to be collected or shared, organizations should make use of state-of-the-art techniques for protecting data. This due-diligence requirement is aimed at minimizing privacy risks that cannot be avoided in the first place.
- **Opt-in and obtain informed consent**. Remaining privacy risks should be communicated transparently to users to enable them to make an informed choice whether to take a privacy risk. This requirement aims at giving users agency in taking privacy risks.

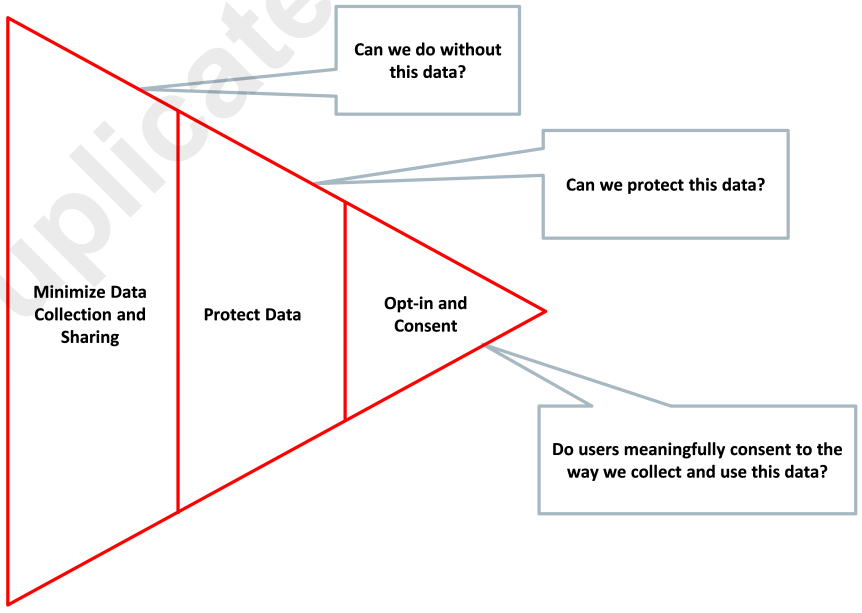


Figure 5-3: Privacy mitigation framework.

Each of these is discussed in detail in the next sections. However, note that following this framework mechanically will not lead to good results. Privacy norms are highly context-specific and evolve over time. You need to think through your specific use case, and the evolving privacy norms for the type of data, given the data subject as well as the sender and recipient. Privacy expectations are violated when the informational norms associated with a specific type of information and social relationship are breached.

To ensure that you follow a thoughtful process when creating a new product, you can follow the ***privacy by design*** approach. Privacy by design calls for privacy to be taken into account throughout the whole engineering process. It is closely related to value sensitive design.

It can also be beneficial to use existing software development kits (SDKs) that have been built with strong privacy protections in mind. Open-source SDKs like HealthKit, ResearchKit, and CareKit allow you to develop apps that track user health and enable medical research while adhering to the demanding standards of user privacy.

Additional Reading

For more information:

- Michael Zimmer: How Contextual Integrity can help us with Research Ethics in Pervasive Data: **https://medium.com/pervade-team/how-contextual-integrity-can-help-us-with-research-ethics-in-pervasive-data-ef633c974cc1**
- ResearchKit:: **http://researchkit.org**
- CareKit: **https://developer.apple.com/carekit/**
- HealthKit: **https://developer.apple.com/documentation/healthkit/about_the_healthkit_framework**

Minimization of Collecting and Sharing Private Data

Not collecting private information in the first place is the most obvious way of reducing privacy risk. For every piece of personal user information, ask yourself if you really need this information to build your product or provide your service. Data sharing between organizations should also be minimized, because data sharing makes it difficult to track how data is used. Just because a user gave consent that their data is shared with some entity, it does not mean that this consent applies to third parties.

Determination of What Private Data Needs to Be Collected

We should not underestimate the complexities in deciding whether to collect or share personal user data. One issue is that what is necessary is not clear-cut. Rather, applying the criteria requires making ethical judgment calls. Consider Facebook. Before Facebook came to the market, online chat rooms and networks allowed people to pick a user name which did not have to be their real name, and usually was not. Facebook deliberately required users to create profiles under their real names. That was a strategic decision to create a new type of social network. Rather than facilitating strangers to encounter each other using pseudonyms, Facebook wanted to make it easy for people who were friends in the real world to find each other. Was it necessary for Facebook to adopt the policy? The issue is not black and white. The ability to find friends by their real names was part of Facebook's value proposition from the start. Yet other social-media platforms function without this requirement.

Moreover, product teams know that features of products often change over time. It is often only through user requests after a product has been launched that organizations discover the functionalities that users value most. Even very mature products using emerging technologies are under constant development. The expectation to pivot makes it tempting to collect more data than strictly necessary for current purposes to keep options open.

Finally, there is sometimes a tradeoff between minimizing the collection of personal data and fair treatment. For instance, a bank might take the necessity requirement too far by making credit decisions based on the socio-economic profile of people living in your postcode, rather than collecting detailed information about your individual financial situation.

Protection of User Data

For all user data that needs to be collected or shared, organizations should make use of state-of-the-art techniques for protecting data. The specific techniques change quickly. Here are three types of techniques to protect user data, along with their associated advantages and disadvantages:

- **Anonymization**: Remove data that may identify a subject.
- **Encryption**: Protect information from access by unauthorized people.
- **Zero-knowledge protocols**: Share relevant information without revealing user data.

Each of these is discussed in detail in the next sections.

Anonymization

Anonymization permanently removes all data that might identify a subject. Anonymization is a form of ***de-identification***: a process to prevent someone's personal identity from being revealed. For instance, you might delete name, email-address, and other columns that hold identifiable information from an employee database. A close cousin of anonymization is ***pseudo-anonymization***, which disguises all data that might identify a subject. For instance, in an employee database you might replace names with numbers and keep a record of the mapping of names to records elsewhere. Another closely related concept is the use of synthetic data. ***Synthetic data*** is data generated by a computer simulation, approximating personally identifiable data, but fully algorithmically generated.

The advantage of anonymization is that it is less likely that data can be traced back to an individual if exposed, reducing the risk of violating a user's privacy. Note, however, that combining information from apparently innocuous columns such as business unit, gender, and age may be sufficient to single out individuals. Moreover, anonymized data may reveal identifiable information if it is linked to another database. For instance, the Ministry of Road Transport in India sold information about vehicle owners to private companies. Through registration numbers available in the dataset, the database can be linked to another database containing driving license records and insurance information.

Additional Reading

For more information about the risks of linking separate databases, visit: **https://www.outlookindia.com/website/story/india-news-legal-or-not-why-has-the-roads-ministry-sold-our-data/334278**.

Encryption

Encryption protects information from access by unauthorized people. Data can be encrypted both at rest and in transit.

- Consider the encryption of data at rest. Regular encryption uses algorithms and a key to encode a message. To decrypt the resulting cyphertext, translating the data back into its original form with the same algorithm and key is required.
- For encryption in transit, VPNs can be used to hide IP addresses and other identifiable information in all network traffic, and to encrypt data; proxies can be used to do the same with web traffic.

Homomorphic encryption is a technique to analyze and manipulate data while it is still encrypted. Only the results are decrypted. It enables you to work on data without sharing it in its unencrypted form.

Encryption of data in transit and at rest is now a standard requirement in privacy regulation and standards. The latest encryption techniques should be used wherever available to protect data from being exposed to unauthorized people. These techniques can reduce the risk of data being exposed inadvertently. However, homomorphic encryption aside, data that cannot be decrypted in any way is useless. Each point at which data is processed in an unencrypted form is an attack vector.

Zero-Knowledge Protocols

Zero-knowledge protocols allow you to prove that you have a piece of information without revealing the information itself. They are techniques to implement ***differential privacy***, as they allow the public sharing of information about a dataset by describing patterns within the dataset while withholding information about individuals in the dataset. These protocols can be used to communicate *about* personal data without revealing the actual data. For instance, rather than sharing bank account statements, zero-knowledge protocols enable users to prove they have a certain level of income without revealing any further details about their income.

Zero-knowledge protocols are useful to transfer relevant information to third parties while minimizing the exposure of private user data. They are a useful addition in the privacy toolbox. However, their application is limited to cases where third parties require less sensitive information about a user than the organization would need to share as proof.

Opt-In and Consent

Informed consent is consent given based upon a clear appreciation and understanding of the facts, implications, and consequences of an action. Only once the previous strategies are exhausted should organizations rely on eliciting informed consent from users about what data is collected about them and how this data is used. But users typically lack a sophisticated understanding about what data an organization may collect about them, and how that data is used. In fact, research shows users lack the knowledge to make informed decisions about privacy options.

This raises a big problem for privacy strategies relying on informed consent: There is insufficient comprehension and willingness in the consent process for users to give informed consent for the collection and management of their personal information. Therefore, obtaining informed consent cannot replace the previous strategies, but should be used only to address remaining privacy risks once the previous strategies are exhausted.

As a result, organizations should not go too far beyond ensuring that users have accepted a data sharing agreement. To make consent more meaningful, the policy you propose to users should:

- Live up to high ethical standards like minimizing data collection and data protection.
- Ask for consent in a way that gives users agency about what happens with their data.
- Connect to the framework for an ethics risk assessment, in that organizations should be wary of taking risks where they are the beneficiaries and decision makers, but not exposed to the risk, whereas users bear the risk but have no decision-making powers.

Establishing meaningful consent and giving users options aims at transforming the risk role of users to make them co-decision-makers about what happens with their data.

Additional Reading

For more information about informed consent, visit **https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/pra2.2015.145052010043**.

Strategies for Obtaining Informed Consent

Organizations need to be explicit about what data they collect, for which purpose, and how the data is collected. Organizations should also provide users with an option to opt in or opt out. Setting up an end user license agreement (EULA), a service level agreement (SLA), and crafting a clickthrough agreement may be sufficient from a legal perspective. But users are unlikely to read long-winded terms of service agreements. Rather, here are some strategies you can use to make consent more meaningful:

- Give short explanations of how data is collected and used in plain language.
- Verify that users understand what you are saying, perhaps using a short knowledge check.
- Point out the privacy risks you identified.

- Set defaults that minimize data collection, even at the expense of product quality, and allow users to select more permissive settings to unlock additional functionality.
- Renew consent agreements regularly and give users the possibility to dial back data collection and sharing on an ongoing basis.

These strategies aim at giving users a meaningful choice as to whether to use your service. The more granular the choice that users can make, the more meaningful their consent, especially to using a product that is difficult to do without. For instance, rather than giving users only the choice to leave or to give advertisers access to their data, consider offering a paid version of your product that does not share data with advertisers. Give users the ability to make granular choices, for instance by selecting not only whether to receive newsletters from you, but also which newsletters specifically, and how frequently.

ACTIVITY 5-4
Discussing Privacy Risk Mitigation

Scenario
Consider either the RudiBrace example or a product you are working on in your own company.

1. Consider two privacy risks associated with the product and ways of mitigating these risks.

2. Which tradeoffs do these mitigation strategies create?

Summary

In this lesson, you identified and mitigated privacy risks. You discussed how privacy affects power structures and emerging technologies, and you investigated how privacy is affected by cultural differences. You also looked at sources for privacy risks and the regulations and standards that have been developed to protect against these risks. After considering some of the tradeoffs you might need to make between privacy and other values, you investigated common risk-mitigation techniques and the idea of informed consent. All of this information will be useful to you as an ethical emerging technologist to ensure that you can protect user data from being used in a way other than it was originally intended.

Have you or your company been involved in a privacy breach? If so, what steps were taken to mitigate the risks?

Which privacy tradeoffs do you think are most important, and why?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

6 Identifying and Mitigating Fairness and Bias Risks

Lesson Time: 1 hour, 30 minutes

Lesson Introduction

Fairness, most people think, means treating people as equals. Yet, what it means to treat people "fairly" or "as equals" is controversial and varies based on context. Data-driven technologies present special risks of bias—of treating people unfairly. Data-driven technologies can reproduce or amplify inequalities in society, as well as introduce new forms of discrimination. Not all forms of bias and unequal treatment are unfair, however. In fact, sometimes treating people unequally is essential to treating them fairly. To identify and mitigate risks of unfairness, we must first understand these complexities. In this lesson, you will define fairness and bias, and identify bias risks and methods to mitigate them.

Lesson Objectives

In this lesson, you will:

- Describe fairness and its application to emerging technologies.
- Identify common sources of bias risks and select appropriate tools for identifying bias risks.
- Describe common fairness tradeoffs.
- Select appropriate tools for mitigating bias risks.

TOPIC A

What Are Fairness and Bias?

Prior to identifying and mitigating bias risks, it is necessary to clearly define the ideas of fairness and unfairness, or bias. In this topic, you will identify the basics of fairness and bias.

Fairness and Bias

Issues of fairness, bias, and discrimination represent some of the most common and disturbing ethical problems with emerging technologies. In this course, we have already encountered several worrisome examples, from predictive policing, to credit scoring, to facial recognition.

Fairness refers to the idea of giving each person his or her due and ensuring that any departures from equality can be justified. Fairness, sometimes called equity or justice, is a natural implication of the moral equality of people that underlies human rights. Recognizing one another as equal moral people requires that we continuously uphold this status by guarding against forms of bias and discrimination in our practices and conduct.

Treating people fairly does not always mean treating people the same. Sometimes recognizing differences in characteristics, privileges, or needs is essential for treating people fairly.

Unfairness is sometimes called ***bias***, discrimination, or injustice; more specific forms of unfairness, related to particular characteristics or groups, may be called sexism, racism, ableism, ageism, and so on. Unfairness has many sources. Often, unfairness comes from prejudiced or hateful attitudes that some people hold toward others. Many times, however, unfairness is not intentional. It comes from inattention to sources of bias or the way our behavior can reproduce or magnify the effects of unfair historical conditions.

Differential Treatment

Not all forms of differential treatment are unfair. For instance, targeting a product to markets where expected demand is high is not necessarily unfair to people who fall outside the target market. And when distributing opportunities of other kinds, it may not be unfair to deny certain benefits to individuals who are already privileged in order to assist individuals who are underprivileged. In many societies, law and social conventions treat certain characteristics—such as gender, religion, or sexual orientation—as "protected," meaning that discrimination on the basis of these characteristics is unacceptable or illegal. These traits are often unchosen and associated with particular disadvantages. Traits that are chosen or associated with advantages are less often recognized as worthy of protection from discrimination.

Bias in Data-Driven Technologies

Technologies that depend on data are especially prone to bias. In statistics, a measurement is biased if it differs systematically from the parameter that it is supposed to estimate (***statistical bias***). Not every instance of statistical bias is discriminatory. For instance, a team may systematically underestimate the true number of trees in a city block by using old data. This statistic is biased, but it is not discriminatory.

Yet many sources of bias in data are related to protected human characteristics. And even biased data that is not related to protected characteristics can lead to discriminatory effects. For instance, if the team's results will be used to inform environmental policy in the city, this biased data may result in policies that are biased against certain city residents.

Data is an approximation of the real world. The real world is characterized by unfairness. Using data to make decisions, therefore, risks reproducing and magnifying various forms of unfairness, as this figure illustrates.

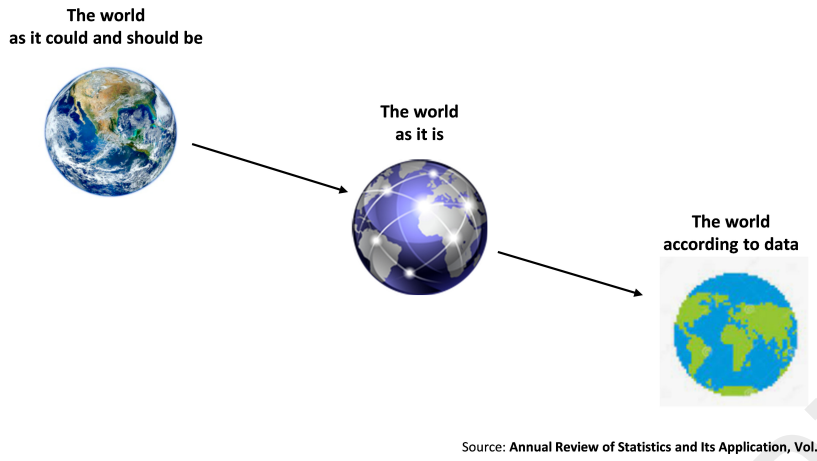


Figure 6-1: Layers of bias in modeling the world with data.

Example of Biased Sampling

Consider the example of CoNLL-2003, an open-source dataset commonly used to build natural language processing (NLP) systems. CoNLL-2003 was developed in 2003 based on 20,000 sentences from newspaper articles. A recent analysis has found, however, that the database contains five times more male names than female names and almost no names that are gender-neutral. Models trained on this dataset are therefore likely to be much better at identifying male names than female or gender-neutral names. Since women and non-binary individuals represent historically marginalized groups, this biased dataset can be expected to exacerbate unfair conditions in society.

Note: For more information about the CoNLL-2003 analysis, visit **https://scale.com/blog/if-youre-de-biasing-the-model-its-too-late**.

Types of Bias

Bias comes in many forms. Other than statistical bias, there is also cognitive bias. A ***cognitive bias*** is a systematic error in processing and interpreting information. We often use cognitive shortcuts (or heuristics) to make decisions more quickly. But these shortcuts can also lead to errors of judgment. These errors are not necessarily related to unfairness, but just like statistical bias, they can lead to forms of unfairness. The following table lists some common statistical and cognitive biases.

Bias Type	*Description*	*Example*
Implicit bias	Stereotypes we may unconsciously hold about certain people or groups.	Instinctively clutching one's wallet when passing a member of another race on the street.
Automation/complacency bias	The tendency to trust decisions produced by automated systems more than human decisions.	Trusting an online symptom checker more than the diagnosis of a doctor.

Bias Type	Description	Example
Sampling bias	Sample data that does not reliably represent the target population, often due to problems in data collection or mistakes in correcting collected data.	Estimating national population characteristics on the basis of data from a single postal code.
Confirmation/reinforcement bias	Discounting information that challenges our prior beliefs or values.	Giving greater weight to scientific studies that support your pre-existing opinion on a topic.
Cultural bias	Assuming aspects of one's own cultural experience are universal.	Academic assessments that test skills or use examples that are more familiar to certain ethnic groups.
Gender bias	Treating the particular interests of one gender as the norm.	Failing to appreciate how pregnancy or care-giving may impose particular burdens on women.
Ableism bias	Failing to account for the interests of disabled persons.	Designing a vehicle based on data from people with full use of their arms and legs.
Temporal bias	Believing or valuing something more because of its location in time.	Being more confident about an exam when it is scheduled in the distant future.
Sociological bias	Stereotyping individuals who are less like oneself.	Holding negative beliefs about employees of a rival company.

Fairness Principles

Many different principles of fairness might be applied to different contexts. People often disagree about how these principles should be defined and which one to apply in a particular situation. Some of the primary principles of fairness used in statistics and computer science are shown in this table.

Fairness Principle	Definition	Example
Group unawareness	Outputs take no account of individuals' membership in protected groups.	A credit scoring model omits race and other parameters that might be proxies for race.
Statistical parity	Rates of success are equal for members of protected and unprotected groups.	Download rates of a dating app are the same for LGBTQ+ and non-LGBTQ+ individuals.
Demographic parity	Rates of success are proportional to a group's prevalence in the broader population.	Company boards are staffed with 50% women.
Formal equality of opportunity	Individuals with the same qualifications have equal chances of success.	Individuals who pass the CEET exam have the same rates of success in job applications.

Fairness Principle	Definition	Example
Equal accuracy	Protected and unprotected groups have the same rate of correct and incorrect classification.	The percentage of mistaken predictions about recidivism is the same for racial minorities and the dominant group.
Group awareness	Success criteria are sensitive to differences among different groups.	Although high schools in a particular state vary widely in test scores and demographic composition, the top students at each school receive scholarships to the state university.

Unfortunately, it is often difficult or impossible for a model to satisfy more than one of these principles simultaneously. Disagreements regarding which principle to use may depend on different underlying assumptions about fairness and bias. Questioning and discussing these assumptions openly can help resolve or at least clarify these disagreements.

ACTIVITY 6-1
Discussing Bias

Scenario
Consider the following as you discuss the basics of fairness and bias. Remember, you discussed the following types of bias:

- Implicit bias
- Automation/complacency bias
- Sampling bias
- Confirmation/reinforcement bias
- Cultural bias
- Gender bias
- Ableism bias
- Temporal bias
- Sociological bias

1. Can you brainstorm one additional example of bias from at least five of the types of bias discussed in this topic?

2. Which of these forms of bias are most relevant in the context of emerging technologies? Why?

TOPIC B

Identify Bias Risks

Before you can act to mitigate bias risks, you need to know what risks you are up against. In this topic, you will identify bias risks and their sources, as well as some tools that can help you identify those risks.

Sources of Bias Risks

As we have seen, bias comes in many varieties. In addition, there are certain conditions, environments, and practices that can increase the risk of bias.

- **Data collection and sampling methods**. Data collection and sampling methods commonly reflect biases. The data may contain missing or erroneous entries. Or, the data may be accurate but sampled from sources that do not represent the target population. The way that data is categorized and labeled may represent biased decisions by humans or AI. Attempts to correct or clean data may involve mistakes or introduce new sources of bias.
- **Overfitting to training data**. Another common risk of bias arises when a model is ***overfit*** to the training data. A model is overfit when it performs well on training data but does not perform equally accurately to new data. Overfitting is often the result of using a model that has too many parameters, a data set that is too small, or a data set that contains too much "noise" (that is, data points that are not relevant to the model).
- **Edge cases**. Edge cases include outliers (data points that are outside the normal range), missing/erroneous data points, and noise. When edge cases are included in a model, they may skew the results in a particular direction.
- **Use/integration of third-party products**. When you are using third-party products that rely on data, the qualities of this data are often unknown. The data or methods may be biased, or they may introduce biases when integrated with other products.

Methods for Identifying Bias Risks

Methods for identifying potential biases include:

- **Analytical techniques** can be used to systematically assess data for appropriate representativeness and document its origins and characteristics. This can involve analyzing the summary statistics, examining edge cases, and noting the methods of collection and any potential problems.
- **Disparate Impact Modeling** involves comparing the results of the model to people who differ just in protected attributes such as gender or religion. Does the model perform well on all groups? Does the model treat some groups differently?
- A model's behavior can also be analyzed in **different environments** to identify biases or potential discrimination. To this end, a model may be tested on different data sets, for different purposes, or with different methods.

Tools for Identifying Bias Risks

Besides the general methods discussed above, there are also numerous tools that can help identify potential biases.

- **Bias bounty** refers to the idea of paying people a reward for identifying significant forms of bias in a dataset or model.
- **Toolkits to identify discrimination and bias** have been developed and open-sourced by numerous organizations. IBM's AI Fairness 360 toolkit, Google's what-if tool, and

ResponsibleAI's responsibly tool are online tools that can analyze common sources of bias in code that is uploaded to them.

- ***Radioactive data tracing*** is a process for identifying whether or not a particular dataset was used to train an ML model. It relies on a technique based on medical diagnostics that involves marking a material, introducing it to the body, and following its course to identify potential pathologies. Data can be marked before it is fed into a model and then followed to determine whether the model is using the data appropriately.

Additional Reading

For more information, visit:

- AI Fairness 360 toolkit: **http://aif360.mybluemix.net**.
- Google what-if tool: **https://pair-code.github.io/what-if-tool/**.
- ResponsibleAI responsibly tool: **https://github.com/ResponsiblyAI/responsibly**.

Fairness and Bias Standards and Regulations

Regulations and standards on fairness and bias are still evolving. Current regulations relevant to these issues include:

- The General Data Protection Regulation (GDPR), which provides recourse for victims of algorithmic discrimination.
- The UK Equality Act prohibits discrimination on the basis of age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation.
- The Civil Rights Act of 1964 prohibits discrimination in the US on the basis of race, sex, gender, religion, national origin, and sexual orientation.
- The IEEE 7000 series includes best practices for identifying and mitigating bias in algorithms.

Additional Reading

For more information about the IEEE 7003 project, which details considerations regarding algorithmic bias, visit **https://standards.ieee.org/project/7003.html?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined**.

ACTIVITY 6-2
Identifying Potential Sources of Bias

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

What are some potential sources of bias in the design and development of this product?

TOPIC C

Fairness Tradeoffs

As with other ethical risks, you will often have to consider other ethical values as you determine how best to mitigate bias risks. In this topic, you will describe commonly encountered fairness tradeoffs.

Common Fairness Tradeoffs

Most people agree on the importance of fairness in emerging technologies. Yet, measures to promote fairness may sometimes conflict with other ethical values such as accuracy, liberty, and utility.

Often, fairness tradeoffs have to do with people's different understandings of fairness itself. For instance, the principle of equality of treatment can, in certain circumstances, conflict with the principle of equality of outcomes.

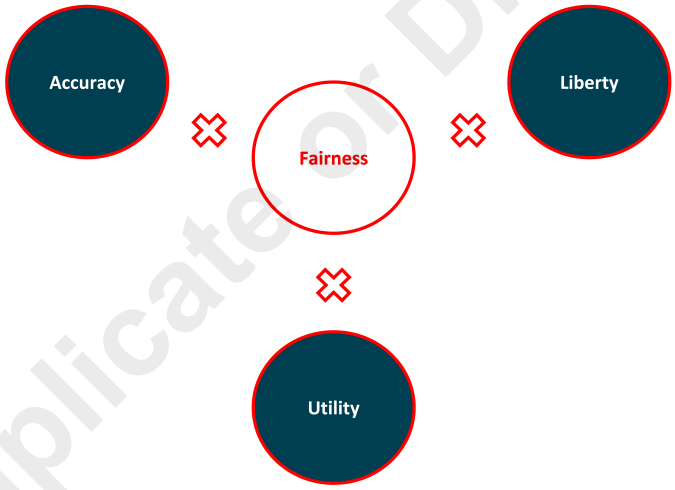


Figure 6-2: Common fairness tradeoffs.

Accuracy Tradeoffs

Accuracy is an ***evaluation metric*** for the quality of a model. Evaluation metrics are metrics that capture an important quality criterion for a model. A subset of these evaluation metrics focuses on the **reliability** of a model. For instance, **accuracy** is the number of correctly predicted data points out of all the data points. Other reliability metrics include **specificity**, which is the proportion of negative cases that were classified as negative. In the context of AI, measures to increase fairness (by reducing bias) can reduce the accuracy of models. A reduction in accuracy, in turn, can lead to negative side effects, such as a reduction of inclusivity.

Consider the case of a product team working on an AI model for mortgage lending. The model produces risk profiles of potential borrowers and assigns mortgages to recipients who pose an acceptable risk. The more accurate the model is, the more people who should be considered as eligible for a loan are considered as such. Hence, greater accuracy means a greater inclusion of eligible borrowers.

However, the AI model uses training data that is biased towards certain minorities, who are more likely to be declined a mortgage application than others. The product team decides to use another

model and dataset that are less biased. Yet, the new model turns out to also be less accurate, turning down more eligible borrowers in total and therefore being less inclusive on the whole.

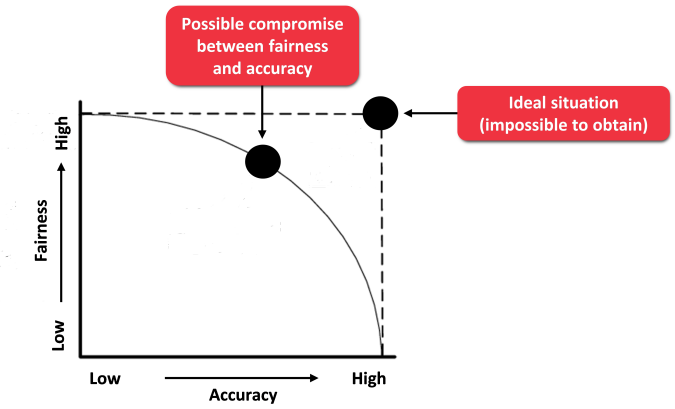


Figure 6-3: Accuracy tradeoffs.

Liberty Tradeoffs

Measures to increase fairness can conflict with some people's understanding of liberty. Liberty is often understood as non-interference, of providing people with the same set of rules that have the fewest possible restrictions on people's choices. Yet, to increase fairness, it is sometimes necessary to interfere; for instance, by redistributing resources from persons who benefit from unjust systems to the victims of injustice.

Consider a product team building an AI model to distribute loans to people based on data about their financial situation. The question that product team members ask themselves is whether they should lower the criteria for getting a loan for historically disadvantaged groups.

From one perspective, the answer might be yes. Since no one chose to be born a member of a historically disadvantaged group, those unfairly disadvantaged should be compensated.

From another perspective, however, giving a boost to disadvantaged people means limiting the liberty of others. Depending on how we define and rank liberty in comparison to fairness, we may find this tradeoff unacceptable.

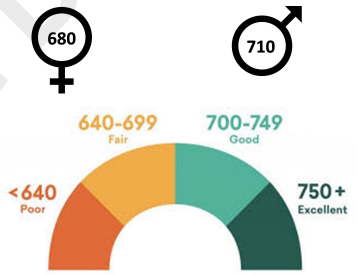


Figure 6-4: Liberty tradeoffs.

Utility Tradeoffs

In some cases, considerations of fairness can conflict with those of overall utility. If this is the case, overall utility can be increased, but only at the expense of discriminating against certain individuals or groups.

Consider a product team working on an application that aims to most effectively allocate intensive care (IC) beds to a group of patients. Ideally, the team members do not want to discriminate between potential patients because healthcare is an important human right, and everyone should have equal access to it.

However, it turns out that allocating available beds to older patients results in more deaths and fewer "quality-adjusted life years." Therefore, to achieve a higher overall utility of the allocation of patients to IC beds, it would make sense to discriminate—to allocate more IC beds to younger patients.

One could see this example as a tradeoff between fairness and utility. But one could also view it as a tradeoff between different ideas of fairness: fairness as non-discrimination and fairness as equal consideration of interests (quality-adjusted life years).

ACTIVITY 6-3
Discussing Fairness Tradeoffs

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

The RudiBrace team members discuss what would be most fair: to boost the activity score of employees with a care duty due to their family situation or to generate the same activity scores for all employees.

1. What argument would you make if you were in favor of boosting scores, based on different notions of fairness?

2. What argument would you make if you were against boosting scores, based on different notions of fairness?

TOPIC D

Mitigate Bias Risks

Now that you know more about bias risks and tradeoffs, you should be better equipped to make informed decisions about mitigating bias risks in emerging technology projects. In this topic, you will mitigate bias risks.

Bias Risk Mitigation Strategies

It is hard, if not impossible, to completely eliminate bias in the context of data-driven technologies. Bias is not only hard-wired in technologies, but also in society itself, and to some extent societal biases will always be reflected in the technologies we develop.

Nevertheless, there are good strategies to mitigate bias risks. Some of these strategies are related to product team management.

- Many biases can be spotted and resolved by having inclusive and diverse product teams.
- Bias risks can also be mitigated through timely engagement with relevant stakeholders.
- There are technical strategies to reduce bias risks, which have to do with AI models and training data.

Product Team Diversity

One way of mitigating bias risks is by ensuring diversity in the composition of a product team. Recall the case of Facebook Portal, in which bias in the product's AI model was detected thanks to a product team member who was personally affected by it. There is no one way to ensure team diversity, but the following guidelines help:

- **Address inherent and acquired diversity**. It is important to try to include people with different backgrounds in a product team—considering diversity in terms of gender, race, age, and socioeconomic background. At the same time, acquired diversity or diversity through experience should also be considered; for instance, by including people who have worked in diverse environments.
- **Think of diversity during team composition**. Diversity can be one of the leading criteria in recruitment processes. At the top of professional competencies, the diverse composition of the team as a whole should be a leading aim.
- **Provide speak-up opportunities**. It is not enough to have a diverse product team. Additionally, room should be created for people to speak up when they identify certain bias risks. It is important therefore to create a team atmosphere in which people feel their views and opinions are valued and taken seriously.

Inclusive Design

Inclusive design is a design process that aims at including as many potential users as possible. It often focuses on groups that are at risk of being excluded or have an unfair disadvantage in using a product.

Generally, it relies on three principles:

- Recognizing exclusion: setting goals in the design process while keeping in mind that certain groups might be excluded by a particular configuration of a technology.
- Solve for one, extend to many: take potential users who are less abled or who are experiencing unfair disadvantages as a starting point for design.
- Learn from diversity: setting design goals while taking diverse perspectives into account.

In developing data-driven applications, the following guidelines for inclusive design help address bias risks:

- Conduct user studies with diverse user groups. Give these users a say in addressing bias risks.
- Test different models to understand how they perform in different contexts. For instance, different facial recognition models can be tested to assess how well they detect faces of different skin tones.
- Validate systems once they are ready to be shipped. This can be done by engaging with external stakeholders, including groups that represent typically disadvantaged users.

Additional Reading

For more information about inclusive design, visit:

- Microsoft inclusive design principles: **https://www.microsoft.com/design/inclusive/**.
- Building inclusive AI at Facebook: **https://tech.fb.com/building-inclusive-ai-at-facebook/**.

Foreseeability of Fairness Impacts: PESTL and STEEPV

As we have seen, fairness risks span a great variety of impacts. For instance, unfairness can be an issue in politics, in economics—concerning the distribution of resources—in law, and so forth. When products and services have a broad range of possible impacts, product teams can use frameworks that capture these impacts in a comprehensive manner. Two such frameworks are known as ***PESTL*** (political, economic, social, technological, and legal) and ***STEEPV*** (social, technological, economic, environmental, political, values), which contain major categories of potential impacts.

Taken together, these frameworks assist product teams in foreseeing the following potential impacts that have implications on fairness:

- **Political** impacts, e.g., on political campaigns.
- **Economic** impacts, e.g., on access to resources like housing and food.
- **Environmental** impacts, e.g., on the emission of greenhouse gases.
- **Technological** impacts, e.g., on international technical standards.
- **Legal** impacts, e.g., on law enforcement efforts.
- **Value** impacts, e.g., on religious values.

Fairness in AI

Bias risks are also mitigated at the technical level. As we saw when discussing the fairness tradeoffs, fairness in AI is a complex and multifaceted problem that does not have a straightforward solution. Algorithmic pattern matching does not equate with human bias or prejudice, but at the same time algorithmic processes can produce outcomes that can reinforce biases. Paying attention to potential bias risks at different stages of product development can minimize bias risks.

The following aspects help mitigating bias risks in developing AI models:

- **Before modeling**: Ensure proportionally equal representation in input data. For instance, this means including representative datasets of skin tones and gender characteristics in facial recognition input data.
- **During modeling**: Verify that the model performs equally well for different groups. For instance, check whether a facial recognition model is equally well equipped to recognize male and female faces.
- **After modeling**: Assess model output based on whether it has an equal probability of false positives or false negatives across groups. For instance, check whether there is an equal proportion of white faces that are not recognized by the model as there are non-white faces.

Establishment of AI Fairness Baseline

To avoid re-inventing the wheel with each new product, product teams can implement a standard for AI fairness across products. To do so, the following technical elements can be optimized:

- Establish a set of techniques for ***statistical calibration*** that help in resampling or reweighing data.
- Research state-of-the-art in bias-reducing ML models and adopt them.
- Set and monitor thresholds for fair outcomes for typically disadvantaged groups.

ACTIVITY 6-4
Mitigating Bias Risks

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. The product team wants to apply inclusive design principles to the RudiBrace product design.

   What groups might be inadvertently excluded from the RudiBrace design?

2. **How could these groups still be included in the design?**

Summary

In this lesson, you identified and mitigated bias risks. You discussed the basics of fairness, bias, and differential treatment. You also looked at sources for bias risks and the regulations and standards that have been developed to protect against these risks. After considering some of the tradeoffs you might need to make between fairness and other values, you investigated common risk-mitigation techniques and the ideas of diversity and inclusivity. All of this information will be useful to you as an ethical emerging technologist to ensure that you can protect user data from being used in a way other than it was originally intended.

Which types of bias risk do you think your organization faces most often, and why?

Which of the bias mitigation strategies do you think your organization is most likely to adopt, and why?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

7 Identifying and Mitigating Transparency and Explainability Risks

Lesson Time: 1 hour, 30 minutes

Lesson Introduction

Emerging technology has the potential to solve big societal problems in new ways, but the solutions it provides are often more opaque and more difficult to understand than the old ones. In this lesson, we will explore to what extent transparency and explainability of new technologies is ethically valuable, or even required. We will dive deep into the problem of black-box systems, such as many ML algorithms, and the ethical risks they pose. We will explore strategies for making black-box systems explainable. We will see that only a part of this task is technical in nature, and that it is a mistake to jump to technical solutions right away. The foundation of improving transparency and explainability consists in understanding the different needs stakeholders have for transparency and explanation, and then finding a solution that meets these needs.

Lesson Objectives

In this lesson, you will:

- Describe transparency and explainability, including their application to emerging technologies.
- Identify common sources of transparency and explainability risks, and select appropriate tools for identifying transparency and explainability risks.
- Describe common transparency and explainability tradeoffs.
- Select appropriate tools for mitigating transparency and explainability risks.

TOPIC A

What Are Transparency and Explainability?

Before we can address ways to identify and mitigate transparency and explainability risks, we must define what transparency and explainability are. In this topic, you will explore the basics of transparency and explainability, including their application to emerging technologies.

What Is Transparency?

Transparency is a way of operating that makes it easy for others to see what actions are performed. As a metaphor, transparency indicates two things: *seeing* something clearly and *understanding* something. But one does not always go with the other: we might clearly see something, but not understand it. Transparency is more than just providing information.

In relationships between people, transparency is about communication. As such, it lays down a double requirement. The sender has to communicate things clearly and understandably, and the receiver should be able to access and understand the message.

As an ethical principle, transparency has much in common with honesty: having a truthful attitude. A transparent conversation is an honest conversation, in which all relevant information is clearly articulated. Similarly, a transparent government is an honest government, which explains relevant decisions to its citizens and allows for regular and visible processes for contestation. Transparency therefore has to do with trust: by communicating in a transparent manner, a person or institution can be deemed trustworthy.

What Is Explainability?

A system is ***explainable*** to someone if the person is in a position to understand which inputs led the system to produce certain outputs of interest. This idea is also sometimes called ***interpretability***. The two terms are sometimes distinguished, but we can use them interchangeably for our purposes. In fact, the ideas of explainability or explicability, interpretability, and auditability are closely related in the context of transparency.

If a system is explainable, this means that it's possible to explain its decisions in a way that is satisfying to people wishing to understand it. For instance, a citizen might ask a municipality worker why a building permit was declined. The answer will in that case involve something like "because the proposed building is taller than the threshold specified by the building regulation, and so would have violated that regulation."

Providing such an explanation enables understanding in communication: The citizen now understands why the permit was declined and what they would need to do for it to be approved. Note that the explainability of a system is always relative to a certain person with particular need. A lawyer may not be satisfied with the explanation given to the citizen for lack of concrete references to the law in question, whereas couching the explanation in legalese may have made it incomprehensible to the citizen.

Not all decisions must be explainable. For example, your car navigation system's decision to take route A instead of route B to get you from home to work usually needs no explanation. Such a decision is not controversial and will not usually need to be explained. Explainability is usually an issue only when decisions affect people in certain ways; for example, when they might cause harm, or when people have a special interest in learning whether or how their interests were taken into account.

Additional Reading

For more information about interpretable and explainable ML, visit **https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48**.

Why Should We Care About Transparency and Explainability?

Transparency enables people and institutions to justify their actions. Justifications are important because they provide a basis for legitimacy. Justifications explain the reasons for acting in a certain way. Moreover, transparency enables people to change things. Knowing what went wrong for what reason enables us to intervene: We can hold people accountable, request changes, and ask for redress. Transparency is especially important when decisions are made that concern the public (e.g., decisions on taxation), when decisions might cause physical or mental harm (e.g., decisions on healthcare), or otherwise affect people's rights.

A typical case of a lack of transparency and explainability is the "computer says no" scenario. A government employee has to provide you with some crucial information about a decision about your social rights, but the only answer you get is that the system says "no," without further explanation.

A lack of explainability in data-driven systems can cause a variety of problems:

- **Algorithmic aversion**: Users don't trust the system enough because of prior decisions that negatively affected them. This can lead to a lack of trust in companies and government institutions.
- **Automation bias**: On the flip side, users might ignore errors of judgment because they trust data-driven decisions automatically.
- **Omission errors**: Errors in decision-making might be missed by human operators because systems do not flag them.
- **Commission errors**: Humans act on erroneous recommendations by the system, failing to incorporate contradictory or external information.
- **Lack of recourse**: When decisions lack explanation, it is more difficult for those affected to advocate against a particular decision.

Transparency in Emerging Technologies

Transparency has become an increasingly salient issue, largely because emerging technologies continue to render certain aspects of society increasingly complex. In the past, transparency was not an issue: Whoever was in power made a decision, and people had to live with the consequences. Today, we live in complex societies with many different institutions and bureaucracies, which can make it difficult to trace who made a certain decision and why. These questions, however, are important in the face of democratic demand for legitimation and accountability.

Data-driven technologies add to this problem because they sometimes make decisions in a different way than humans do. They add to the complexity issue in three ways:

- ***Black box problem***: So-called "Good Old Fashioned AI" (GOFAI) uses decision-trees to come from a certain set of inputs to an output. In principle, this approach can be transparent: A human can trace an outcome back to the inputs, following the logical steps taken. In contrast, modern ML models use hidden layers of information to make decisions, which makes it hard—or impossible—to trace an outcome back to its inputs. This generates the black box problem: The decision-making process between inputs and outputs is not visible or understandable to human beings.
- ***Expert rule***: Data-driven systems become increasingly complex, and their design depends on insights from very specialized scientific disciplines. Because of this, scrutinizing the decisions made by these systems requires expert knowledge. Even if a lay person could see everything that goes on inside a system, they would usually not be able to understand it.
- ***Commercial secrets***: Many data-driven systems are developed by private companies. These companies often have an interest in keeping information secret or confidential. For instance, they

might choose not to disclose information about proprietary algorithms. They might also integrate third-party products into their systems, meaning that these third-party products remain inaccessible for inspection. Finally, companies might have an interest in taking decisions about user experience without informing users, like in the case of shadow banning—partially blocking a user's content without their knowledge.

ACTIVITY 7-1
Discussing Transparency and Explainability

Scenario
Consider the following question as you discuss the basics of transparency and explainability.

1. List some examples of data-driven systems that you use every day, which produce decisions that lack explainability.

2. Discuss whether and why the lack of explainability poses a problem.

TOPIC B

Identify Transparency and Explainability Risks

Before you can act to mitigate transparency and explainability risks, you need to know what risks you are up against. In this topic, you will identify transparency and explainability risks and their sources, as well as applicable regulations and standards. These risks can occur at any stage in the machine learning pipeline, from collecting data, via training the model, to producing operational outputs.

Sources of Transparency Risks

A system is a black box for a certain person if the person does not understand and is not in a position to understand the relationship between inputs and outputs of the system. The ways such systems process information is therefore not transparent to that person, and that person cannot explain the outputs based on the inputs.

There are four major reasons why systems become a black box to a person.

- The first two are technical in nature:
  - Artificial neural networks.
  - Self-learning models.
- The other two are organizational and legal:
  - The integration of third-party algorithms.
  - Intellectual property rights.

Curiously, building explainable AI creates risks of its own, because it may lead people to put too much trust in the system.

Artificial Neural Networks

Artificial neural networks are ML architectures that are modeled on the way animal brains work. They consist of artificial neurons and connections between these neurons, which resemble synapses in biological brains. The strength of the connections between neurons adjusts as the algorithm learns.

Learning proceeds by considering example inputs and outputs, iteratively tweaking the weights of the connections between neurons until the network creates the desired outputs for each of the corresponding inputs.

The result is a complex web of connections between artificial neurons, characterized by the network's architecture and the weights of the connections between the neurons. While artificial neural networks are very potent in solving a large range of tasks, from image classification to natural language processing and computer vision, neural network architectures are difficult to interpret.

To evaluate outputs, ***ground truth datasets*** are used. These datasets contain ideal expected results. Despite its appeal to objectivity, these datasets are often the result of subjective processes. This can involve hand-labeling example data points. And the way in which ideal expected results are arrived at often lacks transparency.

Self-Learning Models

Self-learning models integrate new data into the model in an automatic way. This can be useful to preserve the accuracy of a model over time.

Consider a model predicting foot traffic. Predictions based on data collected before the COVID-19 pandemic will be inadequate to predict foot traffic during a lockdown situation. Self-learning models can preserve performance by regularly or even continuously integrating new data into the model, in this case including how foot traffic declines because of lockdown measures.

Yet self-learning models can also have the opposite effect. If the training process is automated, performance may decline. For instance, new data may have a lower volume and therefore lead to ***overfitting***. Or new features may be added to the training data that might lead to spurious correlations.

Integration of Third-Party Models

Earlier, we discussed the use of third-party data as creating privacy risks. Similarly, using third-party models or integrating them with your own models creates transparency and explainability risks.

A system may be a black box to you, even when it is transparent and explainable for the third party that built it. If the third party does not provide you with sufficient detail on how their algorithm works, you might introduce a black box system into your organization. In addition, the integration process itself might jeopardize transparency and explainability. You may have to modify the third-party solution to make it fit your needs. If these modifications are not well documented and communicated, you may undermine transparency and explainability.

Intellectual Property Rights

Intellectual property rights include copyrights, trade secrets, and trademarks. Many organizations protect their intellectual property to defend against replication of their product or service. Yet for precisely this reason, closed-source intellectual property introduces transparency and explainability risk.

If you use third-party closed source technology, you might not be able to get sufficient insight to understand the way that the technology transforms data. This creates the risk that you are not aware whether the vendor has done their due diligence in creating the technology. This introduces ethical, legal, and compliance risks, and can block independent auditors from certifying your product. In addition, closed-source intellectual property can turn a system that is transparent to you into a black box for users.

Risks of Explainable AI

There are strategies for making AI more explainable. It is important to note, however, that making AI explainable creates risks of its own. In particular, research has shown that explainability tools can make both users and data scientists over-trust and misread algorithms. This can lead to incorrect assumptions about the model and, curiously, its explanation. As a result, explainability tools can instill a false confidence about deploying the models.

Strikingly, explainability tools have been shown to induce a false sense of confidence even when they were manipulated to show explanations that made no sense. This is not to diminish the importance of explainable AI, but rather to extend the principles that are introduced in this course to explainability tools themselves. In particular, the explanations themselves need to be sufficiently transparent to contribute, rather than detract from, the explainability of a black box system.

Additional Reading

For more information about explainability and false sense of confidence, visit **https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explain-itself-can-make-things-worse/**.

Transparency and Explainability in the GDPR

The introduction of the European Union's GDPR has elevated the requirement of explainability in systems processing personal data. Under the GDPR, organizations using systems processing personal data must be able to explain how the system makes decisions (Article 15(1)(h) and Recital 71). This regulatory requirement makes explainability a key requirement for any automated system. For instance, if your organization uses an ML algorithm to screen applications for relevant skills, complying with GDPR requires being able to answer any applicant's request to explain how the screening decision was made.

Moreover, GDPR also grants individuals a "right of human intervention," which entails the right to request human review of automated decisions (Article 22). The interpretation of these requirements is still evolving. In particular, it is not yet settled what exactly needs to be revealed about the decision to meet GDPR requirements, and what level of explanation is required.

Regardless of how these questions are settled, the GDPR gives regulatory weight to the requirement of explainability. The GDPR is applicable to all companies doing business in the European Union. In addition, the GDPR's approach to explainability is widely seen as trailblazing. Hence, even organizations that do not operate in the European Union should pay attention to explainability as a regulatory requirement.

Additional Reading

For more information about explainability and the GDPR, visit **https://academic.oup.com/ijlit/article/27/2/91/5288563?login=true**.

NIST Principles

The U.S. standard-setter NIST has developed useful guidelines for ensuring an appropriate level of transparency and explainability of AI systems. The guidelines follow four principles:

- AI systems should deliver accompanying evidence or reasons for all their outputs.
- Systems should provide explanations that are meaningful or understandable to individual users.
- The explanation correctly reflects the system's process for generating the output.
- The system operates only under the conditions for which it was designed or when the system reaches a sufficient confidence in its output.

The first principle establishes the general presumption that AI systems should deliver explanations for their outputs. The second principle says what qualities a good explanation should have. This is where disagreements arise. What is understandable for some users may not be understandable for others. What makes for a meaningful explanation will hinge on the practical role the explanation plays in the decision. The third and fourth standards require that the explanation aligns with how the system actually makes decisions, and that the system ensures sufficient accuracy of its outputs and accompanying explanations.

Additional Reading

For more information about the NIST principles, visit **https://www.nist.gov/document/four-principles-explainable-artificial-intelligence-nistir-8312**.

ACTIVITY 7-2

Identifying Potential Sources of Transparency and Explainability Risks

Scenario

For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. Can you list two transparency or explainability risks associated with the RudiBrace product?

2. Are there any groups with respect to these risks that are in problematic risk roles?

TOPIC C

Transparency and Explainability Tradeoffs

As with other ethical risks, you will often have to consider other ethical values as you determine how best to mitigate transparency and explainability risks. In this topic, you will describe commonly encountered transparency and explainability tradeoffs.

Common Transparency Tradeoffs

Transparency is often conducive to accountability and trust, and therefore an essential feature of data-driven systems. Yet, transparency is not always desirable and might conflict with other ethical values. In particular, it can stand in the way of confidentiality, efficiency, and privacy.

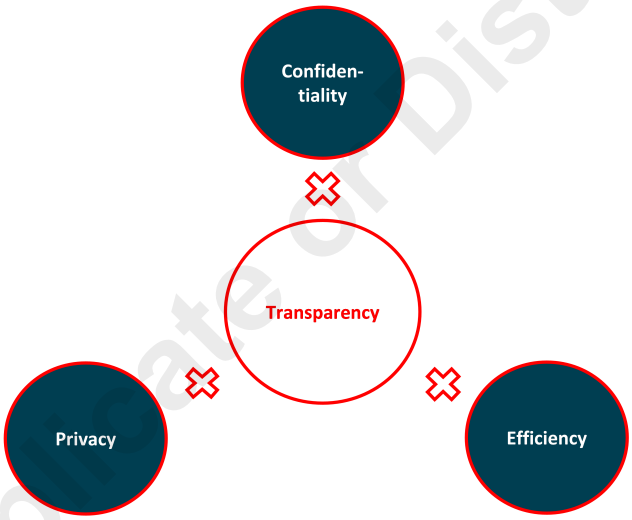


Figure 7-1: Common transparency tradeoffs.

Confidentiality Tradeoffs

Most basically, transparency is about showing and explaining things; making things visible. However, human relations often need a sense of confidentiality: businesses cannot operate without trade secrets, governments need to secure crucial defense information, and relationships between people often involve confidentiality concerning personal information.

Consider the example of agencies that supervise financial actors, like players in the financial technology (FinTech) sector. On one hand, these agencies might allow for full transparency by disclosing all the information collected and managed by companies. This would create a level playing field and would make sure that no player has an unfair advantage or engages in dubious activities. On the other hand, full transparency might conflict with confidentiality agreements between the company and its clients (for instance, about sensitive transactions), and increase the risk of disclosure of personal information.

Once confidentiality is broken for the sake of transparency, this might damage the trust in certain institutions.

Efficiency Tradeoffs

Making our decisions and processes transparent takes a lot of work and can cost significant resources. It can therefore conflict with the goal of operating quickly and efficiently. Transparency can also impose cognitive burdens on an audience. When everything is transparent, it is very hard to focus on the things that really matter. Sometimes, making things transparent can make it even harder for people to understand what is going on. In this way, transparency can conflict with our ability to process information efficiently.

Consider, for instance, the use of AI voice assistants to induce cooperation between humans. Research has shown that voice assistants can be more effective in inducing cooperation than human intermediaries, which would be valuable in cases such as negotiating with malicious actors. However, transparent communication with the voice assistants would require that they disclose their nature, as being artificial rather than human. In this case, increased transparency would defeat the purpose of the voice assistant, thereby reducing its efficiency.

Privacy Tradeoffs

An increase in transparency can promote accountability, but it can thereby also violate privacy. Full transparency would mean who does what is visible, but this also makes it possible to get hold of potentially sensitive personal information.

Consider the case of cryptocurrencies like Bitcoin. Bitcoin's strength lies in the transparency of its shared ledger. Because all nodes have access to the full transaction history of Bitcoin, each user can trust that every transaction is authentic and accounted for. And yet, this transparency comes with a risk. Even though Bitcoin is pseudonymous (transaction addresses are not directly linked to real names), there are ways in which transactions can be linked to individuals. It, therefore, offers full transparency (in terms of transactions) at the cost of privacy.

ACTIVITY 7-3

Discussing Transparency and Explainability Tradeoffs

Scenario

For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

The RudiBrace team considers different tradeoffs of increasing transparency. The team members consider the application context of using RudiBrace to generate activity scores.

Discuss at least one possible way in which more transparency of the RudiBrace might conflict with privacy.

TOPIC D

Mitigate Transparency and Explainability Risks

Now that you know more about transparency and explainability risks and tradeoffs, you should be better equipped to make informed decisions about mitigating those risks in emerging technology projects. In this topic, you will mitigate transparency and explainability risks.

Transparency and Explainability Risk Mitigation Strategies

Explainable AI is AI that is not a black box. A system is explainable to a person if the relationship between inputs and outputs of the system are or can be understood by that person. There is currently a lot of research under way to make ML algorithms more explainable. These tools can help answer questions such as:

- Why did an algorithm make one decision, rather than another?
- How can we correct for failures of the algorithm?
- To what extent can we trust the algorithm?

There is also an important communication aspect to explainable AI. Whether or not a system is a black box can, again, be different for different people. An algorithm that is explainable to the engineers in your organization who made it may not be explainable to users outside your organization, or even to other people within your organization. Not every algorithm needs to be transparent to every stakeholder. Rather, as we have seen, we need to carefully distinguish what the ultimate goals of transparency and explainability are.

Hence building explainable technology goes far beyond tasking software engineers with applying the right set of tools. Building explainable AI starts by determining the transparency and explainability needs of stakeholders.

Determination of the Transparency and Explainability Needs of Stakeholders

There are several steps involved in determining the transparency and explainability needs of stakeholders:

1. Understand which decisions in your organization are made by or with the support of algorithms. For external stakeholders such as users, it is often not even clear which decisions are made by algorithms. But even within the organization, the role of algorithms can easily go unnoticed. Pay particular attention to algorithms that raise flags in terms of one of the four sources of transparency and explainability risk: Models using opaque architectures such as artificial neural networks; self-learning models; models developed by third parties; and closed-source models.
2. Identify whether or not certain stakeholders have particular transparency or explainability needs with respect to an algorithm. As we have seen, explainability and transparency are not necessarily valuable in themselves, but they:
   - Help ensure that decisions are sound.
   - Preempt biases in decision-making.
   - Uphold trust in decision-making
   - Enable self-advocacy.
3. The result of steps 1 and 2 is a list of decision-making processes undertaken or supported by algorithms that are mapped to the transparency and explainability needs of stakeholders. Based on this analysis, you can adopt one of the following mitigation strategies:
   - Explanation of how the system works.
   - Explanation of which factors determined the decision.

- Keep humans in the loop.

Explanation of How the System Works

Recall that explanations may need to look different depending on the stakeholder in question and their explanation needs. Users will often benefit from a narrative explanation outlining the input and output data, including how it is collected and providing a high-level explanation of how data is manipulated by the algorithm. Using visuals like flowcharts is helpful for explaining the process to anyone, not just a lay person. Yet what a meaningful explanation looks like will mostly vary with the need it seeks to address. For instance, to give users an understanding how their behavior influences algorithmic decision-making, they mostly need to know about the input data and how it is collected. By contrast, to enable experts to check whether an algorithm treats different groups fairly, they may require a detailed technical explanation of the model, or even access to the algorithm itself.

Don't expect that you understand all of the explainability needs of users. Often, users will raise concerns and pose questions that you might not have thought of or dismiss as irrelevant based on your intimate knowledge of the algorithm—which users lack. The best way of making the needs of users concrete is to provide them with channels to ask and actively solicit questions, such as by conducting focus groups. In providing explanations, make people aware of potential risks of your product, such as potential inadequacies in training data or limitations in the accuracy of the algorithm.

Explanation of Which Factors Determine a Decision

Another way of approaching the explanation of a black-box system is to focus on which of its inputs drives particular outputs. This is sometimes called the interpretability of a model. There are two kinds of interpretation: global and local. ***Global interpretations*** explain to what extent each input factor contributes to the outputs of a machine learning model. By contrast, ***local interpretations*** explain to what extent each input factor determines a particular prediction.

This table shows different open-source tools that generate interpretations of black-box systems. All tools are open-source and work with Python, the most common programming language used in ML:

Tool	*Use Case*
AI Explainability 360	One of the most sophisticated toolboxes currently available to provide a rich set of local and global interpretation techniques and tutorials.
What-if Tool by Google	Provides a rich set of tools to provide local and global interpretations of models. It specializes in generating high-quality visualizations and requires minimal coding, making it very accessible. It includes a fairness toolbox to check algorithms for different kinds of bias and unfairness.
SHAP (Shapley additive explanations)	A game-theoretic approach to explain the output of any ML model. It provides both local and global interpretations based on Shapley values. The methods are mathematically precise, but they can be slow for some types of models.
LIME	Specializes in local interpretations, using a method that is not guaranteed to be mathematically precise but works well in practice. It tends to be faster than SHAP.

Tool	*Use Case*
ELI5 (Explain Like I'm 5)	Generates local as well as global interpretations of many ML frameworks. It provides a unified API that is easy to learn and has a similar range of features as LIME and SHAP.
ALIBI	Another package with a similar range of features as SHAP, LIME and ELI5 under active development.

Additional Reading

For more information, visit:

- AI Explainability 360: **https://aix360.mybluemix.net/**.
- Google What-If Tool: **https://pair-code.github.io/what-if-tool/**.
- SHAP: **https://github.com/slundberg/shap**.
- LIME: **https://arxiv.org/abs/1602.04938**.
- ELI5: **https://eli5.readthedocs.io/en/latest/**.
- ALIBI: **https://pypi.org/project/alibi/**.

Keeping Humans in the Loop

Keeping a ***human in the loop*** means that a human operator is involved in the decision-making process. Involvement can take different forms. In some approaches, a human is the ultimate decision maker, based on input from the ML system. In others, humans review a subset of decisions, perhaps those that are flagged as potentially problematic. In both approaches, keeping a human in the loop improves transparency and explainability. Because the algorithm is designed from the start in such a way that a human can properly review algorithmic decisions, it is more likely that a wrong decision can be identified and corrected.

ACTIVITY 7-4

Mitigating Transparency and Explainability Risks

Scenario

Use the following tools and questions as you discuss mitigating transparency and explainability risks.

1. Explore the Interactive Demo of the AI Explainable 360 Toolkit, found at **https://aix360.mybluemix.net/data**.
   a) Open a new browser tab or window, and navigate to **https://aix360.mybluemix.net/data**.
   b) On the **Data** page, read the **Introduction**, and then select **Next**.
   c) On the **Consumer** page, select **Bank Customer**, and then select **Next**.
   d) Read the information on the Bank Customer page, and then select one of the customers who is asking for explanations.
   e) Read the information about the customer you selected, and then select **Back**.
   f) Select **Loan Officer**, and then select **Next**.
   g) Read the information on the Loan Officer page, and then select one of the customers.
   h) Read the information about the customer you selected, and then select **Back**.
   i) If you are keen to dive into more technical concepts, check out the explanation for the **Data Scientist** as well.
2. **Is the explanation for the Bank Customer a local or a global interpretation?**
3. **Is the explanation for the Loan Officer a local or a global interpretation?**
4. **Optional: Is the explanation for the Data Scientist a local or a global interpretation?**
5. **In your opinion, which transparency and explainability needs does the explanation for the Bank Customer meet, and which needs does it not meet?**

6. How about the explanation for the Loan Officer?

7. Close the browser window or tab.

ACTIVITY 7-5

Optional: Mitigating Transparency Risks for RudiBrace

Scenario

For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

During the course of developing the RudiBrace product, several ML algorithms were developed, including an algorithm that interprets the sensor data received from the bracelets and computes an activity score that can be viewed in a management dashboard.

1. Think about the algorithm in light of transparency and explainability, and then use the questions in the next two steps to complete this table.

Stakeholder	*Need*	*Mitigation Strategy*

2. What are the transparency and explainability needs of different stakeholders for this algorithm?

3. Can you provide an appropriate mitigation strategy for each stakeholder?

Note: To view a sample completed table, open **C:\095029Data\Identifying and Mitigating Transparency and Explainability Risks\Optional Activity solution table.rtf**.

Summary

In this lesson, you identified and mitigated transparency and explainability risks. You discussed the basics of transparency and explainability, and their relationship to emerging technologies. You also looked at sources for transparency and explainability risks and the regulations and standards that have been developed to protect against these risks. After considering some of the tradeoffs you might need to make between transparency and explainability and other values, you investigated common risk-mitigation techniques. All of this information will be useful to you as an ethical emerging technologist to ensure that you can protect user data from being used in a way other than it was originally intended.

Which types of transparency and explainability risks do you think your organization faces most often, and why?

Which of the transparency and explainability mitigation strategies do you think your organization is most likely to adopt, and why?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

8 Identifying and Mitigating Accountability Risks

Lesson Time: 1 hour, 45 minutes

Lesson Introduction

Emerging technologies challenge the ways in which accountability has traditionally worked. In this lesson, we cover what accountability is, why it matters, and how it applies to emerging technologies. We cover common sources of risks of accountability failures in emerging technologies. Like other values, greater attention to accountability can often come at the expense of other important considerations, and we explore some of those tradeoffs here. Finally, we survey a range of tools for improving accountability and mitigating risks of unaccountable behavior.

Lesson Objectives

In this lesson, you will:

- Describe accountability and its relationship to emerging technologies.
- Identify common sources of accountability risks.
- Describe common accountability tradeoffs.
- Select appropriate tools for mitigating accountability risks.

TOPIC A

What Is Accountability?

Prior to identifying and mitigating accountability risks, it is necessary to clearly define the idea of accountability. In this topic, you will identify the basics of accountability, including its applicability to emerging technologies.

Accountability

Accountability is the practice of holding agents responsible for outcomes, processes, actions, or intentions. It operates in the context of a social relationship: someone is expected to provide an account *to* someone else, *for* some state of affairs, *according to* some set of standards. Accountability comes with the possibility of sanction. Accounts that are inadequate or reveal misconduct invite blame and punishment, while accounts that are satisfactory or reveal excellent conduct invite praise and reward. Accountability operates retrospectively, aiming to assess why and how events occurred. But relationships and mechanisms of accountability are also supposed to incentivize responsible conduct prospectively. Knowing that we may be called to account in the future can and should lead us to act well and prepare to explain our choices.

Accountability can be understood as a virtue and as a mechanism. As a virtue, accountability means reliably answering for our actions. An accountable individual, system, or organization is one that regularly accounts for its behavior to the relevant observers and according to the appropriate standards. Accountability mechanisms are processes and frameworks that facilitate these forms of account-giving and sanctioning.

Accountability is closely related to responsibility, but the two are distinct concepts. Responsibility primarily refers to blameworthiness. To be blameworthy for some bad event, an individual must have been able to foresee the possibility of this event and control the actions that lead to it. In many cases, you can be both responsible and accountable for something. If you borrow a relative's car and damage it through reckless driving, you may be both responsible and accountable. However, you can also be accountable for an event without being responsible for it. If your dog destroys your neighbor's garden, you might not be responsible for this. But you could certainly be held accountable. Similarly, if your subordinate makes a costly mistake in engineering a product, you may be accountable without being responsible.

Why Should We Care About Accountability?

Holding others to account for their actions and decisions is a key part of human relationships—whether personal, professional, or between individuals and companies. We care about accountability for different reasons. Those who are accountable must be able to explain, and justify, their actions and decisions. In this sense, accountability is intertwined with a commitment to truth and transparency. In fact, accountability is a reason we care about transparency in the first place.

Accountability also tells us whom to blame if something went wrong. We all make mistakes. Being accountable means keeping our commitments. When we don't, others can hold us accountable by demanding that we do better. In this way, accountability can promote positive actions and behavior. Sometimes, being accountable means confronting our mistakes—things that didn't go well, and things we should have done better. Accountability is critical to taking meaningful steps to fix our mistakes, provide redress, and work to avoid similar mistakes in the future.

Accountability also limits the unchecked exercise of power. With power, as we're often reminded, comes responsibility, and accountability is a key element in ensuring that power is exercised in the common interest. In this sense, accountability is intertwined with the core values of democracy. Accountability practices enable us to take measures when power is not exercised in the common interest. This means that accountability is vital to limiting the unprincipled exercise of power.

Finally, accountability is intertwined with trust. Stakeholder trust depends on stakeholders' ability to hold organizations accountable to their values and commitments. Would you trust a company—say, with your private data—if you knew there were no accountability mechanisms ensuring it actually complied with its responsibilities? Accountability is a key element of responsible design, and necessary for generating trust in technologies and those who create it.

Accountability in Emerging Technologies

Accountability takes on particular importance in emerging technologies for several reasons.

- Emerging technologies contain tremendous potential for harm and benefit. Those who are harmed are entitled to an answer for what went wrong, to identify and confront responsible parties, to be compensated, and to be assured that the problems have been resolved.
- The complexity of how technologies work creates additional grounds for concern. Those who design, use, or are affected by emerging technologies may not fully understand how they work. Certain technologies, such as those that rely on machine learning, are inherently opaque, making it difficult for individuals to comprehend the reasons behind decisions. When opaque technologies operate in combination with each other or with other complex technologies, tracing mistakes becomes increasingly hard.
- Emerging technologies often involve contributions from many different individuals, organizations, and technical systems. Think especially of the Internet of Things (IoT). The ***problem of "many hands"*** refers to the idea that accountability suffers the more contributors there are involved in a product or event. We may not know why a product malfunctioned, or who is responsible for the malfunction, because the malfunction was the result of the convergence of innumerable actions.
- Decisions in emerging technologies are increasingly made with less and less human control. When decisions are made by AI, it is often not clear what or whom to hold responsible for things that go wrong. This is sometimes called an ***accountability gap***.
- Finally, legal regulation and standards traditionally play a large role in establishing accountability for products and services. Regulation and standards often lag behind the development of emerging technologies, meaning that product developers are left without authoritative guidelines on whom they owe answers, for what, or why. Despite the lack of guidance, technology producers are still expected to take prudent steps toward accountability.

ACTIVITY 8-1
Discussing Accountability Basics

Scenario
Consider the following as you discuss the basics of accountability.

1. Can you describe a situation in your professional life where you observed a lack of accountability?

2. At your workplace, what are some of the barriers to stronger accountability practices?

TOPIC B

Identify Accountability Risks

Before you can act to mitigate accountability risks, you need to know what risks you are up against. In this topic, you will identify accountability risks and their sources, as well as applicable regulations and standards.

Sources for Accountability Risks

Common sources for accountability risks include:

- Technical issues.
- Organizational issues.
- Regulatory issues.

Each of these areas can limit the ability of an organization to account for adverse decisions or consequences. The next few sections describe these sources in more detail.

Technical Accountability Risks

Working with emerging technologies can create numerous accountability risks; i.e., limitations on the ability to account successfully for adverse or controversial decisions or consequences. Risks in product design decisions include those listed in the following table.

Technical Accountability Risk Type	*Description*
Opacity/black box processes	Using techniques like machine learning that are inherently opaque creates more opportunities for accountability failures.
Use of third-party components (open-source libraries, etc.)	The origins and qualities of third-party data sets, products, and tools may be unknown, insecure, erroneous, or biased, making it difficult to trace processes and decisions.
Lack of documentation of decisions and processes	When people are working quickly, it is easy to forget to document decisions made and the rationales behind them; but this creates more problems down the line.
Task delegation to autonomous systems/lack of human control, including delegation of extrajudicial judgments; e.g., by automated weapons	It is often unclear where responsibility lies when autonomous systems result in controversial or harmful actions.
Processes in decentralized systems such as blockchain technologies, including federated learning (decentralized training of algorithms) and executions of smart contracts	Decentralized processes can involve automated execution of code, for which no human user is accountable, and might obfuscate activities of malicious actors.

Organizational and Regulatory Accountability Risks

Accountability risks at the organizational and regulatory levels include those listed in this table.

Organizational/Regulatory Accountability Risk Type	Description
Lack of legal regulation and common standards	Regulations that are vague, contradictory, or missing make it difficult for technology producers to know how to account for their actions, and to whom.
Lack of internal guiding principles	Teams and organizations that do not establish shared guiding principles lack clear expectations of goals, limits, and virtues.
Lack of clear lines of responsibility and liability for decisions	Confusion about roles, or who is liable for what kinds of decisions, makes it difficult to attribute responsibility and sanctions when things go wrong (or right).
Lack of external oversight	Organizations and teams that do not invite scrutiny from external auditors are less likely to establish strong accountability protocols.
Lack of an accountable culture	An organization's culture can lack accountability by promoting behaviors that oppose it. A typical case would be the Silicon Valley mentality of ***move fast and break things***, which held that mistakes are a natural consequence of innovation in a complex, competitive environment.
Case of crises	Risks emerging in times of emergency.

Accountability Standards and Regulations

Standards and regulations are especially underdeveloped for this topic. Law already provides general standards of accountability for legal compliance, including public disclosure requirements, legal liability for harm, government commissions, and public hearings. Civil society holds product developers to account through investigative journalism and watchdog organizations. More specific accountability standards and regulations for emerging technologies are not yet settled. Some of the established regulations and standards include:

- The Sarbanes-Oxley Act of 2002 (SOX) contains various requirements for the governance of U.S. corporations, including requirements on oversight and public disclosure. Although limited to particular kinds of organizations, these requirements may be helpful guidance for organizations of different kinds or operating in different jurisdictions.
- The European Commission's Independent High-Level Expert Group on Artificial Intelligence has proposed various guidelines for accountability of AI-based systems, including auditability, ethics review boards, and "redress by design."
- The Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) organization has proposed a set of principles and best practices for algorithmic accountability.
- The Institute for Electrical and Electronics Engineers (IEEE)'s Global Initiative on Ethics of Autonomous and Intelligent Systems provides principles and practices for accountability.
- Although it hasn't been enacted, U.S. lawmakers proposed the Algorithmic Accountability Act in 2019, which proposed an impact assessment for automated decision-making systems. While not in effect, it gives an idea of the potential direction of accountability regulations in the U.S.

Additional Reading

For more information, visit:

- SOX: **https://www.govinfo.gov/content/pkg/PLAW-107publ204/pdf/PLAW-107publ204.pdf**
- EU Independent High-Level Expert Group on Artificial Intelligence guidelines: **http://doi.org/10.2759/002360**
- FAT/ML principles and best practices: **https://www.fatml.org/resources/principles-for-accountable-algorithms**
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: **https://ethicsinaction.ieee.org/**

ACTIVITY 8-2

Identifying Potential Sources of Accountability Risks

Scenario

Consider the following questions as you discuss the various types of accountability risks presented in this topic.

1. Which types of technical accountability risks do you think are most important?

2. Which types of organizational or regulatory accountability risks do you think are most important?

3. Which types of technical accountability risks do you think need to be mitigated with the greatest urgency?

4. Which types of organizational or regulatory accountability risks do you think need to be mitigated with the greatest urgency?

TOPIC C

Accountability Tradeoffs

As with other ethical risks, you will often have to consider other ethical values as you determine how best to mitigate accountability risks. In this topic, you will describe commonly encountered accountability tradeoffs.

Common Accountability Tradeoffs

Promoting accountability may come at the cost of other values. In particular, accountability may raise concerns in organizations where efficiency, power, growth, and profit tend to be prioritized.

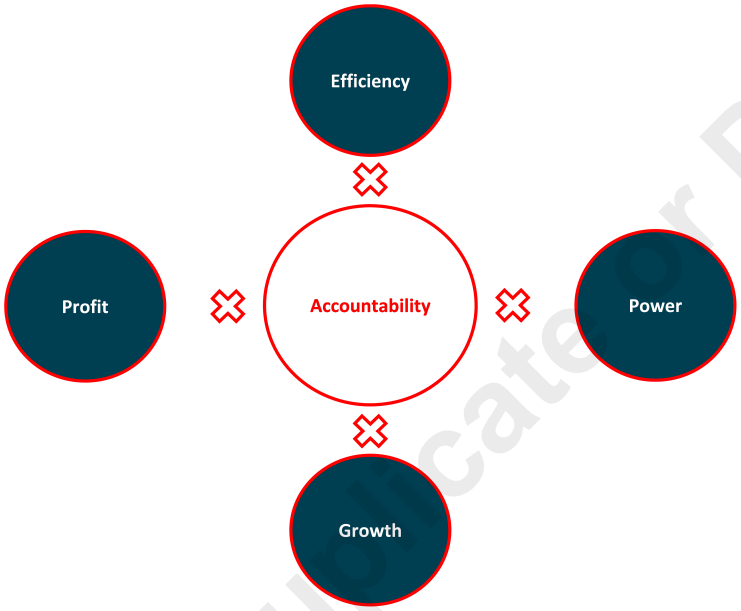


Figure 8-1: Common accountability tradeoffs.

Efficiency Tradeoffs

Data-centered technologies, including big data and AI, enable more efficient decisions. For example, algorithmic decision-making may be faster, more consistent, and more objective than human decisions in many circumstances. However, the opaque nature of these processes makes it more difficult, and in some cases impossible, to understand how decisions are made.

There are many ways and contexts in which increased efficiency complicates accountability. For example, many people worry that we are moving towards a future in which AI systems increasingly decide and act without direct human control, including when it comes to life and death decisions. In extreme cases, one concern is that increased efficiency may also increase the risk of collateral damage. Who should be to blame for a malfunctioning autonomous weapons system causing a civilian massacre? And who should be to blame for a self-driving car running over a pedestrian? The company that made the car, those who designed the software, the authorities that permitted using the car, or the people who used it?

There are immense incentives to create powerful, hyper-efficient decision-making algorithms—even if this comes at the cost of transparency and, ultimately, accountability. Remember the flash crash of

2010, in which the Dow Jones plunged by almost nine percent in just 15 minutes, baffling Wall Street traders and the rest of the world? This was partly a result of the combination of extremely powerful, but also unpredictable and inexplicable, algorithms controlling financial markets.

There is also a more basic way in which efficiency and accountability may come into tension. Taking steps to check and validate work, document processes, review decisions, and prepare materials for oversight takes time, skill, and energy. It is often faster and easier to break things and ask forgiveness later. This attitude is understandable in the face of accountability measures that are poorly designed. But neglecting accountability measures can be extremely dangerous. especially when people's rights are at stake.

Additional Reading

For more information about the flash crash, visit **https://www.businessinsider.com/what-actually-caused-2010-flash-crash-2016-1**.

Power, Growth, and Profit Tradeoffs

Social media companies have enormous power over what people see on the Internet. But major tech companies were not designed to serve the public good. Private companies' primary purpose is to generate growth and profits. To harvest data, for example, the surveillance-advertising business model uses content-selection algorithms to maximize user engagement. These algorithms promote provocative, outrageous content, "fake news," biases, and so on. This has generated calls for better regulation and greater accountability; for example, when it comes to monitoring content on social media platforms.

Greater accountability sometimes means more restrictions. For example, campaigners have pointed out that teenage suicide cases have risen partly as a result of teenagers' ability to view distressing material on social media, as well as instructions on how to take their lives. People have therefore called for social media companies to take greater responsibility, partly by removing more potentially harmful content. Greater accountability in this sense, and more effective content moderation, means more restrictions on what companies allow people to put and see on their platforms.

In many contexts, taking accountability seriously also means that product teams and individuals must disclose details about their products, processes, intentions, and ideas. This may compromise a company's ability to maintain a competitive advantage; it may also invade the privacy of individual people as well as close-knit teams. Such situations require tradeoffs between the needs of technology subjects and third parties for accountability and the needs of technology producers; for instance, with regard to their interest in being able to operate with minimal restrictions.

Additional Reading

For more information, visit **https://ecfr.eu/article/commentary_regulation_and_accountability_how_to_save_the_internet/**.

ACTIVITY 8-3
Discussing Accountability Tradeoffs

Scenario
Consider the following questions as you discuss accountability tradeoffs.

1. If autonomous weapons systems could be shown to significantly reduce civilian deaths in war, does this mean a potential lack of accountability shouldn't stop us from using the technology?

2. If self-driving cars could be shown to significantly reduce the number of accidents in domestic traffic, does this mean a potential lack of accountability shouldn't stop us from using the technology?

TOPIC D

Mitigate Accountability Risks

Now that you know more about accountability risks and tradeoffs, you should be better equipped to make informed decisions about mitigating those risks in emerging technology projects. In this topic, you will mitigate accountability risks.

Accountability Risk Mitigation Strategies

A variety of strategies can help to mitigate accountability risks, including:

- At the **organizational** level:
  - Document company policies clearly, and provide them to all design and development teams.
  - Establish review boards from within or outside your organization to provide oversight of sensitive decisions.
  - Provide opportunities for stakeholders to report concerns, contest outcomes, and interrogate decisions/processes.
  - Require credentialing of technical staff and operators of products whenever appropriate.
  - Cooperate with standard-setting bodies, watchdog organizations, and regulators to improve standards and regulations for your industry.
  - Consider adopting a ***fair competition policy*** to behaving fairly when competing for customers' business and when placing business with suppliers or offset partners. Such policies include not making false claims or remarks that unfairly disparage competitors, or improperly interfering with a competitor's business relationships.
  - Consider adopting an ***open data policy***, which may include a commitment to make data freely available to everyone to use, without restrictions from copyright, patents or other mechanisms of control.
- At the **product design** level:
  - Follow the business conduct guidelines/governance provided by your organization.
  - Establish lines of responsibility and liability for outcomes of each process.
  - Establish standard operating procedures for workflow and interactions with customers.
  - Maintain communication with third-party partners, and clarify division of responsibilities in written contracts.
  - Where appropriate, consider using ***visual contracts*** as a way of creating a binding legal contract without complex legal jargon. A visual contract contains pictures, words and flow charts, which may be easier to understand for non-lawyers.
  - Where appropriate, consider using ***smart contracts***: a self-executing contract with the terms of the agreement between buyer and seller being written in code.
  - Document and record design processes and decisions with the expectation that you may be audited or investigated.
  - Pilot-test all products and document results.
  - Use RACI (Responsible, Accountable, Consulted, Informed) matrices to establish project roles.

Additional Reading

For more information, visit:

- Fair competition: **https://www.ibe.org.uk/knowledge-hub/fair-competition.html**
- Open data: **https://en.wikipedia.org/wiki/Open_data**
- Visual contracts: **https://www.youtube.com/watch?v=YLHsJNkAC9A**
- Smart contracts: **https://www.youtube.com/watch?v=ZE2HxTmxfrI**

RACI Matrices

A ***responsibility assignment matrix (RAM)***, or a ***responsible-accountable-consulted-informed (RACI) matrix***, is a way of allocating project roles and responsibilities. RACI matrices are often part of ***standard operating procedures (SOPs)***, which are step-by-step instructions to help employees carry out routine operations. You may have encountered RACI matrices as a tool in project management and wonder why we discuss them in a course on ethics. The reason is that RACI matrices can be used to enhance accountability, and that is a deeply ethical concern. RACI matrices can help reduce miscommunication and failure to comply with industry regulations. This makes RACI matrices one tool to ensure the human control of technology is responsible and accountable.

Typically, project tasks are arrayed on the leading column and the different people, roles, or departments are arrayed on the header row. Then, use "R," "A," "C," and "I" to populate the boxes.

- Use **R** to match tasks to the person who will primarily accomplish them. It is critical to assign this role with a view to allocating people who have the respective tasks. As we have seen in earlier lessons, ethical failures often happen not because people lack good intentions, but instead they lack the relevant experience or competence. Assigning suitable people to a task is therefore critical for avoiding ethical failures.
- Use **A** to match tasks to the person who will oversee the project and provide an account to external departments for the project outcomes. To foster accountability, tasks should not have more than one **A** entry in the matrix to avoid diffusion of responsibility. If this is the case, the task likely can be separated into multiple tasks. Continue to redefine the tasks until one single person is accountable.
- Use **C** to match tasks to people who will provide input on them. Who is consulted is a deeply ethically charged choice, as this determines which stakeholders will get a say in the decision.
- Use **I** to match tasks to people who will be informed of project progress and completion but not provide input or supervision. Not all stakeholders will be consulted, so keeping those stakeholders informed is critical to avoid surprises and provide a backstop against overlooking an important ethical consideration.

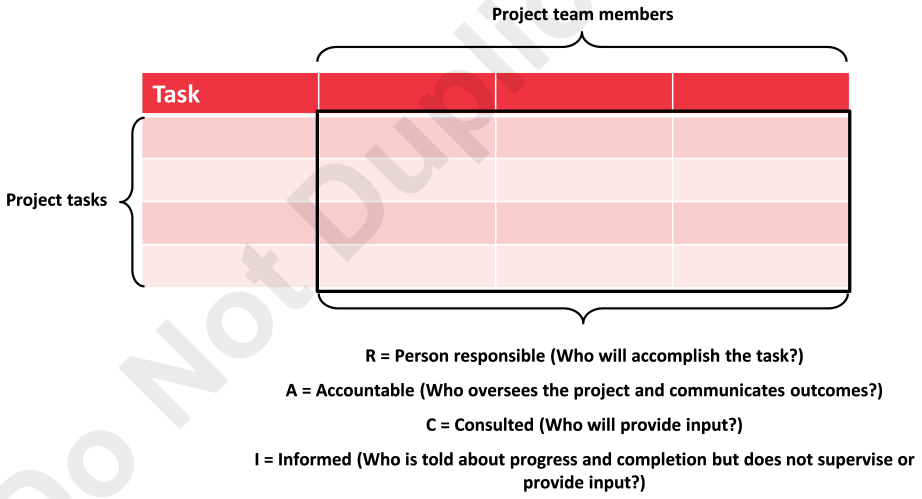


Figure 8-2: A RACI matrix.

Mitigation Tools

Here are just a few of the mitigation tools available to address accountability risks:

- The **Algorithmic Impact Assessment** is a tool that has gained interest in Canada, the EU, and the United States. Modeled on the idea of an environmental impact assessment, it requires

entities to forecast and disclose potential impacts of algorithmic systems, provide opportunities for public comment, and respond to those comments. Although initially devised for use by government agencies, the tool can also be modified for use in other settings.

- The **Responsible AI Design Assistant** from AI Global is an online tool that helps users identify various ethical risks in AI models. It pays particular attention to accountability risks.
- **Data visualization** and **dashboard reporting** can also be used to identify potential errors and malfunctions, as well as to demonstrate performance of products to stakeholders without revealing confidential information.
- Researchers at Google have proposed an **auditing framework** called SMACTR (Scoping, Mapping, Artifact Collection, Testing, and Reflection) for use in various organizations. The framework comes with templates for adaptation.
- **Human in/on the loop** refers to ways of ensuring that human beings are involved in sensitive decision-making by autonomous systems. One way to improve the accountability of autonomous systems is to ensure that they are always under some form of human control.
  - Human in the loop means that only a human agent can make the decision, with advice from an algorithm.
  - Human on the loop means that the algorithm can operate on auto-pilot with human agents standing by to intervene if necessary.

Additional Reading

For more information, visit:

- AIA: **https://ainowinstitute.org/aiareport2018.pdf**
- Responsible AI Design Assistant: **https://oproma.github.io/rai-trustindex/**
- SMACTR: **https://arxiv.org/abs/2001.00973**

ACTIVITY 8-4

Mitigating Accountability Risks

Scenario

For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

Remember, RACI stands for Responsible, Accountable, Consulted, and Informed.

Note: There may be duplications of RACI roles assigned to a particular task. There may also be individuals with no RACI role for a particular task.

1. Complete this RACI matrix for the RudiBrace team as it plans to test its product on volunteers.

	Chief Product Officer	Project Manager	Hardware Engineer	Data Scientist	User Experience Designer	Software Engineer
Create test plan and KPIs/KRIs						
Interview test subjects						
Record and analyze test data						
Monitor software performance						
Confirm device safety						

	Chief Product Officer	Project Manager	Hardware Engineer	Data Scientist	User Experience Designer	Software Engineer
Create test plan and KPIs/KRIs	A	R	C	C	C	C
Interview test subjects	--	A	I	C	R	I
Record and analyze test data	--	A	C	R	C	C
Monitor software performance	--	A	I	C	I	R
Confirm device safety	--	A	R	I	C	I

2. Compare your matrix with other participants, and discuss the differences that you notice.

3. Are there other project team members who should be consulted or informed about any of these risks?

Summary

In this lesson, you identified and mitigated accountability risks. You discussed the basics of accountability, and its relationship to emerging technologies. You also looked at sources for accountability risks and the regulations and standards that have been developed to protect against these risks. After considering some of the tradeoffs you might need to make between accountability and other values, you investigated common risk-mitigation techniques. All of this information will be useful to you as an ethical emerging technologist to ensure that you can protect user data from being used in a way other than it was originally intended.

Which types of accountability risk do you think your organization faces most often, and why?

Which of the accountability risk mitigation strategies do you think your organization is most likely to adopt, and why?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

9 Building an Ethical Organization

Lesson Time: 1 hour, 45 minutes

Lesson Introduction

Developing and deploying emerging technologies in an ethical way requires ethical organizations. This means having an organizational culture that supports ethics: having an ethical purpose, ethical values, awareness of ethical impacts, and standards of professional ethics.

Lesson Objectives

In this lesson, you will:

- Describe an ethical organization, including the aspects of culture and systems that are required for building an ethical organization, as well as some examples of ethical failure due to organizational issues.
- Describe organizational purpose and ethical values, and how these connect to everyday practice.
- Explain the significance of ethical awareness in organizations.
- Develop professional ethics within organizations.

TOPIC A

What Are Ethical Organizations?

The ability to build an ethical culture requires understanding of the current culture of the organization as well as what steps need to be taken to achieve an organizational culture that prizes ethical behavior. In this topic, you will describe ethical organizations.

What Is Organizational Culture?

What does it mean for ***organizational culture*** to be "ethical"? An organization's culture encompasses its values, principles, and practices that guide its members' actions. It is shaped by the organization's leadership, governance structure, policies, and its various stakeholders. An organization's culture is more than its mission statement—it's the result of authentic and consistent conduct and behavior. An organization's culture comprises a set of shared assumptions that guide individuals' actions and specify appropriate behavior in different organizational contexts. Or, put another way, organizational culture is "the way we do things in this organization."

Ethical culture refers to that aspect of an organization's culture that supports the organization and its members in consistently doing the right thing. Key elements of ethical culture include:

- An organization's purpose.
- The ways in which the organization rewards ethical behavior.
- The organization's degree of sensitivity to ethical issues.
- The organization's capacity to engage in ethical deliberation.
- The accountability practices adopted and followed by the organization.

If organizational culture is about "the way we do things in this organization," then ethical culture is about "the way we approach questions about 'what's the right thing to do?' in this organization."

Why Should We Care About Organizational Culture?

An organization's culture is inextricably intertwined with its performance. As we have seen, data-driven organizations face enormous responsibilities as a result of their powers and capacities. Many vital aspects of how organizations behave—from how they respond to challenges and crises, how they adapt to change, and how effectively they advance technological and social progress—are informed by their culture.

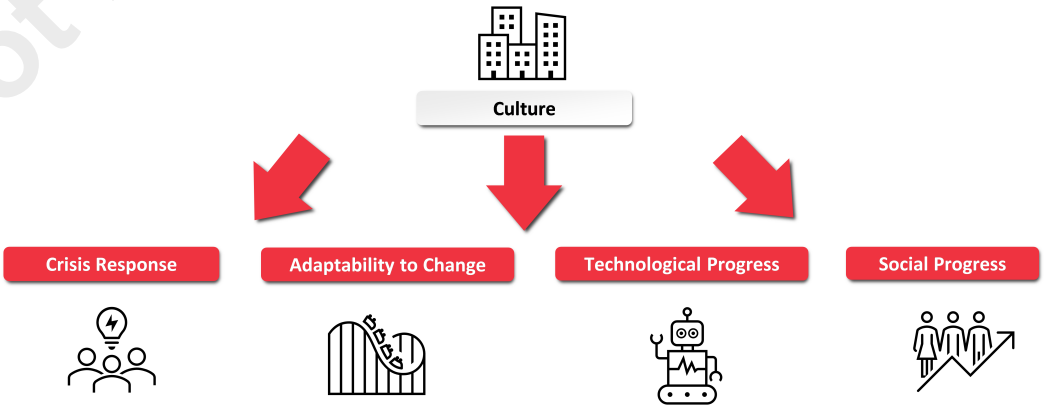


Figure 9-1: An organization's culture affects many facets of the organization.

An organization's culture informs the way in which product decisions are made, which can have a huge impact on what products are created, launched, and how they are ultimately used. For example, the attitude of "move fast and break things" has resulted in a kind of corporate culture in which questions like "*Can* we build this product?" took priority over questions like "*Should* we build this product?" Imagine if we applied the "move fast and break things" mantra to the development of autonomous weapons systems.

Organizational Culture and Ethics in Practice

A healthy organizational culture helps employees act in accordance with basic moral principles and makes it significantly less likely that people will violate legal and regulatory standards. A company's culture can make a significant difference in how likely people are to do the right thing.

Some cultures tolerate, or even encourage, misconduct. One of the most common problems is a kind of corporate culture that focuses exclusively on profits. When this comes at the cost of the interests of customers, suppliers, or other stakeholders, this can have disastrous consequences.

For example, the global oil company BP suffered several crises as a result of deficiencies in ***corporate culture***. The 2010 Deepwater Horizon drilling rig explosion in the Gulf of Mexico was the largest environmental disaster in history, and it cost BP more than $65 billion in total. In previous years, BP had suffered several smaller incidents, including an explosion in a refinery in Texas in 2005, which killed 15 employees and injured 200, in addition to other fatal accidents and spills, to which BP failed to respond. BP's disregard for worker safety, legal requirements, and social and environmental concerns was a concomitant of its exclusive focus on profit. Had BP taken safety procedures, governance, and ethical culture more seriously, lethal incidents and large-scale disasters might have been avoided.

Another case where deficiencies in corporate culture had lethal consequences is that of Boeing. Shortcomings in Boeing's company culture led to a lack of safety and fatal plane crashes. Boeing's culture rewarded efficiency, and the containment of costs, over safety. This ultimately resulted in the design of unsafe pilot software, which in turn caused several planes to crash—including the crashes of two Boeing 737 MAX airplanes in 2018 and 2019, which caused the deaths of 346 people. After the crashes, many stories emerged of employees covering up deficiencies and hiding problems from regulators, as Boeing was trying to evade scrutiny and prevent its employees from speaking up. Messages that the systems were faulty had emerged as early as 2016. Had Boeing's culture encouraged responsible behavior, and prioritized safety over profit, those deadly crashes might have been prevented.

What cases like these illustrate is that, even if harms are not intentional, adequate ethical practices, governance, and management systems are vital to preventing harms. In certain cases, the behaviors encouraged by a company's culture may make the difference between life and death. Culture is not only key to promoting behavior in keeping with ethical standards and legal obligations. It can also save lives.

ACTIVITY 9-1
Discussing Ethical Organizations

Scenario
Consider the following questions as you discuss the contents of this topic.

1. Can you think of examples of company culture influencing employee behavior in a positive manner?

2. Can you think of examples of company culture influencing employee behavior in a negative manner?

TOPIC B

Organizational Purpose

This topic looks at the very heart of the organization and its ethical culture. It asks:

- Why does an organization exist?
- What are its values?
- And what impact does it want to have on the world?

These questions will be answered by discussing organizational purpose, ethical values, and commitments towards society.

The Core Questions

An organization or company also has aim or purpose. In the capitalist economy, for-profit companies need to make a profit to be economically sustainable. Yet purpose points to an organization's reason for being beyond financial objectives. Ultimately, organizations exist to contribute to society in some way. As we have seen, this is one way of understanding what it means to do the right thing.

For this reason, building an ethical organization starts with asking some basic questions. These are:

- Why do we exist? What is the purpose beyond profit of our organization? What do we want to contribute to society?
- How do we behave? What are our core values, around which we model our organization? What constitutes a job well done?
- What do we do? How does the work we do impact the world? And how do we understand and assess this impact?

What Is Organizational Purpose?

To ask for the purpose of an organization, is to ask: Why do we exist? This is a simple question, and yet, it's often difficult to answer. Often, organizations already have a stated purpose, yet it is important to frequently reflect on this purpose with the idea of an ethical organization in mind. The purpose of an organization is often presented in terms of its mission.

1. The first step in formulating an organizational purpose is accepting that the organization exists to make people's lives better, and not just to make profits. It therefore consists in understanding what people would lack if the organization wouldn't be there.

   Philips describes its purpose as: "Improving people's lives through meaningful innovation."
2. The second step in developing an organizational purpose is generating *ownership* of this purpose. This means that an organization's leadership should consult with employees and other stakeholders in the organization to ensure alignment. If the stated purpose of an organization does not resonate with its members, it breeds disengagement, rather than alignment.
3. The third step is considering purpose from different points of view:
   - What does our organization bring to the *customer*?
   - How do we contribute to our *industry*?
   - What *greater cause* does our organization help realize?
   - How do we help our *communities*?
   - And what does the organization bring to its *employees*?

It is important to note that an organization's purpose is not marketing talk. To be effective, it needs to reflect the true ambition of the organization; i.e., be aligned with what the organization does and how people work in a day-to-day fashion.

Sometimes, organizational purpose can be aligned with greater ethical challenges in society, such as the 17 UN goals for sustainable development.

Consider, for instance, the mission statement of Tesla: "To accelerate the advent of sustainable transport by bringing compelling mass market electric cars to market as soon as possible."

Additional Reading

For more information about the UN goals for sustainable development, visit **https://sdgs.un.org/goals**.

Organizational Values

An organization's purpose is broad and future-oriented and needs to be translated into day-to-day reality. The related question to ask is: Given our purpose, how do we think and behave?

It is important to frame organizational values not too broadly, covering everything that humans find important, because this will make them vague and ineffective. Rather, values reflect what an organizational culture is really about, what sets it apart from other organizations. It therefore usually consists of only a handful of statements.

To build an ethical organization, values need to be approached critically. Some values, such as "move fast and break things" can have severe negative side effects. To determine core values, one should therefore ask two questions:

- How should we behave?
- How could this behavior have negative impacts?

Moreover, organizations need to guarantee a level of professional autonomy, to give individuals the capacity to critically reflect on whether or not their organization embodies the right values.

Example Value Statements

Consider some of the core values of Ben & Jerry's Ice Cream:

- We strive to minimize our negative impact on the environment.
- We strive to show a deep respect for human beings inside and outside our company and for the communities in which they live.

These statements effectively communicate the goals of the organization's culture and its vision statement: "Making the best possible ice cream, in the nicest possible way."

How Values Fit the Organization

There are some compelling reasons for organizations to define certain ethical values.

1. First, these values help the organization navigate toward its purpose without causing harm.
2. Second, it makes sure the work done by an organization is aligned with values our societies embrace, such as *human rights* and *principles for medical ethics* like the Belmont Report.
3. Third, there is a growing trend in industries to make business-to-business interactions dependent on having shared ethical values.

Ethical values are organization- and product-specific. When an organization develops a technology with a very narrow function, such as an electronic drill, it should commit to respect safety and accountability, but will not have to deal with themes such as fairness and transparency. In other words, values related to societal impacts are context-dependent, and each organization has the task of understanding its own context.

Additional Reading

For more information about the Belmont Report, visit: **https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html**.

Guidelines for Developing Organizational Values

Note: All Guidelines for this lesson are available as checklists from the **Checklist** tile on the CHOICE Course screen.

Follow these guidelines when you are developing organizational values.

Develop Organizational Values

Organizations can develop their values in the following way:

- Collect and analyze organizational values from existing authoritative sources. For instance, they can consult academic work on AI principles such as the Harvard report on Principled AI; company codes such as Google's AI Principles and the Microsoft responsible AI principles.
- Draft organizational value statements, taking into account the products and services the organization offers and their potential impacts on society. Organizations can use tools such as consequence scanning, scenario analysis, and ethical frameworks and theories to identify and analyze those impacts.
- Identify internal and external stakeholders, such as employees and interest groups, to discuss the relevance of proposed organizational values, and to check whether some organizational values are missing.
- Publish the organizational values, and raise awareness about them within the organization.

Additional Reading

For more information, visit:

- Harvard report on Principled AI: **https://cyber.harvard.edu/publication/2020/principled-ai**.
- Google's AI Principles: **https://ai.google/principles/**.
- Microsoft responsible AI principles: **https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6**.

Making Values Count

Ethical values are important guidelines that shape the work done in an organization; for instance, signaling that the technologies that are developed should respect things like privacy, accountability, and fairness.

Still, organizations should also put effort into translating values into measurable action. One way of doing so is following the ***VCIO model***, outlining how values can be translated into criteria that have certain indicators linked to observable actions.

To use the VCIO model:

1. First, the organization determines its **values**. For example: transparency.
2. Second, each value is translated into **criteria**—requirements of products or services that should be met to fulfill the commitment implied by a stated value. For example: disclosure of original data sets.
3. Third, these criteria are linked to **indicators**—certain features of a product or service that indicate whether criteria are met or not. For example: Is the origin of the data that's used documented?
4. Fourth, these criteria are linked to **observable behaviors** that can be measured in the organization. For example: logging of all training and operating data; version control of the datasets.

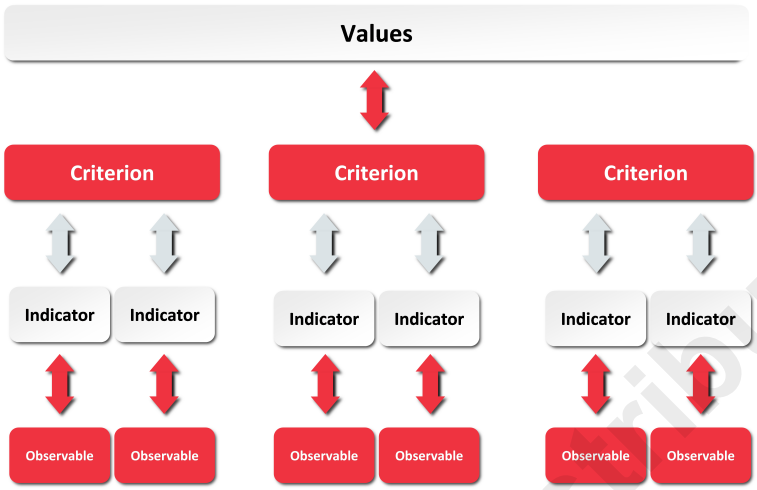


Figure 9-2: The VCIO model developed by the AI Ethics Impact Group.

Purpose and Value Statements and Their Impact

Even though organizational purpose and values are of great importance, they are also often met with reasonable skepticism. History teaches us that some organizations that presented themselves in accordance with very aspirational ideals have caused great harm to people and society. It is one thing for an organization to state that it values ideas like sustainability and well-being, it is another thing to also act on those values.

In many organizations, the leadership and employees are only vaguely aware of the organizational purpose and value statements, and have had little say in how these were formulated. This generates a lack of ownership and effect.

It is therefore important to consider a purpose and values statement as only the tip of an iceberg. These should be the expression of what an organization truly stands for and works towards. What is hidden under the surface is how organizational principles and values affect the day-to-day work of the leadership and employees. This is where ethics starts to count.

ACTIVITY 9-2
Discussing Organizational Purpose

Scenario
Consider the following questions as you discuss the contents of this topic.

1. **Which of these statements of organizational purpose do you think most reflects an ethical organizational culture?**
   - ○ To enrich people's lives with programs and services that inform, educate, and entertain.
   - ○ To make it easy to do business anywhere.
   - ○ Our deepest purpose as an organization is helping support the health, well-being, and healing of both people—customers, Team Members, and business organizations in general—and the planet.
2. **Why did you select that answer?**
3. **Which of these statements of organizational purpose do you think least reflects an ethical organizational culture?**
   - ○ To enrich people's lives with programs and services that inform, educate, and entertain.
   - ○ To make it easy to do business anywhere.
   - ○ Our deepest purpose as an organization is helping support the health, well-being, and healing of both people—customers, Team Members, and business organizations in general—and the planet.
4. **Why did you select that answer?**
5. **(Optional) With what you know about the RudiBrace product, draft organizational purpose and value statements for Rudison Technologies.**
6. (Optional) Open a new browser tab or window, and search for organizational purpose and value statements that you think promote and support an ethical organization. If your instructor requests, you can share your findings with the other participants.

TOPIC C

Ethics Awareness

An organization can spend money and time crafting organizational purpose and value statements, but all that investment can go to waste if employees and other stakeholders are not aware of the organization's posture. This topic focuses on raising awareness of ethical issues and practices.

Why Awareness Matters

At the start of the course, we discussed reasons why ethical risks are often overlooked. We found that while people acting with bad intent sometimes do cause havoc, other reasons are more common, such as:

- Failure to consider key stakeholders.
- Assuming that someone else is in charge.
- Lack of practices for anticipating ethical risks.

These factors all come down to two things: First, a lack of awareness of ethical risks that we have already covered. Second, impacts of and on your business. This table lists some of the impacts that people in the organization need to be aware of to make good decisions.

Type of Business Impact	*Example*
Negative social impacts	How your organization affects issues such as inequality, fairness, or political polarization in society.
Loss of organizational reputation	How a decision to ban messages by certain politicians on your social media platform impacts your organization's reputation.
Loss of consumer trust	How introducing algorithmic decision-making for loan approval impacts trust by customers in your company.
Liability	When acquiring a company, your organization may become liable for harms caused by the acquired company, even if you knew nothing about those impacts.
Legal/regulatory obligations	By processing personal data, you become subject to strict privacy laws like GDPR, causing significant compliance costs.

Drivers of Awareness

The many types of ethical risks and business impacts organizations need to be aware of show that building awareness is a difficult task. Moreover, when dealing with emerging technologies, small decisions by individual contributors can have major impacts. Hence, awareness needs to be built at all levels in the organization. What can organizations do?

- **Clarity of Purpose and Expectations**: Communicate the purpose of the organization and the expectations of employees clearly and regularly.
- **Diversity of Thought**: Hire people that bring different perspectives, and staff teams to create diversity of thought.

- **Organizational Memory**: Learn from mistakes and spread that knowledge throughout the organization.
- **Speak Up!**: Encourage employees to share knowledge and challenge superiors.

The next sections look at these options in detail.

Clarity of Purpose and Expectations

The ability to connect your day-to-day work with the purpose of the organization is one of the most important drivers of ethical conduct in organizations. Awareness of one's contribution to the organization's purpose beyond profit acts as a moral compass for decision-making. It encourages everyone in the organization to hold themselves to high ethical standards. But this sense of contribution to a shared purpose is not achieved by painting a lofty purpose statement on the wall. Organizations need to make values count, as explained in the previous topic. Moreover, leaders need to credibly endorse the purpose. This requires clear and constant communication. But mostly it requires taking action that clearly embodies the purpose. Leaders that credibly endorse the purpose of the organization can share assessments about how the organization is already making good on its purpose, and where it is still falling short. They can use these opportunities to reinforce the ethical expectations of the organization towards every employee, which may also be codified in policies such as a code of conduct.

Three ways in which organizations can drive clarity of purpose and expectations include:

- **Leadership communication**: Leadership should refer to purpose and values as frequently as possible and explain how the organization made crucial decisions in light of purpose and values. For instance, leaders can make it a habit to share a story about a brave decision someone in the organization took that brings to life the organization's purpose and expectations.
- **Iconic actions**: Identify highly visible actions or policies that encapsulate your organization's purpose and the expectations on its people. For instance, Netflix drove home the expectation that employees manage their own time responsibly with their unlimited vacation policy, according to which any employee can take as much paid vacation time as they like.
- **Ethics moments**: Invite members of the organization to raise difficult decisions they faced or currently face in a safe and open forum in the organization. The purpose is to invite members of the organization to reflect about difficult decisions in light of purpose and expectations. For instance, organizations can reserve the first 20 minutes of monthly all-hands meetings to reflect on an ethics moment.

Additional Reading

For more information about the Netflix no-vacation limits policy, visit **https://www.inc.com/justin-bariso/netflixs-unlimited-vacation-policy-took-years-to-get-right-its-a-lesson-in-emotional-intelligence.html**.

Diversity of Thought

Diversity of thought is the idea that people in a group should bring varying, diverse viewpoints to the table. While diversity and inclusion efforts are not sufficient to achieve diversity of thought, they are crucially important to achieve that aim. Recall how Facebook discovered that a prototype of the smart camera in its video-conferencing tool Portal did not track non-white people reliably. It is no coincidence that Technical Business Lead Lade Obamehinti, a person of color herself, noticed the problem. The more diverse the teams working on new technology, the more likely it is that someone will spot an ethical risk.

Yet other types of diversity are also required to achieve diversity of thought. Even if teams are diverse in terms of dimensions like gender, ethnicity, and religion, they may share a similar mindset if they all graduated from the same university programs at the same few universities, or are from a similar socio-economic background. Nor can teams ever hope to represent the full spectrum of relevant perspectives. This makes it crucial to create teams whose members have different cognitive

styles, and to encourage open-mindedness. Moreover, engagement with stakeholders and outside experts is crucial to foster diversity of thought.

Here are four things organizations can do:

- **Diversity, Equity, and Inclusion (DEI) efforts**: Develop a Diversity and Inclusion policy, and provide training on Diversity and Inclusion for all employees.
- **Encourage open-mindedness and intellectual humility**: Insist on every employee displaying respect for other viewpoints; role-model willingness to revise your own standpoint.
- **Engage stakeholders and outside experts**: Invite outside experts or stakeholder representatives to an open discussion to hear their hopes and concerns first-hand.
- **Use ethical foresight tools**: Earlier in the course, we introduced tools to identify ethical risks, such as consequence scanning, scenario analysis, and stakeholder cards. For example, you can conduct a scenario analysis as part of your regular planning process.

Additional Reading

For more information, visit:

- AI and bias: **https://www.nytimes.com/2021/03/15/technology/artificial-intelligence-google-bias.html**.
- Diversity and inclusion: **https://hbr.org/2020/05/diversity-and-inclusion-efforts-that-really-work**.
- Respect for other viewpoints: **https://hbr.org/2018/11/a-new-way-to-become-more-open-minded**.

Organizational Memory

While building ethical foresight into everyday practices allows organizations to anticipate ethical risks, strengthening organizational memory allows them to learn from their mistakes. ***Organizational memory*** is the ability of organizations to record, share, and act on learnings from the past. Organizations can learn from mistakes by reflecting on the root causes of ethical incidents, looking at what could have been done differently, and reviewing what can be done to prevent incidents in the future.

Organizations should take the following steps to strengthen institutional memory:

- **Establish a Review Board**: Establish an external ethics review board and/or an internal ethics function.
- **Investigate ethical failures**: Investigate ethical failures to identify individual, as well as structural, root causes.
- **Share learnings**: Share learnings with employees, and embed learnings in ethics training.
- **Document learnings**: Document ethics review and decision-making processes.

Speak Up!

When organizations fail to anticipate an ethical risk, this does not mean that it went unnoticed by everybody. The Challenger exploded because the O-ring seals used in a critical part of the shuttle were not designed to handle the unusually cold weather on launch day. The night before the Challenger launch, a group of engineers met with NASA officials in an emergency meeting. Despite having full knowledge of faulty O-rings on the shuttle, the engineers remained silent.

To avoid speak-up failures, organizations should create an open culture where people can raise serious concerns and be heard. Leaders should place a responsibility on employees to speak up if they disagree. To be credible, leaders must react to criticism in an open and welcoming way. A no-retaliation policy can also help encourage employees to speak up. However, note that encouraging speak-up can get organizations only so far. Organizations should not rely on employees bravely speaking up to correct for fundamental flaws in the organization's governance or leadership. Here are two things organizations can do to encourage speak-up:

- **Foster a culture of trust and candor**: Employees must trust and respect one another and the organization's leadership for candid feedback to flow freely. Receiving challenging feedback constructively is something leaders must learn.
- **Create safe and confidential reporting channels**: Additionally, an ethics hotline allows employees to speak up anonymously. Ethics hotlines should be operated by a trusted third-party vendor.

Additional Reading

For more information, visit:

- Culture of trust: **https://hbr.org/2016/01/creating-a-culture-where-employees-speak-up**.
- Safe reporting channels: **https://hbr.org/2021/02/how-to-encourage-employees-to-speak-up-when-they-see-wrongdoing**.

ACTIVITY 9-3
Identifying Drivers of Awareness

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. List two or three ethical risks and/or business impacts that you want to drive awareness of.

2. Select two drivers of ethical awareness and explain how they can help improve awareness.

TOPIC D

Develop Professional Ethics within Organizations

This topic looks at the importance of the quality of people in an organization and professional ethics. Even when an organization has an ethical purpose and clear ethical values, as well as strict policies concerning ethics, its leadership and employees might behave badly (both through ill intent and neglect). Therefore, an ethical organization needs ethical people.

People's Qualities and Organizational Culture

Ethical organizations depend on the qualities of their people. We expect people in organizations to have certain dispositions, like collegiality, loyalty, vigilance, and prudence. What this means differs for each organization. For instance, for people working with nuclear power prudence means to be extremely cautious and careful, while for athletes it means responsibly pushing one's performance to the limits.

Unethical organizations cultivate unethical dispositions and therefore "make" bad people. For instance, criminal organizations cultivate dispositions like distrust, dishonesty, and secrecy.

This also means that ethical organizations stimulate good qualities in their people. This can be done in many ways. For instance:

- **Community building**: Make everyone feel they stand for the same ethical purpose and ethical values.
- **Professional skills development**: Focusing on ethical elements in addition to technical elements when developing staff skills.
- **Training and education**: Align people's skills with the ethical values of the organization.
- **Team management**: Build ethical values between collegial relations, ensuring that colleagues will support and challenge each other in making the right decisions.

Each of these is described in the next few sections.

Ethics in Traditional Professions

Professions such as medicine and finance have realized the huge impact they have on society and the responsibility this brings. Consequently, organizations in these fields have made steps towards shaping a sense of professional ethics.

Medicine is a case in point. Already in Ancient Greece, physicians bound themselves to the Hippocratic Oath, named after Hippocrates. By invoking this oath, physicians swear to uphold certain ethical values, such as medical confidentiality and non-maleficence. Today, a version of this oath is still in use in medical schools. The idea is that being a good doctor does not only mean to have the right medical skills, but also to have a good character: to treat patients with respect and dignity.

The tech sector can learn from this example. For instance, to be a good software developer does not only mean to have the right technical skills and build innovative products, but also to pay attention to the potential positive and negative impacts of these products on society.

In part, this shift is visible in a change of the ways in which professional ethics is framed. For instance, Google's motto of professional ethics started out as being "don't be evil" but later changed to "do the right thing."

Community Building

Every organization is also a community: a network of ongoing, relatively stable relations between people. These people hold diverse views. And, yet, they also have a basis of shared ethical values and norms. Community building is about balancing this diversity and shared values: respecting people's individuality and liberty, while at the same time asking them to live up to some basic, shared ethical values.

Ethical community building does not happen at one point in time, during a single meeting, but is a never-ending task. At the heart of community building lies the *conversation*; it requires people to come together and discuss the ethical values that bind them. A difficulty of community building is that it is an emergent process: it happens bottom-up rather than top-down. The leadership can therefore not force it on the rest of the organization.

However, there are some central ways to *facilitate* community building:

- The leadership of an organization should re-imagine its role. It is not the final authority on matters that concern the organization but has a role to *serve* the organization. This serving happens by giving others the chance to speak, and to listen to their opinions and concerns.
- Create a space for organization-wide discussion and deliberation. Such a space is often multi-faceted—it consists of face-to-face conversations between the leadership and other people in the organization, town hall meetings, culture building workshops, and online spaces such as a discussion forum.
- Frame the conversation on community building—*not* around concrete policies and design problems, but around the "bigger picture" issues, which include organizational purpose and ethical values.
- Frame the conversation *not* around issues and problems that are likely to generate consensus, but around those that inspire a diversity of opinions. Disagreement and conflict are essential to deal with to finally build a community.

Professional Skills Development

Most professional skills development focuses on technical abilities that are required to develop products and services. Consider, for instance, coding skills, accounting skills, or planning skills. These hard skills are essential for an organization to function.

Ethics, however, depends largely on communicative skills and critical thinking. Earlier, we explained that ethics has a lot to do with finding good reasons for acting in a certain way. Soft skills are needed to find and express such good reasons.

Organizations can do different things to align professional skills with the skills needed for ethics:

- **Ethical oath**: As we saw earlier, physicians declare an ethical oath to align their technical work with ethical values. An organization can write an ethical oath by listing the core virtues it wants its members to have as imperatives: e.g., "in my profession, I will act transparently." No official ethical oath for the emerging technologies sector exists, but organizations can gain inspiration from the Archimedean Oath for engineers, or the data science oath.
- **Training**: Organizations can facilitate internal trainings for their members to gain the essential soft skills needed for ethics. This involves both knowledge-intensive trainings, like trainings that link to this instruction material, and practical engagement with ethics, such as ethical foresight workshops.
- **Organizational resourcing**: Organizations can invest in ethics by aligning professional roles with the ethical impact they want to have. For instance, they can hire or train ethics experts who have an exemplary function, inspiring others to behave ethically and educating them.

Additional Reading

For more information, visit:

- Archimedean Oath for engineers: **https://en.wikipedia.org/wiki/Archimedean_Oath**.

- Data science oath: **https://www.nap.edu/read/24886/chapter/7#32**.

Training and Education

Scientific studies have shown that professionals who receive ethics education have a significantly higher ability to perceive ethical issues and are reported to act more frequently to uphold ethical conduct. In other words, ethics education works.

In selecting an ethics training and education program, organizations can take the following points into account:

- Ensure that the program is organized around industry-specific, practical cases. Cases help people to easily understand what might go well or wrong in organizations in terms of ethics.
- Use materials from organizations that specialize in professional education. A lot of material on organizational and technology ethics is written in an academic style that is hard to access. Organizations that specialize in professional education offer state-of-the-art ethics curricula that are aligned with the practical reality of organizations working with emerging technologies.
- Make use of practical tools and heuristics to discuss ethics, such as the Ethical OS Toolkit or the Ethics Canvas.
- Customize the program for different roles in the organization. Some people will need to be educated as ethics experts, which requires intensive training, while others will need to be acquainted with the basics, which can be taught in less intensive teaching blocks.

Additional Reading

For more information, visit:

- Scientific studies about ethics education: **https://www.tandfonline.com/doi/abs/10.1080/15236803.2002.12023540**.
- Ethical OS Toolkit: **https://ethicalos.org/**.
- Ethics Canvas: **https://ethicscanvas.org/**.

Team Management

Cultivating the right dispositions among members of an organization is primarily a matter of organizing day-to-day work in a team setting. How teams are organized is therefore a pivotal issue.

In order to organize teams in an ethical manner, managers can take the following into account:

- To take tasks and practices in teams beyond a narrow technical focus and link them to people's ethical beliefs and aspirations. For instance, consider that an engineer is not just responsible for coding, but for coding to achieve certain positive impacts in society.
- To allow for participatory decision-making in teams. This means making sure everyone is heard when making important decisions, and giving everyone a say in steering the work done by the team.
- To channel ethical concerns that people have into procedures that turn these concerns into positive change. This includes allowing for internal whistleblowing when there is organizational failure.
- To give people the opportunities to excel in an ethical way. This means allowing for initiatives that go beyond narrow business interests and benefit society.

ACTIVITY 9-4
Discussing Professional Ethics

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

The RudiBrace team has decided that its members will declare an ethical oath that fits the work they do. Remember that primary concerns that they have are related to privacy and fairness.

With this information in mind, can you suggest two statements for the RudiBrace ethical oath?

Summary

In this lesson, you identified the requirements for building an ethical culture. By describing the correlation between organizational culture and ethics, examining organizational purpose and values, raising ethical awareness, and fostering an environment that rewards and builds professional ethics, you can be sure that your organization is well on the way to having an ethical culture.

At your workplace, is there a culture or value statement in place? How does it align with the information discussed in this lesson?

At your workplace, what suggestions might you make to foster an environment that values professional ethics?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

10 Developing Ethical Systems in Technology-Focused Organizations

Lesson Time: 1 hour, 45 minutes

Lesson Introduction

The formal design of an organization includes various systems, such as bylaws and policies, as well as monitoring, communications, and feedback mechanisms. These systems play a significant role in determining how an organization recognizes and responds to ethical questions. This lesson covers some of the systemic dimensions of an ethical organization. As this is the final lesson, it also draws together some of the previous points to conclude with a discussion of ethical leadership and innovation.

Lesson Objectives

In this lesson, you will:

- Describe how policy and compliance support ethical organizations.
- Describe how metrics and monitoring can contribute to an ethical culture.
- Describe how communicating and engaging with stakeholders contribute to an ethical culture.
- Describe ethical leadership.

TOPIC A

Policy and Compliance

A large facet of organizational design deals with the scaffolding that holds the organization "in place." In many instances, this scaffolding is composed of a series of policies and compliance guidelines. This topic describes a range of elements to help you use policy and compliance to support the goals of ethical organizations.

What Are Policy and Compliance?

In an organization, ***policy*** can refer to:

- The rules imposed by external authorities. Organizations and the people within them must choose how to interpret regulations, which regulations to contest, and which new regulations to advocate. These questions become especially important in the area of emerging technologies, where the regulatory environment is still developing.
- The internal legal design of the organization. Organizations must also decide on an internal legal structure that establishes functions, roles, and responsibilities; even organizations that are long-established must continually assess these features to ensure they continue to respond appropriately to current conditions.
- Rules governing conduct within that structure. Additionally, specific policies can be used to address particular ethical questions that pertain to emerging technologies. These include many items covered in this course, such as data privacy and security policies, fairness and non-discrimination policies, and policies covering algorithmic audits and impact assessments.

Compliance refers to the process of enforcing and abiding by policies. Many organizations have a compliance unit that monitors the organization's fidelity to external laws and standards. But in the more general sense, aspects of compliance concern everyone in an organization, as everyone is obligated to follow the rules that apply to them. (As we will address in this topic, however, noncompliance may be justified when the rules themselves are illegitimate.)

Why Do Policy and Compliance Matter?

To see why policy and compliance matter, consider the debate over "net neutrality," a debate over whether Internet service providers should be prohibited from privileging certain forms of traffic over others. Many technology-driven organizations have felt compelled to develop a position on this debate and to take steps to advocate or challenge different policy proposals.

Consider also OpenAI, an organization focused on AI research and development that combines nonprofit and for-profit business structures. This model seeks to combine the financing and efficiency of a business with the social accountability of a charity. It shows how the choice of organizational form can facilitate or frustrate ethical goals.

Finally, consider how specific policy tools can assist with resolving ethical challenges. The genetic testing company 23andMe's privacy policy seeks to balance customer interests in privacy with business and scientific interests by requiring customers to opt into any secondary use of their (anonymized) data.

Internal Structures and Systems

Organizations must decide on an internal legal structure that establishes functions, roles, and responsibilities; even organizations that are long-established must continually assess these features to ensure they continue to respond appropriately to current conditions. Although some of these decisions may concern technical questions of efficiency and functionality, many are also deeply

ethical. The way an organization is designed determines how power is distributed and regulated. Inequitable power distribution within organizations is one significant cause of ethical failures. While these failures are not specific to organizations dealing with emerging technologies, many recent ethics scandals at technology companies involve questions of organizational design in some way or another, such as the balance of power between leaders and rank-and-file employees at Amazon and the role of ethics advisory boards at Google and Facebook.

One essential component of organizational governance is having clear, consistent, and fair rules, in the form of ***standard operating procedures (SOPs)***. SOPs that are clear, consistent, and fair reduce inefficiencies, limit arbitrary decisions, and help to curb abuses of power. This contrasts with organizations run by a fiat of individuals, who make decisions secretly and capriciously, without clear reasons, and in ways that unjustifiably favor some interests over others. Analogies to software code are apt here. Much like a piece of software, an organization can be programmed to operate effectively on the basis of protocols that are thoughtfully chosen, transparent, and equitable. Hard cases and new situations will still require manual deliberation or refining the rules. But the SOPs themselves provide the baseline for further development.

Policy Tools for Emerging Technologies

Organizations that deal with emerging technology increasingly make use of several policy tools, including:

- Codes of conduct.
- Ethical sales and use policies.
- Ethics review boards.

Codes of conduct establish behavioral expectations for employees generally or employees occupying certain roles. Many people who work in emerging technology are already bound by professional codes of ethics established by membership bodies such as the Association for Computing Machinery (ACM). Organizational codes of conduct can help to fill gaps left by formal rules, especially in situations where conditions are highly variable or sensitive to individual judgment. But they should not be (as they often are) a substitute for thoughtfully designed systems and procedures.

Another set of tools worth emphasizing are policies on ethical sales and product use. Sales policies establish to whom an organization will and will not sell its products. For instance, it might determine that a product is only fit to be sold to military entities and not to individuals. It might determine that a product should not be sold to organizations with poor compliance records or questionable intentions. Similarly, acceptable use policies, such as Salesforce's Acceptable Use Policy, establish how technology products may be permissibly used. Acceptable use policies are sometimes called end user licensing agreements (EULAs). They may also be called Terms of Service policies. Such policies are especially relevant in cases where the product is an ongoing service that the vendor can continue to monitor. If the product is a platform, for instance, the vendor will need to establish guidelines for permissible usage of the platform and interaction with other users. If the product is a device, the vendor may restrict use to licensed operators or for certain purposes. Product policies can be owned by particular units within an organization or by cross-functional teams or committees.

Despite their importance, ethical sales and use policies can also be controversial. In some cases, restricting sales of a product from certain entities can be objectionably discriminatory (or perceived as such), and restricting how customers use a product can be objectionably paternalistic (or perceived as such). Ensuring these policies are successful requires carefully weighing the different interests involved and ensuring that the terms can be justified to all affected.

A further tool worth highlighting is the use of ethics review boards. An ***ethics review board*** is a standing body that adjudicates sensitive ethical issues, such as research on human subjects or products with risks of harm. Their advice can be binding or non-binding. They can be composed of internal organization representatives, external experts, or a combination of the two. How these boards are designed and staffed are key to their success. Facebook's Oversight Board is staffed by external experts and has the power to overrule company executives on content moderation positions. Though it has only recently begun operating, it is widely regarded as a more promising model than Google's AI ethics board [Advanced Technology External Advisory Council (ATEAC)],

which was dissolved immediately after its launch in 2019 after complaints about the ideological views of certain members.

Additional policy tools include:

- **Chief Ethics Officer**: A senior leadership position responsible for managing ethical issues throughout the organization.
- **Cross-functional ethics committees**: A group composed of representatives from different functions and convened to develop policies on cross-cutting ethical questions.
- **Ethics hotlines and complaint reporting**: Allows stakeholders to get answers to ethics policy questions and report potential misconduct.

Compliance, Enforcement, and Contestation

Policies can serve their purposes only when people comply with them. It is thus important for organizations to take steps to monitor and ensure that rules are followed. As controversies over law enforcement make plain, however, enforcing policies is a delicate matter, even within organizations. Justice requires that the enforcement of policies be conducted equitably, treating each subject as an equal; it also requires following due processes, which assure various rights to individuals accused of violations.

The flip side of compliance is ***contestation***. Stakeholders may not agree with all of the rules and decisions they are expected to follow, and providing channels for contesting and revising rules can acknowledge the value of productive disagreement. In some cases, the organization as a whole may be called on to contest external policies that it regards as unconscionable. These decisions should be taken carefully, but they are an integral part of the management of an ethical organization.

Internal Policy Design Process

When designing a code of ethics, a privacy policy, an ethical usage policy, or similar rules for the conduct and operation of an organization, the same basic set of steps can be useful.

1. Identify the need for the policy.
2. Identify the owner(s) of the policy.
3. Gather information about the policy, including existing templates, standards, and best practices.
4. Consult appropriate stakeholders to ensure that each party with a significant stake or expertise in the matter has fair opportunities for input.
5. Draft the policy and invite further comments.
6. Establish procedures to support the policy, such as integrating it with existing systems and establishing compliance mechanisms, such as training procedures.
7. Ensure that roles for implementation are carefully defined.
8. Create procedures for monitoring implementation, usage, and compliance, and for gathering feedback.
9. Establish a process for handling policy disputes fairly.
10. Approve the policy.
11. Publish the policy and determine best means of communication.
12. Implement the policy and monitor results.
13. Review and revise the policy at regular intervals.

ACTIVITY 10-1
Discussing Policy and Compliance

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. Outline the basic elements of an internal or external acceptable use policy for RudiBrace.

2. What constraints (if any) should RudiBrace place on how corporate customers use the RudiBrace product?

3. Why did you answer the way you did?

TOPIC B

Metrics and Monitoring

All organizations value and strive for success, but what is the best way to measure success? If success is defined as sales revenue, it's pretty easy to identify data that can indicate whether or not revenues have increased over a given time period. When success is defined as being an ethical organization, finding data to support the claim can be quiet difficult. This topic focuses on how you can take the ideas of metrics and monitoring, and apply them to ethical goals within the organization.

What Gets Measured Gets Done

Tech companies in particular bank heavily on metrics to steer their organizations. Compared to traditional organizations, tech companies tend to be less hierarchical, and product teams have more freedom to self-organize. Beyond pure tech organizations, this way of managing is widespread even in parts of traditional organizations that develop emerging technologies. The reason is that the majority of the knowledge about execution lies with engineering staff and other individual contributors.

Steering by metrics allows product teams to experiment, rather than follow a plan devised by senior managers who are likely less knowledgeable about an area of technology than they are. To align the efforts of product teams with the goals of the organization, tech companies rely heavily on metrics. Until relatively recently, product teams at Facebook working on the social media platform were incentivized to increase engagement on the site.

The platform used this metric to figure out what features to develop, and how to prioritize them. To incentivize teams through metrics, organizations can, for instance, tie financial rewards to that metric and promote staff that successfully increase the metric. "What gets measured gets done"—a quote often attributed to management consultant Peter Drucker—encapsulates why relying on metrics can be an effective way of steering an organization.

From Metrics to Monitoring

Yet, reliance on metrics has a dark underbelly. We have seen that making ethical decisions is a matter of becoming aware of a range of different considerations and balancing sometimes difficult tradeoffs. Working to maximize a single metric cuts out all of this complexity. A focus on metrics as the sole objective can create a culture of myopia and short-termism. Meeting metrics often does not translate in responsible decision-making. ***Goodhart's Law*** applies. According to this law any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes. As a result, the organization is unlikely to be successful in the longer term. When metrics become the primary objective, they can also distort the meanings of the work itself. The problem is that metrics are not neutral. No single metric however, well intended, will fully align with responsible decision-making.

This is also what Facebook found: As research has shown, maximizing engagement tends to promote posts that negatively impact people's well-being and drive political polarization. In response to this criticism, Facebook has replaced the engagement metric with the metric of meaningful interactions. One year into the transition to the new metric, one study has found that the results are not satisfying.

This does not mean that metrics have no role to play. Rather, it illustrates that organizations should not exclusively rely on metrics to ensure responsible decision making. Processes embedding ethical foresight and nourishing a strong professional ethos are important drivers as well. In particular, there are three ways in which metrics can support building an ethical organization:

- Measuring organizational culture.
- Monitoring use of foresight tools.
- Monitoring the impact on society.

Additional Reading

For more information, visit:

- Facebook engagement metric and the risk of negative effects: **https://dl.acm.org/**, and search for the research article named ***Social network activity and social well-being***.
- Facebook metric of meaningful interactions: **https://www.niemanlab.org/2019/03/one-year-in-facebooks-big-algorithm-change-has-spurred-an-angry-fox-news-dominated-and-very-engaged-news-feed/**.

Measuring Organizational Culture

Metrics can be helpful in assessing the quality of an organization's culture. Conducting an ethics culture survey is a popular method of assessing an organization's ethical culture. The survey typically consists of questions that ask employees to rate the extent to which their organization practices ethical decision-making, as well as questions to assess root causes of ethical and unethical decision-making.

While good culture surveys can provide nuanced insight into different drivers of ethical and unethical behavior, there are also limitations to a purely quantitative approach. A more comprehensive method of assessing organizational culture is to use a ***culture audit***. A culture audit is an examination of an organization's culture that typically includes an ethical culture survey, but also a review of organizational documents, and interviews with key managers and employees. A culture audit can help to identify areas where the culture needs to be changed to promote better ethical decision-making before ethical issues arise. The culture audit can also help to identify the cause of unethical behavior and help to develop a plan for changing the culture to promote ethical behavior.

Monitoring the Use of Foresight Tools

Metrics can also be useful to ensure that teams go through the right process to surface and address ethical issues. Earlier in the course, we introduced a range of ethical foresight tools. To ensure that these tools get used, organizations can track adoption and use by teams. You may also decide to track metrics like the number of potential ethical issues surfaced during product development, and the proportion that was addressed and mitigated before the product shipped. Reviews of these metrics can be built into regular project reviews, creating a space to talk about potential ethical issues and escalate any issues that the team cannot address by itself.

These discussions can create a culture of shared responsibility for product ethics. They also send a clear message throughout the organization that ethical product development is not a one-time thing. It is a continuous process and a mindset that should be built into the culture of your organization. Process metrics can support establishing this culture, enabling organizations to recognize and respond to ethical issues more effectively. More importantly, you will be able to identify ethical issues early and mitigate them before they become ethical failures.

Monitoring the Impact on Society

Metrics can also be useful to measure the impact of your products on society. This is an important way to check whether you are living up to your social responsibility as a company. There are currently several initiatives underway to measure and benchmark the ***Environmental, Social, and Governance (ESG)*** impact of organizations. ESG criteria are standards for organizational operations that socially conscious investors can use to screen potential investments.

For example, the Global Reporting Initiative (GRI) provides a set of guidelines for reporting on ESG activities. In the last few years, many organizations, including companies, governments, and civil society organizations, have started to report their impact using this framework. These publicly available reports provide a good starting point to measure whether your products are having a positive or negative social and environmental impact. You can also use these reports to develop KPIs that your organization can use to track its progress.

Another framework to explore is the UN's SDG Compass. The SDG Compass provides guidance for companies on how they can align their strategies, as well as measure and manage their contribution to the realization of the SDGs.

Additional Reading

For more information, visit:

- GRI guidelines: **https://www.globalreporting.org/**.
- SDG Compass: **https://sdgcompass.org/**.

ACTIVITY 10-2
Discussing Metrics and Monitoring

Scenario
For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. List two impacts on society that the organization should be conscious of. (There are no right or wrong answers.)

2. Come up with a creative metric to measure each of the impacts. (There are no right or wrong answers.)

3. For each metric, think about how single-minded pursuit of the metric might lead to negative side effects.

TOPIC C

Communication and Stakeholder Engagement

Organizational leadership cannot work in a vacuum and expect that all parties will be satisfied with the results of their efforts. In today's world of instant information exchange, it is even more important than ever to share information with those who are affected by the organization's practices and decisions. For that reason, we'll now explore ways to communicate and engage with stakeholders about ethical issues and other issues.

Why Do Communication and Engagement with Stakeholders Matter?

Successful business practice is intertwined with recognizing and engaging with stakeholder claims. Stakeholder engagement matters for different reasons. Engaging stakeholders can be warranted because such individuals have crucial information or expertise to contribute to company policy or product design. Engaging with stakeholders is also essential to generating and maintaining their trust, which may be essential for the success of a business. Employees who don't trust an organization are unlikely to stay long. Customers who don't trust a product won't buy it.

But there are also many circumstances in which engaging with stakeholders is morally required simply because they have a right to be informed or to have their voices heard. Keeping customers in the dark about a data breach or a product malfunction may violate their rights. Employees may have a right to participate in certain decisions that affect their lives. Even when certain stakeholders can't be said to have a right to participate, seeking their input can be an important sign of respect.

Identification of Relevant Stakeholders

Different stakeholders have different priorities, given different interests and varying degrees of power. Effectively engaging relevant stakeholders therefore requires identifying and understanding different perspectives. For example, shareholders have an interest in company shares doing well, while members of the community who reside near a company may have a strong interest that the company not pollute the environment. And employees might prioritize rewarding tasks, job security, or a safe work environment.

The first step in engaging with stakeholders effectively is identifying different stakeholders and understanding their various interests and expectations. If not all interests can be respected, organizations must decide which stakeholders' claims to prioritize, and how to justify these decisions. This is where the importance of *ethical reasoning*, which we encountered early in the course, comes in, and where effective communication with stakeholders is of the essence.

Different stakeholders might take priority in different contexts, and who counts as a stakeholder might also evolve over time.

Mechanisms for Engaging Stakeholders

Stakeholder interests and concerns are usually identified through data gathering and analysis. Companies may conduct surveys or focus groups with customers, suppliers, members of the community, civil society groups, or other stakeholders. Sometimes, sales figures, product usage, or marketing data will reflect relevant information. For example, user data in web and mobile applications may provide information about how users engage with a company's services. Social media is another source of information about stakeholders' interests.

Engaging stakeholders may take different forms. In addition to direct communication, stakeholders may be given a voice through voting and representation mechanisms, in surveys, hotlines,

designated working groups, town hall meetings, and deliberative juries. Stakeholder engagement can range from responding to media queries and customers' comments on Twitter to setting up a designated space to deliberate over stakeholder concerns, such as Facebook's Oversight Board. This independent board was set up to review Facebook's content moderation decisions.

Communication During and After a Crisis

Despite your best efforts, your organization may experience an ethics crisis at some point. Even well-managed organizations with the best intentions and absolute integrity may face unexpected, ethically challenging situations.

- A scandal in a leader's personal life may raise questions about their ability to lead.
- A data breach may call attention to privacy and security policies.
- A product may malfunction or be used in the commission of a crime.
- An employee may be caught making morally unacceptable remarks in or outside the workplace.

A common response by organizations is to close ranks, clamp down on information, and shield themselves from scrutiny. Some may engage in symbolic gestures—such as firing scapegoats or increasing charitable donations—to deflect responsibility.

While there are risks of communicating information before it has been verified and the full scope of the problem is clear, in most cases, those affected by a scandal have a right to be informed as soon as possible. This should include information on the source of the problem, efforts underway to contain and investigate it, plans for corrective, compensatory, or disciplinary action, and opportunities for affected individuals to voice additional questions or concerns. If the situation is ongoing, the organization should provide timely updates on the progress of the resolution.

ACTIVITY 10-3

Discussing Communication and Stakeholder Engagement

Scenario

For this activity, you can use the RudiBrace example you were introduced to earlier, or you can select a product you are working on in your own workplace. Sample responses are provided for RudiBrace.

1. How should the RudiBrace team engage with stakeholders before the product launch?

2. How should stakeholders be identified?

3. What considerations will the team need to weigh in cases of conflict?

4. How should the RudiBrace team engage with stakeholders after product launch?

5. How should stakeholders be identified?

6. What considerations will the team need to weigh in cases of conflict?

TOPIC D

Ethical Leadership

We've now reached the final part of the course, and come full circle. We began this course with questions about what constitutes ethical behavior—questions about what we ought to do. These are some of the most important questions faced by leaders. Leaders have to make hard choices. Throughout their careers, leaders of organizations, small and large, encounter ethical challenges. Sometimes the difference between right and wrong isn't obvious, sometimes it's fiercely debated, and often it involves conflicting values.

What Is Ethical Leadership?

There are many ways to define and understand ***ethical leadership***. At the broadest level, we might say that leadership—specifically *ethical* leadership—involves exercising influence to guide others in the ethical pursuit of aims.

Being a leader involves deliberating with others about ethical problems, and assessing ethical perspectives of various stakeholders, as we saw earlier. Ethical leadership also involves thinking through moral disagreements and ambiguities. The ethical reasoning skills we explored at the beginning of this course are essential to this.

Note: Ethical questions can be very broad, or they can be very specific.

Methods for Increasing Ethical Leadership

Ethical leadership in organizations can also be about implementing new organizational forms, systems, or policies. Many organizations are experimenting with new ways of blending profit and purpose by design, by adopting social missions or placing constraints on profit-seeking. Others are experimenting with methods of decentralizing authority, either by introducing democratic (not autocratic) governance measures into the workplace, starting businesses owned by employees themselves, or experimenting with flat structures that reduce reliance on traditional hierarchies.

Corporate Social Responsibility and Taking a Stance

Increasingly, those who work for powerful companies expect their organizations to speak out on issues like immigration, racial equality, and criminal justice. Many believe that corporations have a duty to use their tremendous wealth and power to give back to society, to challenge systems and policies of oppression, and to lend a hand where they can.

Common pitfalls result when companies take stances on issues for mainly performative reasons (recall the ideal of virtue signaling) or for marketing purposes. In other cases, companies may have good intentions but limited expertise in the particular area where they seek to intervene. Often, a stance is championed by certain stakeholders within an organization, such as employees or an outspoken executive, and may not fairly represent the views of all the relevant parties, who may have fewer opportunities to make their positions known. Another standing worry in discussions of corporate activism is that the voice of powerful actors like corporations may drown out the voices of the less powerful, such as experts, people with opposing viewpoints, and members of the affected communities themselves.

We might think that, when companies engage in activism, they should do so for the right reasons. If they are to engage with broader social issues, they should do so sincerely. They should prioritize issue areas where they have relevant expertise or connection. They should consider carefully how

different viewpoints within the organization are to be represented. And they should be mindful of the volume of their voice in public discussions.

Leadership through Standard Setting

One challenge that organizations might face in showing ethical leadership is that, in the existing business environment, being ethical might be costly. If one organization invests in ethics and raises its standards while other organizations do the opposite, this might lead to a reduction in market share for the ethical organization.

Ethical leadership can thrive when an entire industry moves in the same direction. Organizations can influence this by means of outreach and communication to the public about their efforts in ethics. Another way to have lasting influence on an industry as a whole is through engaging in ***standard setting***, or getting directly involved in the development of rules and guidelines for your industry.

Organizations can lead by participating in standard setting activities such as:

- Keeping close contact with national and international standard setting bodies, such as ISO or the IEEE. Getting involved in international standard setting often requires going through the national bodies first.
- Joining committees of independent technical experts that draft and vote on new standards in the area of ethics.
- Linking activities concerning ethics within the organization with standard-setting activities; encouraging members in the organization to join the standard-setting efforts.

As we saw in previous lessons, standards are important, authoritative tools to unify organizations around common ethical purposes and values.

Additional Reading

For more information about standard setting with the ISO, visit **https://www.iso.org/get-involved.html**.

ACTIVITY 10-4
Discussing Ethical Leadership

Scenario
Consider the following questions as you discuss the contents of this topic.

1. What is your personal definition of ethical leadership?

2. Do you think corporations that work with emerging technologies should try to influence public policy outcomes?

3. Can you think of two or three items discussed in this course that you want to focus on at your workplace?

4. How have your views on ethics in emerging-technology-focused organizations changed because of this course?

Summary

In this lesson, you identified tools and methods for developing ethical systems in technology-focused organizations. By implementing thoughtful and relevant policy and compliance measures, adopting metrics and monitoring practices that can help support an ethical culture, making communication and engagement with stakeholders a priority, and finding ways to strengthen ethical leadership qualities, you can help ensure that your organization is prepared to address ethical issues as they arise.

At your workplace, what suggestions might you make for adopting metrics and monitoring for ethical behaviors?

At your workplace, what suggestions might you make to strengthen the ethical leadership qualities of the organization?

Note: Check your CHOICE Course screen for opportunities to interact with your classmates, peers, and the larger CHOICE online community about the topics covered in this course or other topics you are interested in. From the Course screen you can also access available resources for a more continuous learning experience.

Course Follow-Up

Congratulations! You have completed the *Certified Ethical Emerging Technologist™ (CEET): Exam CET-110* course. You have successfully identified and mitigated ethical risks, by using ethical frameworks and other tools, to incorporate ethics into data-driven technologies such as AI, IoT, and data science. The knowledge and skills you have acquired also help you to prepare for taking the CertNexus Certification Exam CET-110: Certified Ethical Emerging Technologist.

What's Next?

This is the only course in this series.

If you would like to know more about the technical side of emerging technologies, you might want to look into the following Logical Operations and/or CertNexus courses:

- *Certified Artificial Intelligence (AI) Practitioner (Exam AIP-110)*
- *Certified Data Science Practitioner (CDSP) (Exam DSP-110)*
- *Certified Internet of Things (IoT) Practitioner (Exam ITP-110)*
- *Applied Data Science with Python and Jupyter*
- *Certified Internet of Things (IoT) Security Practitioner (Exam ITS-110)*

You are encouraged to explore ethics and emerging technologies further by actively participating in any of the social media forums set up by your instructor or training administrator through the **Social Media** tile on the CHOICE Course screen.

A Mapping Course Content to Exam CET-110: Certified Ethical Emerging Technologist Certification Objectives

Obtaining the CertNexus **Certified Ethical Emerging Technologist** certification requires candidates to pass Exam CET-110.

To assist you in your preparation for the exam, Logical Operations has provided a reference document that indicates where the exam objectives are covered in this course.

The exam-mapping document is available from the **Course** page on CHOICE. Log on to your CHOICE account, select the tile for this course, select the **Files** tile, and download and unzip the course files. The mapping reference will be in a subfolder named **Mappings**.

Best of luck in your exam preparation!

Solutions

ACTIVITY 1-1: Discussing What's at Stake

1. **Which of the cases discussed do you find most disturbing? Why?**

   A: Responses will vary, but should show that learners can support their opinions with reasons. For instance, if a learner decided that the Myanmar Facebook issue was the most disturbing, the reasoning might involve the widespread nature of social media platforms and the propensity of humans to accept much of what they see as truth.

2. **In hindsight, sometimes it's easy to identify what went wrong and what should have been done instead. Other times, the complexity of the issues makes it harder to judge what's right or wrong. Which case do you find most difficult to evaluate, and why?**

   A: Responses will vary, but should show that learners can support their opinions with reasons. For instance, if a learner decided that the CRISPR babies case was the hardest to evaluate, the reasoning might be based on the perceived benefits of genetic engineering and the potential harm of He's actions affecting widespread use of the technology.

ACTIVITY 1-2: Identifying Ethical Issues

1. **Think of a hypothetical personal or business situation where someone faced a difficult ethical choice. What made it difficult?**

   A: Responses will vary depending on the specific circumstances. For instance, if you found a wallet with ID and cash in it when you and your infant daughter were a few days away from being evicted because of owing rent, you could find it difficult to return the wallet with all the cash intact. Knowing that the cash could help you protect your daughter might make it harder to do the right thing.

2. **Try to think of a time when your judgment about what's right clashed with a law, policy, or social norm. How did you decide what to do?**

   A: Responses will vary depending on the specific circumstances. For example, in the late 1960s, many young men in the US were drafted into military service and sent to Vietnam. If you were eligible for the draft but did not think that the U.S. military involvement was warranted, you might have decided to "dodge" the draft and move out of the country. Or you could have tried to obtain a medical or student deferment, or even declared yourself a conscientious objector. The decision you made would be quite personal and probably encompass not only your beliefs about U.S. military action in the Far East, but also your beliefs about honor, patriotism, family, honesty, and many other intangibles.

3. **Why are ethical considerations so important when it comes to emerging technologies?**

   A: Responses will vary depending on the specific circumstances, but should include the idea of computer technology making decisions that affect humans, and where the accountability lies for erroneous decisions.

ACTIVITY 1-3: Discussing Ethical Decision-Making

1. **In the Portal case, what everyday design decision posed a serious ethical risk?**

   A: Portal works with deep learning, which uses large quantities of data to learn how to recognize faces. The dataset used was non-representative, including mostly white faces. This resulted in Portal expressing a racial bias against Black people.

2. **Have you been involved in any projects where the use of technology introduced ethical risks? How can you guard against these risks?**

   A: Responses will vary, but might include using ML and depending on historical data to gain insights on new product acceptance. Using data primarily from the past does not take into account the current societal climate or any technological changes that might help determine a more accurate assessment of product acceptance.

ACTIVITY 1-4: Identifying Causes of Ethical Failures

3. **Can you think of an example of a product using emerging technology that uses tech for good?**

   A: Responses will vary, but might include products or services from companies like Change.org, WhiteHat, or Bethnal Green Ventures.

Note: Some suggested phrases to search include "tech for good" or "tech startup benefit company."

ACTIVITY 2-1: Discussing Ethical Reasons

1. **Can you describe a situation when different ethical reasons came into play and conflicted with one another?**

   A: Responses will vary, but might include situations describing conflicts between rights, duties, or broader values. Responses may also describe conflicts between different types of reasons; for example, tensions between social, cultural, or religious values and rights.

2. **In your opinion, what are the values that are widely agreed on, and which are less so?**

   A: Responses will vary, but might include broad ideas like fairness as being more widely accepted, and particular interpretations of such ideas as being less widely accepted. Other examples may include fundamental rights, like the right to life, as being widely agreed on, and social and cultural conventions as being less widely agreed on.

ACTIVITY 2-2: Discussing Ethical Reasoning Stumbling Blocks

1. **How would you respond to a colleague who insists that ethics is "just a matter of personal opinion"? Try to think of this in the context of a business need.**

   A: The colleague's position is an expression of moral relativism, which assumes that ethics is subjective. Your response might acknowledge what is clearly true about moral relativism: that there are ethical disagreements on many important ethical risks, as well as emphasize the importance of staying clear of moral righteousness. Your response should also discuss to what extent you can assume that there are better and worse ethical positions in any given case. From the standpoint of when a business decision needs to be made, you might ask your colleague how she would justify the decision.

2. **How would you respond to a colleague who wonders whether more effective technology should make ethical reasoning obsolete?**

   A: Answers will vary, but might include the view that ethical reasoning will always be necessary, no matter how effective the technology is one creates. Answers might also lead to a discussion about the extent to which elements of ethical reasoning may be "built into" certain emerging technologies, and whether or not participants think this is a good idea.

ACTIVITY 2-3: Identifying Potential Risks the RudiBrace Team Might Face

1. **Are there any stakeholders whose interests, rights, or values the RudiBrace team might fail to consider? If so, whom?**

   A: RudiBrace is a product that creates ethical risks for the people using the wearable bracelet, and the people who make decisions based on the data that the device produces. People on the product team likely have little familiarity with the interests or value of either of these groups. This creates risks that they will get missed.

2. **Is there a risk that ethical issues might fall through the cracks due to the way the team is structured? If so, what might you suggest to reduce that risk?**

   A: It is typical for product teams not to have a member whose explicit responsibility is ethics, though some organizations are beginning to change that. Since it is nobody's primary responsibility to identify and address ethical issues, such issues might easily fall through the cracks. This is particularly the case given that everyone on the team will work under high pressure to meet their explicit goals and metrics.

ACTIVITY 2-5: Using Regulations, Standards, and Human Rights to Identify Ethical Risks

2. **Do you think the assessment list mentioned in the final paragraph is a regulation or a standard?**

   ○ Regulation

   ◉ Standard

3. **Why did you answer as you did?**

   A: The guidelines were developed by the High-Level Expert Group on AI. The European Commission appointed this group of experts to provide advice on its AI strategy. Members include representatives from academia, civil society, and industry. The overall work of the AI HLEG has been central to the development of the Commission's approach to AI. But the AI HLEG is not a legislative body, and therefore the document does not have the force of law. Because this tool was created by a group of experts, it is considered to be a standard.

4. **What are some practical learnings the RudiBrace team can take from this?**

   A: Responses will vary, but might be similar to the following: For human agency and oversight, the RudiBrace team should make sure that the solution they build allows for human oversight and enhances the agency of people rather than constraining it. For technical robustness and safety, the RudiBrace team needs to make sure that personal data collected is properly secured.

ACTIVITY 3-1: Applying Ethical Theories in Decision-Making

1. **What is one ethical concern that consequentialist reasoning might raise about RudiBrace?**

   A: Responses will vary, but might include: How do the benefits created by RudiBrace compare to the privacy risks it creates? Are there additional product benefits that we did not anticipate?

2. **What is one ethical concern that deontological reasoning might raise about RudiBrace?**

   A: Responses will vary, but might include: What limits do the rights of employees impose on the data the wearable device may collect? Is there a bigger audience for this product than we originally planned for?

3. **What is one ethical concern that virtue ethics might raise about RudiBrace?**

   A: Responses will vary, but might include: How might RudiBrace affect the social relationships that people have with each other at work? How might RudiBrace affect the character of its users? Will we be proud of the product when it is released?

ACTIVITY 3-2: Identifying the Best Framework for a Situation

1. **Which framework or frameworks are most suited to this situation?**

   A: EU Guidelines for Trustworthy AI or IEEE Ethically Aligned Design.

2. **Which framework or frameworks are most suited to this situation?**

   A: Beijing AI Principles.

3. **Which framework or frameworks are most suited to this situation?**

   A: Government AI Readiness Index.

4. **Which framework or frameworks are most suited to this situation?**

   A: IEEE Ethically Aligned Design.

ACTIVITY 3-3: Selecting Strategies to Resolve Ethical Risks

1. **Can you suggest an action to remedy one of the ethical risks the team identified for RudiBrace?**

   A: Responses will vary depending on the ethical risk you selected. A possible response for mitigating the security risk would be to apply full end-to-end encryption of personal data.

2. **Can you suggest a possible tradeoff for the action you identified?**

   A: Responses will vary depending on the ethical risk you identified in the previous step, but might include the need to balance requirements for securing private data with the need to secure the entire organizational network (since encrypting data can cause issues related to identifying cyberattacks).

3. **Can you suggest a strategy for resolving the tradeoff?**

   A: Responses will vary depending on the ethical risk and tradeoff you selected in the first two steps, but might include providing detailed information to users to support informed consent, in lieu of encrypting the data, in order to balance the interests of employees and managers.

ACTIVITY 3-4: Establishing an Ethical Decision-Making Process

1. **Which people or groups should be involved in setting company policy about these questions?**

   A: Responses might include people or entities such as customers, health professionals, non-governmental organizations (NGOs) focusing on privacy, an ethics board if it exists, the U.S. Food and Drug Administration (FDA) or equivalent governmental agency, and advocacy groups for patients with rare diseases.

2. **What are some ways that these companies could share or give up power to improve the legitimacy of these decisions?**

   A: Answers may vary. Companies might commit to external standards and monitoring processes, or they might set up independent ethics committees with final authority on these questions.

ACTIVITY 4-1: Discussing the No Harm Principle

1. **How safe do driverless cars need to be to satisfy the no-harm principle?**

   A: There are no right or wrong answers to this question. For driverless cars, absolute security is not the only possible threshold. Alternatively, we might require that driverless cars are at least as safe as conventional cars. Does it make an ethical difference whether humans impose risks on each other, or risks are imposed by driverless cars?

2. **Consider the RudiBrace example. Imagine that bad actors are trying to access the RudiBrace to harm the users. Discuss what types of bad actors might want to attack RudiBrace, and what reasons would motivate them.**

   A: There are no right or wrong answers to this question. A possible answer would be that cyber criminals might attack the RudiBrace to get access to user data in order to sell it on the black market (i.e., for financial gain).

ACTIVITY 4-2: Identifying Security Risks

1. **List at least one security risk associated with the product and the risk roles that your organization has with respect to these risks.**

   A: For RudiBrace, one salient security risk is that user data could be stolen. The RudiBrace team is a decision maker and beneficiary with respect to imposing this risk, while the employees whose data gets collected by RudiBrace are risk-exposed.

2. **Are there any groups with respect to these risks that are in problematic risk roles?**

   A: The RudiBrace team is a decision maker and beneficiary with respect to imposing this risk, and this combination of risk roles raises special concerns. The employees are exposed to the risk, but they may not benefit to the same extent.

ACTIVITY 4-3: Identifying Security Tradeoffs

1. **Consider at least one security risk associated with the product and ways of addressing the security risk.**

   A: For RudiBrace, one salient security risk is that user data could be stolen. One way of addressing this risk would be not to collect any identifiable information from users, so that leaked data cannot expose individuals.

2. **Which tradeoffs with privacy, accountability, fairness, or the environment do these strategies create?**

   A: The suggested tactic described in the previous step creates a tradeoff with accountability. Because no identifiable data is collected, management cannot use the data to hold individual employees to account.

ACTIVITY 4-4: Identifying Security Risk Mitigation Techniques

1. **For the RudiBrace or your own product, can you describe some baselines for normal system behavior that can be used to detect security problems?**

   A: Responses will vary, but may include bandwidth consumption, user access and behavior, processing speed, etc.

2. **For a security risk associated with the product, list some techniques that can help to mitigate the security risk.**

   A: For RudiBrace, one salient security risk is that user data could be stolen. The RudiBrace team can establish the baseline systems behavior, especially user access behavior; create a rapid response team to help clients address breaches; take measures to protect data both in storage and in transit; and use threat modeling and attack simulations to anticipate attack vectors.

ACTIVITY 5-1: Discussing Privacy Basics

1. **Have you ever experienced an invasion of privacy through an emerging technology? Did you take any measures in response?**

   A: Answers will vary, but might include invasions of privacy in the online sphere, through particular apps, through surveillance technologies, and through smart devices.

2. **Do you think privacy matters for its own sake, or because it protects other important interests?**

   A: Answers will vary, but might include the view that privacy matters for its own sake, regardless of other interests, because it's a self-standing right; or the view that privacy matters primarily because it's essential to people's ability to pursue other vital interests.

ACTIVITY 5-2: Identifying Privacy Risks

1. **List two privacy risks associated with the product and the risk role that Rudison Technologies (or your organization) has with respect to these risks.**

   A: RudiBrace is ripe with privacy risks. One risk is for what purpose data will be processed and analyzed. RudiBrace has the mission to help organizations become more productive by fostering collegiality and collaboration. This open-ended mission might lead the team to want to process and analyze the data RudiBrace is collecting in ways that employees may not understand or condone.

2. **Are there any groups with respect to these risks that are in problematic risk roles?**

   A: Responses will vary, but may include the fact that users of RudiBrace face privacy risks but are not themselves decision-makers about these risks, and that the companies that deploy RudiBrace are beneficiaries of the technology but do not face significant privacy risks of their own, because those risks are born by employees.

ACTIVITY 5-3: Discussing Privacy Tradeoffs

1. **What limits to privacy should we accept for the sake of national security, public health, and the quality and efficiency of services?**

   A: Answers will vary, but might include that privacy can be sacrificed in the case of cyberterrorist threats, or for essential technical infrastructure like email and Google Maps to function well.

2. **In what contexts do you think privacy overrides other values?**

   A: Answers will vary, but might include when information concerns very personal, intimate things such as health records.

ACTIVITY 5-4: Discussing Privacy Risk Mitigation

1. **Consider two privacy risks associated with the product and ways of mitigating these risks.**

   A: In the case of RudiBrace, one privacy risk is that the data will be processed and analyzed in ways that employees will not understand or condone.

2. **Which tradeoffs do these mitigation strategies create?**

   A: This problem cannot usefully be addressed by having users accept a data sharing agreement. Rather, RudiBrace needs to define clear purposes for collecting and processing data, minimize data collection and sharing in light of these purposes, and take all reasonable steps to protect the data collected.

ACTIVITY 6-1: Discussing Bias

1. **Can you brainstorm one additional example of bias from at least five of the types of bias discussed in this topic?**

   A: Responses will vary but might include: for implicit bias, prejudging drivers because of their age (very young drivers are too reckless; older drivers are too slow); for automation/complacency bias, trusting the automotive diagnostic machine over the opinion of the human mechanic; for sampling bias, the Facebook Portal example discussed earlier in the course; for confirmation/reinforcement bias, searching for news stories that reinforce your beliefs and views; for cultural bias, being unaware of differing communication styles and assuming values are shared among people of different cultures; for gender bias, the packaging and marketing of children's toys; and for ableism bias, the presence or absence of handicapped parking.

2. **Which of these forms of bias are most relevant in the context of emerging technologies? Why?**

   A: Responses will vary but might include sampling bias because of its connection to data, complacency bias because of the increasing use of automated systems, or cultural bias because of the dominance of certain countries in technological development.

ACTIVITY 6-2: Identifying Potential Sources of Bias

What are some potential sources of bias in the design and development of this product?

A: Valid responses might include, but are not limited to, sampling bias from the way the underlying data is collected, cultural bias from assumptions about workplace norms, and gender bias from failing to appreciate differences in the burdens that employees face.

ACTIVITY 6-3: Discussing Fairness Tradeoffs

1. **What argument would you make if you were in favor of boosting scores, based on different notions of fairness?**

   A: Answers will vary. For instance, if fairness implies assigning the greatest benefit to the least advantage, it makes sense to give employees with care duties a boost in their activity scores.

2. **What argument would you make if you were against boosting scores, based on different notions of fairness?**

   A: Answers will vary. For instance, if fairness implies having a level playing field that does not incorporate external factors, assigning such a boost would be unfair.

ACTIVITY 6-4: Mitigating Bias Risks

1. **What groups might be inadvertently excluded from the RudiBrace design?**

   A: Valid responses include but are not limited to: people with physical disabilities for whom wearing a bracelet is not an option, or people with flexible contracts who might be excluded on the basis of their employment status.

2. **How could these groups still be included in the design?**

   A: Responses will vary, but might include: People with disabilities could be included in the design by offering different wearable solutions (e.g., also headsets); people with flexible contracts could be included by offering them RudiBraces that rotate between employees.

ACTIVITY 7-1: Discussing Transparency and Explainability

1. **List some examples of data-driven systems that you use every day, which produce decisions that lack explainability.**

   A: Answers will vary, but possible system-made decisions might include recommendations made by a social network or decisions made by a car's onboard computer.

2. **Discuss whether and why the lack of explainability poses a problem.**

   A: Answers will vary. For the social network recommendations, if your request is casual, there is probably little need for transparency. For decisions made by an onboard computer, you might want to be able to understand why a decision was made to ensure the physical safety of those in the car.

ACTIVITY 7-2: Identifying Potential Sources of Transparency and Explainability Risks

1. **Can you list two transparency or explainability risks associated with the RudiBrace product?**

   A: Answers will vary. For Rudison Industries, the RudiBrace product will include a management dashboard to monitor employee engagement and activity across the firm. None of these concepts are directly observable. Rather, they need to be inferred based on the interaction data the wearable collects, based on a theory of what meaningful engagement and activity consists of. Transparency and explainability risks emerge if managers interpreting the dashboard have different notions of engagement and activity in their heads than what RudiBrace actually measures.

2. **Are there any groups with respect to these risks that are in problematic risk roles?**

   A: Answers will vary. For instance, RudiBrace may operationalize activity time as time that employees spend sitting in the break room or in meeting rooms, whereas managers might think about activity in terms of output created. If RudiBrace is not sufficiently transparent about how it measures activity, managers may inadvertently reward teams that spend a lot of time sitting around and socializing, rather than the teams that get the most work done.

ACTIVITY 7-3: Discussing Transparency and Explainability Tradeoffs

Discuss at least one possible way in which more transparency of the RudiBrace might conflict with privacy.

A: Answers will vary, but might include: More transparency would mean giving insight into individual factors that lead to the generation of a certain activity score. This might include, for instance, information about work times or whether employees arrive early or late at the office. Making this information transparent in the organization potentially conflicts with people's privacy.

ACTIVITY 7-4: Mitigating Transparency and Explainability Risks

2. **Is the explanation for the Bank Customer a local or a global interpretation?**

   A: The explanation for the Bank Customer is a local interpretation, because it explains the classification of the model for a particular observation.

3. **Is the explanation for the Loan Officer a local or a global interpretation?**

   A: The explanation for the Loan Officer is a local interpretation, because it explains the classification of the model for a particular observation.

4. **Optional: Is the explanation for the Data Scientist a local or a global interpretation?**

   A: The explanation for the Data Scientist is global, because it provides an explanation of the behavior of the model as a whole.

5. **In your opinion, which transparency and explainability needs does the explanation for the Bank Customer meet, and which needs does it not meet?**

   A: The Bank Customer gets relevant information to understand which variables contributed to the decision to deny the loan application, and how much of an influence they have. Moreover, it also lays out how these variables would need to change for this applicant to approve the loan. This goes some way towards helping customers to understand what they can do to improve the likelihood that they will get a loan. But there are also three shortcomings with this explanation: First, the meaning of the variables themselves is obscure. Typical customers won't know what, for instance, a "consolidated risk marker" is, and how they can influence it. Here, an additional layer of explanation is needed. Second, the explanation does not indicate all the ways in which a customer could become eligible for a loan. If they changed a factor not listed in the explanation, it is possible that the requirements listed would change. This limits the ability of customers to find the best way for them to become eligible for a loan. Third, the explanation provides no insight into the quality of the algorithm, measured in terms of accuracy or fairness.

6. **How about the explanation for the Loan Officer?**

   A: The Loan Officer gets information about whether or not similar applicants to the applicant under consideration repaid their loan. This meets the need of an explanation that can inform the final decision whether or not to grant the loan. The Loan Officer can see immediately how many similar customers are in the database and what their detailed characteristics are. This is useful because it gives the Loan Officer a basis grounded in data for making a judgment call. Do some of the dissimilarities tip the balance of the decision for this customer? Have things changed so that a factor that was important in the past is no longer relevant? The explanation also has weaknesses, however. First, seeing the detail behind the algorithm may trigger automation bias—the tendency to defer to an algorithm. Hence the explanation may make the decision worse, rather than better. Second, the explanation gives no indication of how important different features typically are for predicting default risk. Third, the explanation obscures a major limitation in the dataset, namely that it only covers people who are successful applicants.

ACTIVITY 8-1: Discussing Accountability Basics

1. **Can you describe a situation in your professional life where you observed a lack of accountability?**

   A: Answers will vary, but might refer to the kinds of accountability risks and failures discussed in this lesson, such as lack of clear documentation to describe decisions and processes or lack of oversight.

2. **At your workplace, what are some of the barriers to stronger accountability practices?**

   A: Answers will vary, but might refer to some of the mechanisms discussed in the next topic, such as black box processes or use of third-party data.

ACTIVITY 8-2: Identifying Potential Sources of Accountability Risks

1. **Which types of technical accountability risks do you think are most important?**

   A: Answers will vary, but may include any risk discussed in the section titled Technical Accountability Risks. Learners may motivate their answers by means of their personal experiences; for instance, when they work a lot with black-box AI models, or assessment of negative impact; for instance, by considering the possible harm that can be done through lack of documentation.

2. **Which types of organizational or regulatory accountability risks do you think are most important?**

   A: Answers will vary, but may include any risk discussed in the section titled Organizational and Regulatory Accountability Risks. Learners may motivate their answers by means of their personal experience; for instance, when they perceive a lack of accountable culture in their organization, or assessment of negative impact; for instance, by considering the possible harm that can be done through a lack of legal regulations for mitigating certain accountability risks.

3. **Which types of technical accountability risks do you think need to be mitigated with the greatest urgency?**

   A: Answers will vary, but may include any risk discussed in the section titled Technical Accountability Risks. Learners may motivate their answers by referencing actual, urgent cases in which certain risks manifest themselves. For instance, they can point at unintentional discrimination through predictive policing software as a way to prioritize black-box/opacity risks.

4. **Which types of organizational or regulatory accountability risks do you think need to be mitigated with the greatest urgency?**

   A: Answers will vary, but may include any risk discussed in the section titled Organizational and Regulatory Accountability Risks. Learners may motivate their answers by referencing actual, urgent cases in which certain risks manifest themselves. For instance, they can refer to voter manipulation on Facebook as a way to prioritize legal regulation of accountability of social networks.

ACTIVITY 8-3: Discussing Accountability Tradeoffs

1. **If autonomous weapons systems could be shown to significantly reduce civilian deaths in war, does this mean a potential lack of accountability shouldn't stop us from using the technology?**

   A: Answers will vary, but might include the view that minimizing innocent deaths is a weightier moral duty than preserving traditional accountability mechanisms, given the fundamental right to life; or the view that the risk of a lack of accountability should stop us from using autonomous weapons systems irrespective of the potential benefits of reducing civilian deaths. The discussion may focus on questions about how to weigh the right to life against risks involved in undermining accountability mechanisms.

2. **If self-driving cars could be shown to significantly reduce the number of accidents in domestic traffic, does this mean a potential lack of accountability shouldn't stop us from using the technology?**

   A: Answers will vary, but might include the view that self-driving cars should be used despite a potential lack of accountability in the event that something goes wrong, because governments have a duty to protect their citizens which outweighs the potential risk of a lack of accountability; or the view that the risk of a lack of accountability outweighs the potential benefits of reducing traffic accidents because functioning accountability mechanisms are essential to preserving vital functions of society.

ACTIVITY 8-4: Mitigating Accountability Risks

1.

	Chief Product Officer	Project Manager	Hardware Engineer	Data Scientist	User Experience Designer	Software Engineer
Create test plan and KPIs/KRIs						
Interview test subjects						
Record and analyze test data						
Monitor software performance						
Confirm device safety						

A:

	Chief Product Officer	Project Manager	Hardware Engineer	Data Scientist	User Experience Designer	Software Engineer
Create test plan and KPIs/KRIs	A	R	C	C	C	C
Interview test subjects	--	A	I	C	R	I
Record and analyze test data	--	A	C	R	C	C
Monitor software performance	--	A	I	C	I	R
Confirm device safety	--	A	R	I	C	I

3. **Are there other project team members who should be consulted or informed about any of these risks?**

   A: Responses will vary, but might include that an internal or external review board should be consulted, or public relations staff should be informed.

ACTIVITY 9-1: Discussing Ethical Organizations

1. **Can you think of examples of company culture influencing employee behavior in a positive manner?**

   A: Responses will vary, but might include companies that encourage their employees to perform acts of community service and other good works.

2. **Can you think of examples of company culture influencing employee behavior in a negative manner?**

   A: Responses will vary, but might include examples of companies that place profits before safety.

ACTIVITY 9-2: Discussing Organizational Purpose

1. **Which of these statements of organizational purpose do you think most reflects an ethical organizational culture?**

   ○ To enrich people's lives with programs and services that inform, educate, and entertain.

   ○ To make it easy to do business anywhere.

   ◉ Our deepest purpose as an organization is helping support the health, well-being, and healing of both people—customers, Team Members, and business organizations in general—and the planet.

2. **Why did you select that answer?**

   A: Responses will vary, but might include that it indicates concern for people and the planet.

3. **Which of these statements of organizational purpose do you think least reflects an ethical organizational culture?**

   ○ To enrich people's lives with programs and services that inform, educate, and entertain.

   ◉ To make it easy to do business anywhere.

   ○ Our deepest purpose as an organization is helping support the health, well-being, and healing of both people—customers, Team Members, and business organizations in general—and the planet.

4. **Why did you select that answer?**

   A: Responses will vary but might include that it focuses only on the ease of doing business without reference to any ethical values.

5. **(Optional) With what you know about the RudiBrace product, draft organizational purpose and value statements for Rudison Technologies.**

   A: Responses will vary, but might include: for the purpose statement, "RudiBrace aims to strengthen collegial bonds in a safe way, respectful of privacy and data integrity;" and for the values statement, "We strive to maximize safety and privacy; we strive to treat users fairly and to give them control over their data."

ACTIVITY 9-3: Identifying Drivers of Awareness

1. **List two or three ethical risks and/or business impacts that you want to drive awareness of.**

   A: The RudiBrace team might focus on raising awareness of privacy risks and consumer trust, because of the sensitive personal information that it is dealing with.

2. **Select two drivers of ethical awareness and explain how they can help improve awareness.**

   A: Two important drivers to achieve awareness are fostering diversity of thought and ethical foresight. We have already discussed how RudiBrace can use ethical foresight tools to surface ethical risks. To foster diversity of thought, the RudiBrace team can look to hire team members from diverse backgrounds and engage with stakeholders. In addition, they should regularly engage with representatives from key stakeholder groups, especially with employee representatives.

ACTIVITY 9-4: Discussing Professional Ethics

With this information in mind, can you suggest two statements for the RudiBrace ethical oath?

A: Responses will vary, but the following suggestions are based on the data science oath: "I will respect the privacy of my data subjects, for their problems are not disclosed to me that the world may know, so I will tread with care in matters of privacy and security." "I will remember that my data are not just numbers without meaning or context, but represent real people and situations and that my work may lead to unintended societal consequences, such as inequality, poverty, and disparities due to algorithmic bias."

ACTIVITY 10-1: Discussing Policy and Compliance

1. **Outline the basic elements of an internal or external acceptable use policy for RudiBrace.**

   A: Responses will vary, but might include that the product cannot be resold, that the product requires certification of those who operate it, and that the product cannot be used for non-business purposes.

2. **What constraints (if any) should RudiBrace place on how corporate customers use the RudiBrace product?**

   A: Responses will vary, but might include that the product can be used only to improve workplace collaboration, that the product cannot be used to discriminate among employees, and that the product cannot be used to infringe on the privacy or autonomy of employees.

3. **Why did you answer the way you did?**

   A: Responses will vary, but may include: the fact that Rudison does not want to be complicit in uses of its products that violate fundamental ethical principles; the responsible use of this product requires shared commitments between buyer and seller; the product contains significant risks of abuse, so it is important to make buyers aware of these risks.

ACTIVITY 10-2: Discussing Metrics and Monitoring

1. **List two impacts on society that the organization should be conscious of. (There are no right or wrong answers.)**

   A: Rudison Technologies' mission is to help organizations become more productive by fostering companionship. At a minimum, the organization should have a positive impact on society on companionship and efficiency.

2. **Come up with a creative metric to measure each of the impacts. (There are no right or wrong answers.)**

   A: To measure companionship, the organization may conduct surveys in organizations that use their technology, asking about the frequency and quality of interactions between employees since the introduction of RudiBrace, compared to the period before. To measure efficiency, Rudison Technologies can gather trend data on metrics like output per time unit.

3. **For each metric, think about how single-minded pursuit of the metric might lead to negative side effects.**

   A: The organization should monitor whether the product has unanticipated negative impacts on society, especially in dimensions where ethical risks were identified, such as privacy.

ACTIVITY 10-3: Discussing Communication and Stakeholder Engagement

1. **How should the RudiBrace team engage with stakeholders before the product launch?**

   A: Before launch, the team might identify Rudison's leadership, board, regulators, employees, shareholders, customers, users, suppliers, and third parties affected by product use.

2. **How should stakeholders be identified?**

   A: The team might consider tools such as surveys and focus groups, product data, social media, voting and representation mechanisms, hotlines and customer service platforms, working groups, town hall meetings, and deliberative juries.

3. **What considerations will the team need to weigh in cases of conflict?**

   A: The team might need to weigh competing demands of each of the relevant groups, such as customer interests in affordability and control, user interests in privacy and autonomy, and shareholder interests in profit.

4. **How should the RudiBrace team engage with stakeholders after product launch?**

   A: After launch, the team might engage with anyone affected by the product, by conducting surveys or focus groups with those using the product, suppliers, members of the community, or civil society groups who have an interest in RudiBrace's work.

5. **How should stakeholders be identified?**

   A: After launch, the team might consult user data in web and mobile applications, sales figures, product usage, or marketing data, as well as potential social media sources. (Also see previous question—discussion of Questions 4 and 5 will likely be related.)

6. **What considerations will the team need to weigh in cases of conflict?**

   A: After launch, the team might consider different stakeholders' interests and potential conflicts between them. For example, those who introduced the product idea might have different interests from those actually using the product. Responses may also discuss what tradeoffs encountered in previous lessons may reemerge in this context.

ACTIVITY 10-4: Discussing Ethical Leadership

1. **What is your personal definition of ethical leadership?**

   A: Responses will vary, but might include: My personal definition of ethical leadership is proactively integrating ethics in my everyday work. This means organizing ethics workshops with my team, being vigilant in spotting ethical risks in a timely manner, and establishing direct channels with leadership to communicate important ethical risks.

2. **Do you think corporations that work with emerging technologies should try to influence public policy outcomes?**

   A: Responses will vary, but might include: On certain topics, tech companies have such a significant influence (for instance, on the topic of misinformation) that they need to influence public policy for the better—for instance, to safeguard democracy. However, they should do so only in a democratic way, through stakeholder consultation and other democratic mechanisms.

3. **Can you think of two or three items discussed in this course that you want to focus on at your workplace?**

   A: Responses will vary, but might include: I will organize a value-sensitive design workshop with my team; we will have an in-depth discussion at work about the different kinds of bias our products might have; we will approach management and ask for a revision process of our organizational purpose and values statement.

4. **How have your views on ethics in emerging-technology-focused organizations changed because of this course?**

   A: Responses will vary, but might include: I now understand that our organization is about more than just making money and profit. We have certain ethical responsibilities when it comes to the products and services that we make, to make sure that people's privacy is respected, that our processes are transparent, and so forth.

Glossary

accountability
The practice of holding agents responsible for outcomes, processes, actions, or intentions.

accountability gap
A condition that can occur when decisions are made by AI, because it is often unclear what or whom to hold responsible for things that go wrong.

AI
(artificial intelligence) The capacity of machines to exhibit human-like intelligence.

ambient intelligence
An electronic environment that is sensitive and responsive to a person's presence by using sensors, pervasive computing, and AI.

anonymization
The process of permanently removing data points that might be used to identify a subject.

applied ethics
The element of moral philosophy that deals with concrete issues that pose ethical challenges.

artificial neural network
An ML architecture that is modeled on how animal brains work.

availability
The component of data security that ensures that authorized users can access the data whenever they need to do so.

bias
Prejudice in favor of or against a person, group, or thing as compared with another, usually in an unfair way.

black box problem
An explainability issue that arises from using ML models, in which the decision-making process between inputs and outputs is not visible or understandable to human beings.

CIA
(confidentiality, integrity, availability) An information security model that focuses on data security.

cognitive bias
A systematic error in processing and interpreting information.

cognitive security
The use of artificial intelligence to detect security threats. It implements big data analytics to find connections and vulnerabilities in systems that are very difficult for humans to detect.

commercial secrets
An explainability issue that arises from using ML models, in which an organization limits the information shared about an ML project. Often referred to as trade secrets.

compliance
The process of enforcing and abiding by policies.

confidentiality
The component of data security that ensures that only authorized users and processes should be able to access or modify data.

consequence scanning
A tool for identifying ethical risks in a product development lifecycle that often uses key ethical considerations to categorize possible issues and their effects on the product.

consequentialism
An ethical theory that maintains that the rightness or wrongness of actions depends exclusively on the goodness or badness of consequences they bring about.

consequentialist
A person or characteristic that adheres to the ethical theory of consequentialism, where the rightness or wrongness of an action depends entirely on the consequences of that action.

contestation
The act or process of disputing policy and compliance rules.

contextual integrity
A theory that states that privacy requirements have the tendency to change according to different situations.

corporate culture
See *organizational culture*.

cost-benefit reasoning
A form of deduction that applies the principle that a decision or action should provide a greater benefit than the cost required to implement the decision or action.

CRISPR
(Clustered Regularly Interspaced Short Palindromic Repeats) A family of DNA sequences found in the genomes of organisms, such as bacteria, that can help combat viral infections and other disorders.

CRISPR technology
A tool for editing genomes whose potential applications include correcting genetic defects, preventing the spread of diseases, and improving crops.

culture audit
An examination of an organization's culture that typically includes an ethical culture survey, but also a review of organizational documents, and interviews with key managers and employees.

de-identification
Any process to prevent someone's personal identity from being revealed.

deep learning
A form of ML that uses layers of information of increasing complexity to make decisions, which constitutes a training model.

Delphi method
A systematic, interactive forecasting tool that gathers the opinions of multiple individuals without necessarily bringing them together face to face. A group of experts reply to a series of questionnaires, with responses shared among all parties. The questionnaires are revised based on the results of previous questionnaires until consensus is reached.

deontology
An ethical theory that states the right thing to do depends on complying with the right kinds of rules and principles, no matter the outcome of one's actions.

devil's advocate
Someone who expresses a contrary opinion to provoke debate or test the strength of an opposing argument.

differential privacy
A system for sharing information about a dataset by describing patterns within the

dataset and withholding information about the individuals whose data is in the dataset.

digital signature
A form of data protection that ensures that a digital message or document has not been tampered with or altered, whether intentionally or unintentionally.

distributive justice
A term used to describe the general problem of resolving conflicting interests.

diversity of thought
A situation where the people in a group bring varying, diverse viewpoints to the table.

duties
Moral debts or obligations that correlate to rights and are recognized by society.

emerging technologies
Technologies that are currently being developed or are expected to be available in the near future, and that may have significant economic and social impacts.

encryption
The protection of information from access by unauthorized people, often accomplished by encoding data with a masking algorithm so that it is not easily read by unauthorized persons.

end-to-end encryption
A form of data protection that decrypts and encrypts data messages at the sending and receiving points, no matter how many links are encountered along the way. Optimal for providing secure, private communication.

engineering activism
A movement that attempts to focus efforts on social justice issues throughout the engineering process.

ESG criteria
(environmental, social, and governance) Standards for organizational operations that socially conscious investors can use to screen potential investments.

ethical culture
The aspect of an organization's culture that supports the organization and its members in consistently doing the right thing.

ethical decision-making
The process of identifying ethical risks, developing options for action, and selecting an option that is supported by the best available reasons.

ethical leadership
The act of exercising influence to guide others in the ethical pursuit of aims.

ethical reasoning
The process of assessing considerations that speak in favor and against different courses of action, and weighing these considerations to form a judgment about what to do.

ethical reasons
Considerations that count in favor of, or against, a certain course of action.

ethical risk
A potential negative impact on people, society, or the environment.

ethical theories
Systematic efforts to understand moral concepts and justify moral principles.

ethical tradeoffs
Situations when alternative options for action are each supported by good reasons.

ethics
The philosophical study of the concepts of moral right and wrong (or moral good and bad).

ethics by design
An organizational approach to product development that focuses on addressing ethical risks early on, between the ideation and design phases of the product development lifecycle.

ethics review board
A standing body that adjudicates sensitive ethical issues, such as research on human subjects or products with risks of harm.

ethics washing
The practice of engaging in ethical reasoning without being motivated to do the right thing.

evaluation metric
A means of measuring the quality of a model.

expert rule
An explainability issue that arises from using ML models, in which complex data-driven systems require special knowledge to understand the decisions made during an ML project.

explainability
A system characteristic that enables a person to understand which inputs led the system to produce certain outputs of interest.

explainable AI
(artificial intelligence) A form of AI that does not contain a black box problem.

fair competition policy
An organizational policy that sets limits on behavior when competing for customers' business and when placing business with suppliers or offset partners.

fairness
The idea of giving each person his or her due and ensuring that any departures from equality can be justified.

fallacy
A general pattern in which a bad argument will fall.

GDPR
(General Data Protection Regulation) A pan-European data protection law that is widely adhered to by organizations around the world and that expands the rights of individuals to control how their personal data is collected and processed. It also places a range of obligations on organizations to be more accountable for data protection.

general AI
A type of artificial intelligence that is designed to be able to solve any problem.

global interpretation
An explanation of the extent that each input factor contributes to the outputs of an ML model.

Goodhart's Law
An adage that states that any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.

governance
The system of rules, practices, and processes by which a firm is directed and controlled.

ground truth dataset
Datasets that contain the ideal expected results that are used to evaluate the output of an artificial neural network.

groupthink
A form of bias that occurs when group members fail to voice opposing views.

human in the loop
A form of ML that requires a human to be involved in the decision-making process.

human rights
Rights accorded to human beings that create a protective zone around persons and allow them the opportunity to further their valued personal projects without interference from others.

human rights due diligence
A way for organizations to proactively manage the effect of any potential or actual adverse human rights impacts with which they are involved.

impact assessment
A technology design approach that focuses on assessing different options for action to resolve ethical risks.

inclusive design
A design process that aims to include as many potential users as possible, often focusing on groups that are at risk of being excluded or that have an unfair disadvantage in using a product.

informational privacy
The ability to control who has access to personal information, and to what extent.

informed consent
Consent given based upon a clear appreciation and understanding of the facts, implications, and consequences of an action.

integrity
The component of data security that ensures that data is maintained so that no improper modifications, whether accidental or malicious, can be made to it.

intellectual property
Any product of the human intellect that the law protects from unauthorized use by others.

interests
Ethical reasons that reflect concern for tangible benefits or intangible values.

interpretability
See ***explainability***.

ISO 27701
(International Standards Organization) A standard that provides guidance on how to collect personal data and prevent its unauthorized use or disclosure.

ISO/IEC 27000
(International Standards Organization/ International Electrotechnical Commission) The family of standards that provides best practice recommendations on information security management.

key ethical considerations
Concepts that tend to capture the most important areas of ethical concern in emerging technology.

link encryption
A form of data protection that decrypts and encrypts data messages at every link, node, or switch between the sender and the receiver.

local interpretation
An explanation of the extent that each input factor contributes to a particular prediction made by an ML model.

metaethics
The element of moral philosophy that deals with the nature of ethics itself.

ML
(machine learning) A process whereby large quantities of data are processed by algorithms to make intelligent decisions about the behavior of the software in a certain environment.

moral philosophy
The study of ethics.

moral relativism
The view that ethics is completely subjective, and that trying to identify the right thing to do is therefore pointless.

moral righteousness
The attitude that all who disagree with one's own ethical position are wrong.

move fast and break things
A development premise that indicates mistakes are a natural result of innovation in a complex competitive environment. The phrase is attributed to Mark Zuckerberg; it was the motto of Facebook from 2009 to 2014.

narrow AI
A type of artificial intelligence that is designed to solve particular problems and is the most common variety of AI currently in use.

normative ethics
The element of moral philosophy that deals with the basic elements of morality.

open data policy
An organizational policy that guides the sharing of data outside the organization.

organizational culture
A set of shared assumptions (values, principles, and practices) that guide individuals' actions and specify appropriate behavior in different organizational contexts.

organizational memory
The ability of organizations to record, share, and act on learnings from the past.

overfitting
A situation where a data model performs well with training data but not with new data.

overgeneralization
The tendency to make assumptions about a group of individuals or a broad range of cases, based on an inadequate sample.

paternalism
A situation where decisions about a person's own interests are made by someone else.

personal information
Any information that can directly or indirectly be linked to someone's identity.

PESTL
(political, economic, social, technological, and legal) A framework for comprehensively identifying the effects of fairness and bias risks.

pluralism
The view that there can be conflicting moral views that are each equally worthy of respect.

policy
The rules imposed by external authorities, the internal legal design of the organization, and rules governing conduct within that structure.

privacy
Freedom from observation or intrusion of someone's personal life by others.

privacy by design
An organizational approach to product development that focuses on incorporating privacy considerations into every stage of the development process.

problem of many hands
A condition that can occur when multiple individuals contribute to an outcome and it becomes difficult to determine who is responsible for what.

products
Commercial and noncommercial applications of an emerging technology.

pseudo-anonymization
The process of disguising data that might identify a subject.

RACI matrix
(responsible-accountable-consulted-informed) A tool for promoting accountability when assigning project roles.

radioactive data tracing
A process for identifying whether or not a particular dataset was used to train an ML model.

RAM
(responsibility assignment matrix) See ***RACI matrix***.

regulation
A set of rules made by a sovereign legislative body, often in consultation with subject matter experts.

rights
Entitlements to act or to be treated in certain ways.

risk
The likelihood of experiencing any negative outcome.

risk pattern library
A database that includes reusable use cases that outline threats, weaknesses, and countermeasures.

risk roles
The possible positions of different stakeholders for a particular risk, namely risk-exposed, beneficiary, and decision-maker.

scenario analysis
A tool for identifying ethical risks in a product development lifecycle by asking "what if" questions about the product.

security fetishism
The perception that security is paramount in value, often leading to the failure to consider any other value as being nearly as important as security.

self-learning model
An ML architecture that automatically integrates new data into the model and adjusts outputs according to the additional data.

slippery slope
A type of argument that claims an otherwise permissible action would lead to a chain reaction with ultimately catastrophic consequences.

smart contract
A self-executing contract with the terms of the agreement between buyer and seller being written in code.

SOP
(standard operating procedure) An organizational method intended to be routinely followed to perform specified actions or in specified situations. SOPs should be clear, consistent, and fair.

SSH
(Secure Shell) An encryption protocol that protects data in transit, particularly traffic related to remote access.

SSL TLS
(Secure Sockets Layer, Transport Layer Security) An encryption protocol that protects data in transit, particularly traffic related to web services.

stakeholder prompts
Tools for identifying ethical risks through stakeholder engagement.

standard
A rule or guideline created by industry and civil society organizations to establish common norms for interoperability, product quality, and professional conduct.

standard setting
The practice of developing industry-wide rules and guidelines.

statistical bias
A situation where a measurement differs systematically from the parameter that it is supposed to estimate.

statistical calibration
The process of adjusting the values of the parameters of a model to ensure the model will output data that, for a given set of input data, matches as closely as possible data found empirically.

STEEPV
(social, technological, economic, environmental, political, values) A framework for comprehensively identifying the effects of fairness and bias risks.

super AI
A potential type of artificial intelligence that is foretasted to be designed to be able to outsmart humans.

synthetic data
Data that is generated by a computer simulation and that approximates personally identifiable data.

technical solutionism
The assumption that technology alone suffices to solve any ethical issue.

technocracy
Rule by experts, in contrast to democracy, which is rule by the many.

technological determinism
The view that technology-based progress cannot be controlled.

Toronto Declaration
A statement on human rights standards for machine learning that focuses on avoiding discrimination.

tradeoff
The balance achieved between multiple desirable outcomes that are not mutually compatible.

transparency
A way of operating that makes it easy for others to see what actions are being performed.

ultimate attribution error
The tendency to attribute others' negative behavior to bad intentions while excusing our

own bad behavior as a result of the circumstances.

value-sensitive design
An implementation of ethics by design (often used in technology design) that accounts for human values in an ethical and systematic manner.

values
Ethical reasons that reflect concern for intangible benefits, or what people find desirable.

VCIO model
(values, criteria, indicators, observable behaviors) A way for organizations to ensure that their values translate into observable actions.

virtue ethics
An ethical theory that emphasizes virtues of mind, character, and sense of honesty.

visual contract
A binding legal contract without complex legal jargon.

vulnerability scoring
A way to assess the severity of a potential threat or weakness.

zero-knowledge protocols
Encryption protocols that enable you to prove you have information about something without revealing the information itself.

Index

D

E

F

G

H

I

R

S

T

U

V

095029S rev 1.0
ISBN-13 978-1-4246-4089-8
ISBN-10 1-4246-4089-X