Foresight into AI Ethics in Healthcare (FAIE-H):

A toolkit for creating an ethics roadmap for your healthcare AI project

Version 1.0

January 2020

Table of Contents

Introduction

Unlocking the power of data using algorithms and intelligent systems has the potential to help tackle some of the world's biggest challenges. Most of us set out to launch AI projects because we want to make a positive impact in the world. However, regardless of our intent, if we are not careful in making the incremental design and deployment decisions, our well-intentioned technology can fail or have serious negative societal and ethical consequences. Thinking through what those consequences could be can give you the foresight you need to avoid such failures and maximize the benefits of the technology for everyone.

We designed Foresight in AI Ethics (FAIE) as a toolkit to help you build an ethics roadmap that is tailored to your particular project. The process highlighted in toolkit is born out of our work in providing AI ethics consulting in 2017. It is a systematic way you can follow to uncover what key ethics issues are relevant to your project and strategize how to better anticipate, manage, and act to mitigate the ethics issues. This toolkit can introduce AI ethics foresight early in the design and deployment process, rather than serve as an auditing or evaluation tool.

Who the toolkit is for

This toolkit for anyone who is actively involved in the development and deployment of data-driven technologies for healthcare.
This includes data scientists, engineers, product managers, business leaders, or entrepreneurs looking to incorporate ethics into their AI project. It also includes people who have recently been assigned to assess the ethical integrity of a new AI system being developed.

Who we are

Open Roboethics Institute (ORI) is a nonprofit think tank in Canada. We specialize in studying social and ethical implications of robotics and AI technologies. In 2017, we launched an AI ethics consultancy, Generation R Consulting, with the dream of helping businesses address AI ethics issues they face today. In the course of our work, we developed a systematic process to create an AI ethics strategy for our clients. Generation R is now fully part of ORI, so that the lessons learned and research from our consulting work, such as FAIE, can be shared openly and for free.

A note before you begin

Many experts across the world are working to devise processes that can guide our design and deployment decisions related to AI projects. FAIE is inspired by these efforts. What is missing though, are examples of ethics challenges that are particular to the use cases and creative solutions to address them, so that the community can learn from and inspire each other.

In this healthcare edition of FAIE, the original toolkit has been modified to highlight a fictional case study in healthcare. We invite you to try FAIE and let us know what works and doesn't

work by submitting your own case study. These case studies will be made public for everyone to benefit from.

Why AI ethics?

When we account for the ethical dimensions of technology, especially during the design phase, we can shape people's reactions to and opinions of the technology. It addresses the need for good design, which helps to lower the risks inherent in AI projects. Sources of these risks include liability, stakeholder perceptions/reactions, employee/public perception and trust. Doing due diligence can help manage liability, anticipate and manage stakeholder and public perceptions/reactions. It allows for a smooth rollout and engender trust in the technology. This is in addition to medical ethics that gives us confidence in healthcare practitioners and the healthcare decisions they make. Just like how we need to be able to trust our doctors as patients, healthcare practitioners need to be able to trust the technologies they use.

We believe that it also helps businesses gain a competitive edge: as more and more people seek to work on projects that are aligned with their values, proactive stance on ethics by businesses can help attract and retain talent for tech companies; since governments across the world are actively deliberating regulation of AI, anticipating future regulatory changes by enculturing shared social values early can help reduce the impact of regulatory risks; AI policies, processes, and products that are value-aligned will be better positioned to earn and retain consumer trust; lastly, ethics assessments introduce new innovation opportunities.

Has this process been used in real life?

Yes. We used the process to provide an assessment for an actual organization, Technical Safety BC (Canada), with great success. Take a look at the full report on our website.

What you can expect

There are three main phases to FAIE, and a total of eleven steps to follow. Throughout the process, it forces us to think about the following three things in a systematic manner:

- **People**: How the stakeholders are related to each other and the technology,
- **Values**: What values are important for each stakeholder groups and the society, and
- **Technology**: How the new technology is related to the people and their values, and what impact it will have on them.

At the end of this process, you will have a strategy that enables key stakeholders of the technology you are designing to become knowledgeable stewards of the technology. You will have a mapping of foreseeable ethics issues specific to your project in one of three categories of things (make design, business, or communication decisions) you can do to manage or address the identified set of issues. Due to the many different techniques and perspectives that are used to enrich the final outcome, we recommend an interdisciplinary team (e.g., data scientist, business leader, product manager, communications manager) to take on this project together. Unlike checklists or technical evaluation tools that can be used in a few minutes, this toolkit requires you and your team to take the time to explore, investigate, and think. In order to better demonstrate the process, we walk you through a fictional example of a hospital that is creating a new AI division to innovate their operations.

How to use this toolkit in a healthcare context

To help walk you through the eleven steps of the Foresight into Artificial Intelligence Ethics (FAIE) toolkit, we created a case for you to follow along with. In this fictional scenario, we will assume the role of third-party consultants who have been hired to create an artificial intelligence (AI) ethics roadmap for a made-up healthcare organization as they embark on an AI project.

A case study: Mapleville Hospital

Mapleville Hospital is a large teaching hospital located in downtown Toronto, Canada. Since the hospital was founded nearly 70 years ago, it has been a champion for patient safety, compassion, teamwork, integrity, and innovation in healthcare. The hospital is well-known for its Institute of Endocrinology & Metabolism, a global leader in the industry that is committed to providing the highest quality care to patients with diabetes, endocrine, and metabolic disorders.

Recently, senior leadership has identified an opportunity to implement a **clinical decision support system (CDSS)** as a means of supporting healthcare providers in the treatment of patients with type 1 diabetes. The CDSS, known as "Sugar Mate", uses patient healthcare data to recommend personalized treatment plans, interventions, and educational material.

In order to bring this initiative to life, Mapleville Hospital has partnered with a small AI and data analytics start-up company based out of Montreal, Canada. The start-up, GroupThink Inc., provides data analytic services and insights to a number of healthcare partners, including hospitals and outpatient clinics. Since their inception in 2018, GroupThink Inc.'s vision of providing customized data insights has evolved to incorporate more than just open data options. For the purpose of creating the CDSS, Mapleville Hospital has agreed to supply GroupThink Inc. with anonymized patient data derived from the hospital's electronic medical record (EMR). The hospital's Chief Information Officer (CIO) has hired a Lead Data Scientist to oversee the Digital Technology Team's design and implementation of Sugar Mate.

How to use this document

Throughout this document, you will find examples of the FAIE toolkit being applied to the Mapleville Hospital case in blue boxes. You can also find general tips in purple boxes throughout.

Phase 1: Identify your use case & stakeholders

In this beginning phase, your primary goal is to identify the scope of your AI ethics analysis and set the scene. The following two steps are involved:

1. Identify the primary use case
2. Identify stakeholders of the use case and select key stakeholders

Step 1. Identify the primary use case

Typically, people envision multiple uses for the same AI technology or dataset. However, each use of the same technology can have a unique set of challenges. Therefore, the first step of FAIE is to identify the primary use case where the technology applies. A **use case** describes how an outcome of a data-driven algorithm is intended to be used. Talk to the people designing the technology or leading the project (e.g. product manager, lead data scientist, UX designer) and create a list of use case (i.e. potential uses of the technology).

Decide which of these people will be your **reference stakeholder**. A reference stakeholder is the go-to person who should be available to work with you for the entirety of the FAIE process. We recommend selecting someone who has decision-making authority within the healthcare organization.

Ask them the following questions to identify the use cases.

- What are the most immediate and intended uses of the technology you are developing?
- Why is the technology being developed? What is the ultimate goal of the project?
- What is the intended impact of the technology in the organization and for the users? Who are the people that will likely be most impacted by the new technology?

Example: Use Cases for Mapleville Hospital

Use Case 1 (primary): The clinical decision support system (CDSS) algorithm will be used by endocrinologists to provide customized care for type 1 diabetic patients at Mapleville Hospital.

Use Case 2: The output of the CDSS algorithm will be used to inform patients of best practice treatment options to improve patient engagement, decision-making, and autonomy.

Use Case 3: The data science team at Mapleville Hospital will use the CDSS outputs to create clustered patient profiles, which will be disseminated to Mapleville's clinical research team for type 1 diabeties research.

In this case study, the CIO (our reference stakeholder) wants to take an incremental approach to the overall AI strategy. The CIO envisions that Use Case #1 is the foundation for extracting value from the CDSS algorithm.

Step 2. Identify stakeholder groups

Now that you and the reference stakeholder agree which use case you will be focusing on, let's identify who are the stakeholders of the technology within the primary use case.

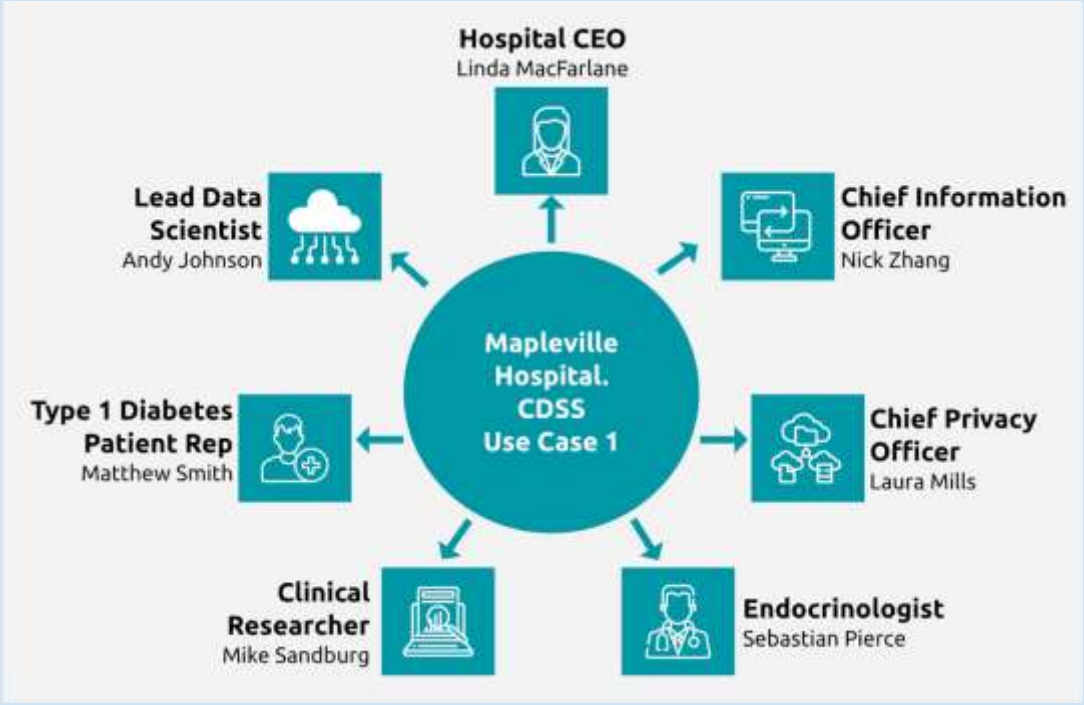
Stakeholders refer to people or organizations who are directly or indirectly impacted by the AI technology, or who are closely related to its design and deployment in some way (e.g., data scientists, other practitioners). List as many groups of **stakeholder groups** (the natural grouping of individual stakeholders who belong in the same category) as possible, as it will help you to understand the landscape of stakeholders relevant to the use case.

Prioritize the list into key stakeholder groups. The **key stakeholder groups** will be the ones from whom you draw in-depth understanding about people, values, and technology. Therefore, it is important to select stakeholders who are likely to give you the most diverse perspectives on the technology and its use. We recommend selecting at least three key stakeholder groups.

Identify one or more representatives from each key stakeholder group. If there are too many, then work with your reference stakeholder to narrow it down. She/he likely has an understanding of who should be consulted in this process and can introduce you to those individuals for Phase 2.

Example: Stakeholder Groups in Mapleville Hospital

In collaboration with the Chief Information Officer, we identified seven key stakeholder groups.



Phase 2: Discover ethics risks

Now that we know what use case and which set of stakeholders are the focus of our analysis, we can start to delve deeper into the analysis. **Your primary goal in Phase 2 is** to understand the people, organization, technology, and values involved in the use case. Here, we conduct an analysis of people's values, their relationships, and technology. This analysis will lead us to identify what kind of risks are applicable to the primary use case, such that we can address or manage them. The following steps are involved:

1. Listen to the key stakeholders
2. Map personas and identify values
3. Discover tensions in people's values
4. Discover tensions in people's activities
5. Discover tension in technology

Step 3. Listen to the key stakeholders

All technologies are put in a context where a shared set of values of the society exist. These set of values are what we call **societal values.** Here, we focus on the values of **transparency**, **trust**, **fairness & diversity**, **accountability**, **human rights (e.g. the right to privacy)**, and **human autonomy**.

In addition to the societal values, we need to understand how people and their values are related to each other, and what values are important for each stakeholder groups and the context of technology use. These are what we call **stakeholder values**. There are many different ways to take this step. The scope of each technique can be adjusted based on availability of resources and needs of the project.

Here, we highlight three different techniques (secondary research, interviews, and day-in-the-life) you can use. Use any combination of the three techniques to find your key stakeholders' values and learn about who they are.

Tip: Selecting Societal Values

You can also find a list of top societal values that could better apply to your project. There are many places where you can find hints of these values. For projects in Canada, the societal values would include human rights, gender equality, respect for the law, and diversity. You can also find these values by referring to the principles your company or your community ascribe to. Prof. Alan Winfield, a thought leader in the field, also has a handy list of existing AI ethics principles (e.g. Montreal Declaration) you can refer to. You can also use the core values used in Value-Sensitive Design processes: autonomy, community, cooperation, democratization, environmental sustainability, expression, fairness, human dignity, inclusivity and exclusivity, informed consent, justice ownership, privacy, self-efficacy, security, trust, and universal access. Learn more about Value-Sensitive Design here.

Secondary research

A lot of the information you are looking for may already be readily available. Save time and build on existing knowledge by looking for and reviewing the following:

- **Mission/vision/value statements**: these statements can help you understand the organization's expression of themselves and their goals/values. You can build on these during the interviews and analysis.
- **Organizational chart**: this chart can help you understand how various stakeholders are related to each other.
- **Corporate policies**: these existing policies that deal with user consent, use of and access to data, privacy, security practices, technology use, and human resource management are likely to provide information about what values are prioritized and what policies may be missing.

Interviews

Interviews are a great way to gain in-depth perspective into a particular stakeholder. They are often conducted in a private setting and allow the interviewee to voice their concerns and opinions. These interviews may be received differently depending on the company/team culture. Therefore, it is important to frame the interview as a constructive process, rather than a process of criticizing the technology or the organization.

Tip: 1-on-1 vs. Group Interviews

Group interviews are possible; however, make sure to consider the following:

- All the interviewees need to be comfortable to share their experiences with each other.
- It is important that the power dynamics in the group do not make any one of the people uncomfortable.
- It is important that the interviewer establishes norms/rules for the interview session so that people are aware of the expectations.

Schedule interviews with key stakeholders, making sure to interview at least one person from each key stakeholder group. Conduct the interviews with the following themes of questions. If you can, record the interview so that you can take detailed notes from it later. Stakeholders may find some of the questions easier to answer than others, and some people may answer the questions rather indirectly. Encourage them to share stories and anecdotes in these scenarios as they can be powerful tools for you to better understand their perspectives.

Interview Themes:

Social	**Value discovery**
We asked questions about the stakeholder's history with the organization, what their typical workday looks like, and what their role within the organization and	Building on what you now know about their role, ask them questions that can reveal values related to the use case and their engagement with the organization, and their understanding of the organization's values. Questions about their likes and dislikes related to their role typically helps to reveal their values. **Example questions:** • How would you describe a typical workday?

use case entails. This will give us an understanding of the social context within which the technology is to be deployed.	• What do you like best about your current role? • What do you like least about your current role? • How would the use of a CDSS impact your work experience? • What is the culture like at Mapleville regarding AI and decision support systems?

Functional
If the new AI system is meant to replace something (or someone) that exists already, ask questions to reveal the stakeholder's knowledge and use of the existing technology,
people, processes, and policies. This will give us an understanding about the stakeholder's relationship with existing technology solution and its relevant processes/policies.
	Value discovery
	Now that you understand their relationship to the existing technology or processes, ask them questions that can reveal their perspective on how and if societal values such as transparency, trust, fairness, accountability, human rights, human autonomy, and diversity are considered in the existing technology/processes. **Example questions:**
	• What types of policies exist within the organization regarding CDSS technologies, their required inputs, and the expected use of the data outputs?
	• What policies and procedures are in place to obtain patient consent?
	• How will the Digital Technology Team ensure that the CDSS algorithm is trustworthy and fair?
	• What risks and mitigations strategies has Mapleville identified to ensure that:
	  o Human rights are respected
	  o Clinician decision-making autonomy, consequences, and potential liability issues for recommending treatment interventions based on CDSS outputs are considered

Future Ask questions to better understand the stakeholders' knowledge and perception of the new technology and processes being developed. This will help us understand the stakeholder's relationship with and attitudes about the new technology project.	**Value discovery** Ask a series of ***what if*** questions that can reveal the stakeholder's perspective on how and if values such as privacy, trust, transparency, accountability, and fairness ***should be*** considered in the new technology. Questions about potential benefits and harms help reveal these perspectives. **Example questions:** • What would happen if societal values such as human rights and fairness were not considered during the design and implementation of the CDSS algorithm and the Sugar Mate application? • How would patients react to a privacy breach of their personal health data relating to this project? • What would happen if the clinical benefits and insights of the CDSS outputs are superior to clinician insights?

Day-in-the-life

Understanding a day in the life of one or two of the key stakeholders can be an effective way to understand where the technology fits into people's daily tasks and how the values of the stakeholder and organization come into effect when using a set of technologies. It should also help you understand why your stakeholder makes certain choices in using the technology, its outputs, and in providing data to the system.

A. Identify the one or two stakeholders whom you want to understand in depth

B. Meet with the stakeholder(s) and ask them to verbally walk you through their day, especially how they use the technology in question in their day-to-day work. Alternatively, you can shadow them for a part of their day and ask questions to learn more about their activities during that time.

Tip: Selecting Stakeholders for the Day in the Life Study

To select the stakeholders you'd like to study in depth, consider the candidates and answer the following:

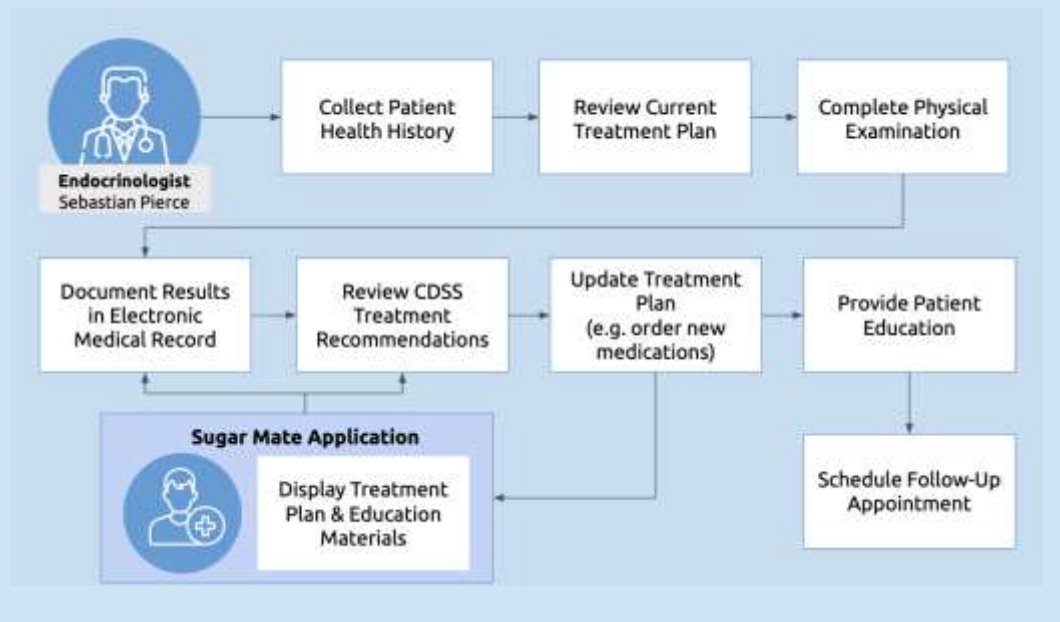
- Are they users of or directly affected by the AI system?
- Are they direct data providers for the AI system?
- Do they make decisions based on the outcome of the AI system?

If you answered yes to any of the above questions, then they are likely good candidates.

Example: A day in the life of an Endocrinologist at Mapleville Hospital

Other techniques

Group interview (a.k.a. focus groups), supplementary surveys and research (e.g., internal employee feedback on Glassdoor or online review/social channels), expert interviews and other more advanced ethnographic techniques could be used to understand people and their values. The techniques outlined above are simply those that we find to be the simplest to execute in short periods of time.



Step 4. Map personas and identify values

In this step, we use the information we gathered in Step 3 to develop **personas** for each key stakeholder using the template provided below. A **stakeholder persona** is used to organize the information you gathered about people through interviews, day-in-the-life, and secondary research. It not only highlights the details of what each stakeholder does under what constraints, but also helps to extract the values that are important for the stakeholders. Answer the questions in the stakeholder person template.

Once you've answered the questions, review your answers to identify which societal values are implicitly or explicitly highlighted by the stakeholder. Society values include **transparency**, **trust**, **fairness & diversity**, **accountability**, **human rights**, and **human autonomy.** Afterwards, identify what other values, apart from the societal values, seem to be important for the stakeholder. We call these **stakeholder values**. Stakeholder values can include a variety of things that are specific to the individual or groups of individuals. These include values such as effectiveness, human relationships, and job security. Take note of these stakeholder values in your persona table.

In the following stakeholder persona template, we have provided an example persona of an Endocrinologist practicing at Mapleville Hospital.

Template: Stakeholder Persona

Name	*Sebastian Pierce*, Endocrinologist	What is this person's primary goal?	*To diagnose and support patients living with endocrine disease and to help them manage their symptoms*
What do they like about their job and their interaction with the company?	• *Providing long-term care to patients is very rewarding* • *Challenges related to translating biochemistry, cell biology, and genetics into comprehensive patient care* • *Consistent research advances in the field of endocrinology*	What other stakeholders support them in achieving his/her goal?	• *Other healthcare practitioners (nurses, referring clinicians, laboratory staff, pharmacists)* • *Clerical staff* • *Patients & their families*
What do they dislike about their job and their interaction with the company?	• *Long working hours* • *Difficulty coordinating care with other health practitioners/organizations* • *Increasing amount of time spent documenting electronically as opposed to spending time connecting with patients and their families*	How do they use the technology to achieve their primary goal? If the technology has yet to launch, how are they planned to be used?	*The CDSS system pulls patients' health data from the EMR and uses it to determine a customized treatment plan and education materials that have been tailored to meet the needs of that specific patient. The endocrinologist uses the built in CDSS during each patient visit for improved treatment decision-making. At the point of care, the patient will be informed of the treatment options and will have the opportunity to engage in discussions about changes to their treatment plan.*
What are the company/group values?	*Mapleville Hospital's values are: Safety, Compassion, Teamwork, Integrity, and Innovation*	What policies, regulations, social norms or technical constraints do they need to work with?	• *Patient safety and consent* • *Privacy/security of information* • *Equitable practices* • *Hospital policy* • *Recommendations made by the CDSS are only as good as the quality/completeness of EMR data*
Stakeholder values based on the persona	*Patient-centred care, patient safety, efficiency, innovation*	Societal values extracted from the persona	*Privacy, transparency, fairness, accountability*

Step 5. Discover value tensions

Values are the glue that holds people and technology together. In the previous steps, we discussed a set of societal values: **transparency**, **trust**, **fairness & diversity**, **accountability**, **human rights**, and **human autonomy**, and identified stakeholder values.

Review the list of societal and stakeholder values identified in Step 4. You'll notice that there are some values that are in conflict with other values. These are what we call **value tensions**. These value tensions can exist between stakeholder value and societal value of the same person, or stakeholder values between different people. Sometimes, the same values may be in tension because of the different ways in which different stakeholders interpret the values. We provide two examples of value tensions here.

Identify as many value tensions as you can and take note of which values are in tension and how.

Value Tension #1

Value tension: Which values are in tension with one another?	*Efficiency vs. transparency*
Tension description: How are the values in tension with one another?	*The CDSS aims to provide the endocrinologist with evidence-based recommendations for treatment plans, which increases the efficiency with which the endocrinologist can treat patients. However, it is not always clear how the CDSS makes recommendations due to the "black box" nature of the system, which means that the endocrinologist must remain critical of the recommendations being made.*

Value Tension #2

Value tension: Which values are in tension with one another?	*Innovation vs. accountability and patient safety*
Tension description: How are the values in tension with one another?	*Mapleville Hospital generally embraces new technologies, especially when they produce benefits for both practitioners and patients. However, the hospital is ultimately accountable for patient safety, and, as such, they have to exercise caution when implementing new technologies that may affect patient care outcomes. At the same time, leadership at Mapleville Hospital do not want to stifle innovation and are open to piloting new projects.*

Step 6. Discover tensions with stakeholder persona

Take a look at each persona you developed in Step 4. In reviewing each stakeholder's goals, activities, and roles, and what constraints they work with, think about how these elements support or clash with the stakeholder and societal values. List the tensions that correspond to people's activities.

Example: Discovering Tensions with a Mapleville Stakeholder

Let's take a look at the persona of an Endocrinologist in the Mapleville Hospital case study as an example. Through developing the Endocrinologist persona, we found that the recommendations made by the CDSS are only as good as the quality/completeness of EMR data *(i.e. blood glucose, vitals, labs, medication orders, medical history, progress notes, etc.)*. These inputs are used by the CDSS in order to determine treatment plans for type 1 diabetic patients. If the EMR data is poor in quality or lacks completeness, this could result in inaccurate recommendations being made by the CDSS. This is in tension with the value of efficiency.

Step 7. Discover tensions in technology

We now need to understand how the technology is related to the people and our societal values. Our primary goal in this step is to **understand what the input** (e.g., what kind of training data) **and output** (e.g., a numeric score or a category) **of the technology are**, and **how they are related to the people** (e.g., who is providing the data? who is receiving the output?), and **societal values**.

A. Understand the technical components

Consult with the reference stakeholder or someone who knows the ins-and-outs of the technical components of the project, including the dataset being used for the project. Here, we want to understand what the main sources of data are, and to map the overall information flow of the technology you are analyzing. Find out the following sets of information to better understand the technology:

Input

Understand what the **input** of the technology is, and how it's related to the stakeholders. To do so, you need to ask and answer the following questions:

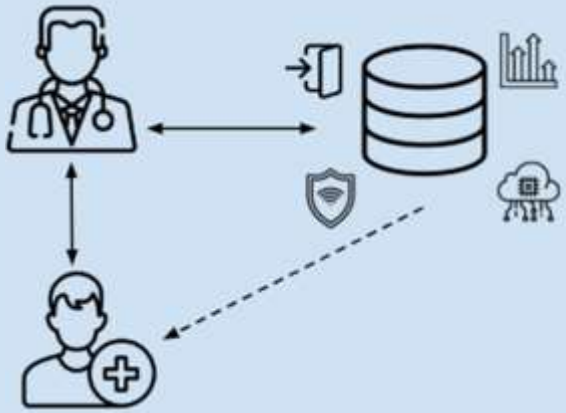
- Data Fields: What is the list of data fields being used (or that has the potential to be used) for the use case?
- Data Source: Who or what is providing the data?
- Input Surface: What interface, if any, do people use to provide the data?
- Format: In what format is the data stored (e.g. numerical, text, photo, date)?

- Nature of Data: What is the nature of the data? Could the data be used to identify people (i.e. personally identifiable information)? Does the data infer a judgement or outcome related to a user's or stakeholders' performance or other qualities?

If you have a large set of data fields, organize the above information into a table so that you can fill out the information for each data field identified. Add any descriptive notes to help you recall important details about the data later.

Example: Understanding the Input for Mapleville

GroupThink Inc. gathers data from multiple sources. However, for this particular use case, the data used for the project comes from Mapleville Hospital's EMR where patient health information (e.g., medication history, lab results) is being stored.



Model and the output

To understand how the input is processed and what the output is expected to be produced.

- What are some core pieces of the technology? Is it predictive? Does it use data-driven approaches? What are the parameters that are within the designers/technologists' control? Is it run in real-time? If it's an adaptive, machine-learning system, how often does it update the model and when? Does it use only historical or the most recent data, or does it use a hybrid (weighted) model?
- How is it connected to other technical systems or products? Does it have physical actuation capabilities (e.g., smart systems or robots)?

Post-output interaction

Understand the **output** of the technology (e.g., a numeric score or a category) and how it's related to the stakeholders. The project may be in the early stages and may be considering multiple options. In this case, jot down all possibilities being considered:

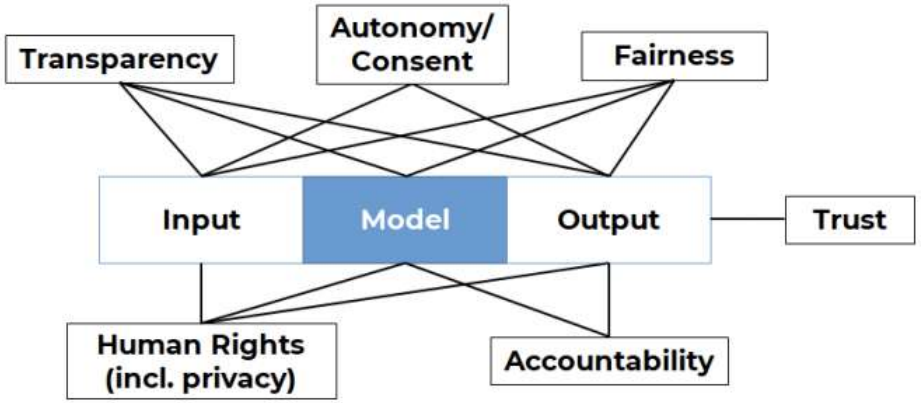
- Output: What is the main output?
- Output Source: What produces the output? This can be from the algorithm and any other relevant information pulled from a dataset.
- Output Interface: What interface is used to present the output to people?
- Format: In what format is the data presented (e.g., numerical, categorical)?
- Who has access: Who has access to the output produced?

B. Analyze the technology against values

We want to determine how the values are related to the input, model, and output of the system and evaluate where value tensions may arise. Not all values are equally important across the three parts of the technology. The figure below illustrates a mapping of which societal values are typically most relevant to input, model, and output of a data-driven technology respectively.[1,2]

Value Questions

Now, thinking about the key stakeholder groups and values related to the input part of the technology, answer the list of value questions that can help evaluate how well the value is served in the input process. A sample set is provided below. Feel free to add or modify the value questions as you see fit.[3] Answer the value questions considering each of the key stakeholders in mind.



Work with a technical stakeholder if you need further information to answer some of these questions. You will naturally start to see where there are ethical issues, or where there are actions that can be taken to mitigate a risk. Take a note of them, as you'll take a holistic look at these along with value tensions identified in previous steps. You might also be inspired to ask additional questions as you start to answer the existing questions. Add them to your list of value questions and continue until you've considered all stakeholders in answering all the value questions.

We provide our answers for the Mapleville Hospital example along with the value questions below. In our example, we use hashtags **#risks**, and **#need4action** to mark where the ethics risks are and where actions are needed to address challenges. Notice that, depending on how far along the design of the technology is, the value questions can be modified to ask about the plans for the future, rather than the present state.

---

[1] If you have chosen a different set of societal values, you can build a mapping similar to the one we have here.

[2] As you follow through the steps below, you might find the need to modify this mapping for your specific use case. Feel free to modify it. You can do so by considering how each societal value might be affected by different design, communication, or business decisions made about the input, model, and output of the technology. If the decisions have the potential to impact the value related to any of the three parts of the technology, map the value respectively.

[3] In modifying or adding new value questions, choose close-ended questions (i.e., those that lead to yes or no answers) over open-ended ones. This will help you pinpoint where ethics risks are, and what kind of solutions may be needed to address the risks.

Reference: Value questions

	Value Questions		Mapleville Hospital Answers
Input	Accountability	Is the team or person responsible for overseeing the data input processes for the new technology clearly identified?	*Yes. The lead data scientist is responsible for overseeing the CDSS system's build and implementation.*
		Is there a team or person responsible for cleaning input data?	*Yes. There are a team of data scientists who are in charge of ensuring the input data is standardized and appropriate for use in the CDSS.*
	Autonomy/ Consent	Is there an informed consent process in place for the data collection that outlines the fact that the data can be used for this use case?	*Yes. For new referral patients, consent is obtained at the time of pre-registration. The patient has the option to revoke their consent at any time in person, or via the patient-facing application.*
		Can the stakeholders decide not to have their data used in the algorithmic system?	*Yes. Patients have the opportunity to decline consent at any time. Moving forward, continuous algorithmic use of their data will be ceased.*
		Are there any elements in the data collection process (e.g. user interface used for inputting data) that could result in unintended outcomes?	*Yes, potentially. Since the CDSS output is reliant on clinical EMR data, algorithmic training will be highly dependent on data accuracy. #risks*
	Fairness	If people are involved in directly collecting data from someone/something, how diverse are these people in terms of race, gender, age, class, and other socioeconomic factors? Teams of people who are similar to one another can lead to similarly biased observation and data entries.	*The lead data scientist will be using any patient's data (from the EMR) provided that consent was obtained. Healthcare clinicians are responsible for inputting the EMR data, which should contain mostly objective information. This will likely not be an issue.*
		Are certain groups of stakeholders' information collected disproportionately more than others? If so, does this fact support or conflict with the societal and stakeholder value?	*The information collected for the CDSS algorithm is dependent on the EMR data, which contains a high proportion of affluent individuals. This may conflict with stakeholder values. #risks*
	Human Rights (incl. privacy)	Does the input data include sensitive/identifying information (e.g., gender, race/ethnicity, religion, location of work/residence, education, social and professional associations/groups)?	*Yes. #risks*

		Can the stakeholders opt not to enter the sensitive/identifying information?	*No. The algorithm is dependent on specific input parameters and variables. Removing the data would impact algorithm output quality. #risks*
	Transparency	Do the relevant stakeholders know how/when the information is collected/changed/used?	*Yes. Both patients and clinicians using the input and outputs of the CDSS are aware.*
		Are the data provided by the stakeholders used to collect any secondary sources of information (e.g. connected to social media profiles, external online platforms)? If so, are the stakeholders informed of this?	*Yes. The information collected from the patients will be used to inform future research within Mapleville Hospital's research department. This information will not be shared with external sources as per custodianship ownerships under the PHIPA laws.*
	Trust	How well do those who will be using the output understand the new technology?	*The endocrinologists will receive general training on how the CDSS works. However, they are not experts in data science. #need4action*
		What kind of decisions or reactions will the results be used to inform?	*The outputs will inform patient treatment plans, as well as suggest customized education materials designed to meet each patients' respective needs.*
		What are the stakeholder groups' feelings around trusting the output of the new technology?	*The main endocrinologist has embraced the idea of using CDSS technology; however, there have been some questions from patients and their families who want to ensure that their plans are reviewed by a physician prior to being ordered. #need4action*
Model	Accountability	How often is the model updated/re-trained and is the frequency adequate for the use case?	*Yes. The model is retrained every night. The model will ensure to check updated/revoked consents.*
		Who oversees the model training/updating process and are they the right people who can detect new problems and act upon them?	*The lead data scientist oversees this process. They are the most responsible person at the organization to oversee the model training/ updating processes.*
	Autonomy/ Consent	Are there concerns that the model could infringe upon any of the stakeholder groups' autonomy or decision-making control?	*Yes. One endocrinologist has been vocal about their concerns that the CDSS system will override their authority as a physician. #risks*

	Fairness	Are there sources of bias that could lead to unfairly discriminate against individuals/groups, especially against specific gender, race/ethnicity, religion, social class or otherwise marginalized group?	*Yes. Mapleville Hospital is located in a densely populated area in downtown Toronto. Patients at this hospital tend to be of higher socio-economic status. Using patient health data may discriminate against those of lower/middle socio-economic status. #risks*
		Are there any parameters or technical aspects of the system that can contribute to biases in the output against specific gender, race/ethnicity, religion, social class or otherwise marginalized groups?	*No. All information obtained from the EMR is parsed the same.*
	Human Rights (incl. privacy)	Is the model designed to reveal or predict an individual's identity (e.g., sexual orientation), potential (e.g., a child's probability of success in life), such that it contradicts with stakeholder and societal values, including human rights?	*No.*
	Transparency	Is the model and its performance understandable to and monitored by those training it?	*Yes, by the lead data scientist. However, the model will require ongoing monitoring, which will need to be delegated to someone on the Digital Technology Team. #need4action*
		If there is a questionable/erroneous outcome or an incident in the future, is it possible to explain to a third party what aspects of the model led to the outcome/incident?	*Yes, the model output is explainable through a supervised algorithm.*
	Trust	Have the developers been briefed in the range of potential implications this model could have on decision making?	*Yes. The developers were part of a working group that explored the different implications that could result from the CDSS output, which they factored into building the model.*
Output	Accountability	Who is responsible for acting on the output, and does this stakeholder group have ways to remedy or override erroneous or questionable output?	*Endocrinologists are the only ones responsible for acting on the output of the CDSS model. Ultimately, the endocrinologist can override the CDSS recommendations prior to updating the care plan.*

		Is there a communicated and unobstructed means for different stakeholder groups to raise an alarm on possibly dangerous use of the technology?	*No. But this is a topic Mapleville should explore in future, maybe through a steering committee. #need4action*
		For cases where sensitive findings arise from the outcome, is there a clear means for different stakeholder groups to deal with the potentially uncomfortable truths (burden of knowledge)?	*No. #need4action*
		What are the implications of false positives? What are the implications of false negatives? Are the appropriate decision makers aware of the balancing of risks between the two?	*The endocrinologist will still be responsible for the final treatment decisions transcribed to the patient's care plan. If the endocrinologist is uncertain with the output, he/she will require to seek further information for clarification.*
	Autonomy/ Consent	Is the output connected to another process or technology without human intervention being necessary? If so, are the risks from worst case scenarios minimal and acceptable?	*No. The output is always displayed for the endocrinologist to review, interpret, and accept/decline the recommendations.*
		Is the technology designed to replace or assist human decisions? If it is meant to replace them, is it meant to support the overall function of the stakeholders whose decisions are being replaced?	*The CDSS outputs are meant to assist in the endocrinologists' decision-making, not replace them.*
	Fairness	Are the primary users of the technology aware of the potential biases that may have contributed to the output?	*Yes.*
		Are the stakeholders who are subjected to the technology given a means of remedy?	*Not yet. #need4action*
		Does the output produce the same result for all users? Does it lead to unfairness or discrimination?	*It produces the same results for any endocrinologist.*
		Does the output lead to fair distribution of wealth, opportunity, or other positive outcomes?	*It is hard to determine this yet, since it is in the early stages of technical implementation. #risks*

	Human Rights (incl. privacy)	Does the technology suppress or protect fundamental human rights, such as the right to life, liberty, security, freedom of movement and of expression, among others?	*No.*
	Transparency	Is the output from the algorithm presented in such a way that is understandable to its audience?	*Yes. The CDSS output is presented within the EMR as an embedded graphical user interface (GUI).*
		Is the output from the algorithm presented to the stakeholders in a way that allows them to understand how/why the system has produced the specific output? Is this important for them to understand?	*Not yet. #need4action*
	Trust	Is the output from the algorithm translated from a probability score to a categorization (e.g., 90% probability of being X is presented as being X)? Is the translation of the probability to categorization appropriate for the use case and trustworthy?	*Yes.*
		Does the technology and its output have the potential to lead to a destructive cycle of behaviours or operations (e.g. reinforcing gender bias of those who are the primary source of input data)?	*Yes. The endocrinologists may become extremely dependant on the algorithmic outputs to inform decision-making. #risks*
		If someone were to take the outputs from the system and generalise it to the other use cases, is it reasonable to foresee problematic interpretations or increase in distrust among stakeholders?	*Yes, the output interpretation will be different for each stakeholder group.*

Step 8. Synthesize into ethics challenges

Take a look at the value tensions, risks and needs for action you've noted from previous steps. Taken together, you'll be able to see how some of these items can naturally be grouped together. Organise them thematically as much as you can. Depending on your use case, you might find it easiest to organise them by the values that are most relevant to the tensions, risks, and needs. Afterwards, prioritise them so that you have an idea as to what the most pressing challenges are for the use case.

Example: List of Ethics Challenges for Mapleville Hospital
Accountability: Currently, there is no way for endocrinologists at Mapleville Hospital to report or raise alarms around potentially dangerous uses of the technology from their perspective. This is a topic that Mapleville should explore in the future, perhaps through the use of a steering committee. This steering committee may also help stakeholders to confront potentially uncomfortable truths that may arise while using the CDSS, which is currently a gap in this project. **Autonomy and consent:** Since the CDSS output relies on clinical EMR data, algorithmic training will be highly dependent on data accuracy, which means if patients are opting out of the program, they will not be represented in the data set, which could potentially skew results. **Fairness:** The CDSS algorithm is highly dependent on EMR data. Currently, Mapleville Hospital mainly serves the affluent population in downtown Toronto, which could translate to skewed data resulting in a biased outcome. Input data could lead to discrimination against lower- and middle-income individuals, as they are not accounted for in the algorithm's training. Using location data, gender, race, and age data could lead to discriminatory practices. **Human connection:** The patient-physician relationship is an important part of delivering health care. While the intent of the CDSS is to provide physicians with the best possible treatment plan for an individual patient, the patient should still be an active participant in this conversation. **Human rights:** The algorithm includes sensitive patient data, which includes gender, age, level of education etc. Patients are made aware of this and can opt out of giving their data. However, at the same time, patients benefit directly from giving their data as the endocrinologist can provide customized care and treatment plans. Their rights are being accounted for, as they have the opportunity to opt out, but ultimately they will lose out on the potential advantages the CDSS has to offer. **Job security:** Physicians need to be made aware that the intent of the CDSS is not to replace them, but rather to supplement their knowledge and expertise which will allow improved and more timely decision-making. **Transparency:** The conclusions of the CDSS algorithm are not made explicitly apparent to the endocrinologists who are using the outputs to tailor treatment plans. This may limit the endocrinologists' ability to explain and make the output explicit to patients.

Trust: The interpretation of the output may vary between stakeholders. Also, the output has the potential to lead to destructive behaviours by enabling the endocrinologist to become reliant on the CDSS output, which may harm the patient-provider relationship and impact trust

Phase 3: Create a Roadmap and Implement

Now that we have a set of ethics challenges identified, it's time to do something about it. While AI ethics is a rapidly evolving field, there are still many solutions to AI ethics issues that have yet to be explored, and many that are difficult to generalize. The good news is that a customized solution can lead to more practical and actionable solutions, rather than solutions that do not really meet the needs of a specific application.

In this phase, your goal is to develop solutions to either address or help manage or address the risks, issues, and needs identified in Phase 2. This involves the following three steps:

1. Create value alignment
2. Co-create and iterate
3. Finalize and communicate

Step 9. Create value alignment

For each issue you've identified, you now have a full understanding of how different values and people related to them. Considering these values that are in conflict and interests that are opposed, brainstorm what **organizational**, **design**, or **communication** decisions could be made to help align the values and interests.

Examples of organizational decisions include:

- Consider different organizational decision-making models
- Enable the internal auditing team to keep people accountable for the AI ethics issues and solutions identified
- Create new corporate policy or modifying existing ones
- Identify individuals or groups to take on specific roles or tasks

Examples of **design decisions** include:

- Design of interfaces to collect specific feedback from stakeholder groups
- Implement technical solutions to de-bias a known bias from a database
- Modify how, when, or where different information about the system is presented to whom

Examples of **communication decisions** include:

- Change the key message used to market the technology
- Hold training sessions with key stakeholder groups to demystify the in-and-outs of the technology

- Present the findings from the AI ethics assessment at the annual general meeting of the company

Focusing on the organization, design, and communication decisions will help to frame solutions that are **practical**, **immediately implementable**, and **action oriented**. This can help dissuade you from thinking about superficial fixes (band-aid solutions) or solutions that require a long time and lots of people to deliver (e.g. solutions that require new global standards or national policies to be developed, or new regulatory bodies established).

Depending on the appropriateness and the resources you have, feel free to involve as many key stakeholders (especially the reference stakeholder) as you can in the brainstorming process. Learning from an existing and growing set of best practices can also inspire you as well (check out the growing list of AI ethics guidelines global inventory).

Step 10. Co-create and iterate

Like any good prototyping process, it is important to test your ideas and make sure it is indeed **practical**, **implementable**, and **action oriented**. With the results of your first brainstorming session(s), touch base with decision makers and future actors who would be involved in implementing your brainstormed solutions and make sure they are indeed implementable solutions. Iterate over the solutions based on the feedback from these decision makers and, if possible, co-create new solutions with the decision makers.

This serves two purposes: it allows you to test your idea, and it fosters buy-in from those who will be taking the decisions and actions forward to implement these solutions later.

Step 11. Finalize and communicate

Once you have solutions to either manage or address each issue, document it in a format you can share with decision makers and future actors of the solutions. The more widely you share your solutions as possible, the better (as long as this is appropriate, and you have everyone's permission to do so). Sharing of an AI ethics assessment is important because it not only boosts **transparency** of the whole project to the stakeholders involved, it also serves to keep everyone accountable to implement the solutions. In addition, it also inspires others to identify new issues that may not have been obvious at this time, and build on your work to develop practical, implementable, and action-oriented solutions to the new issue. As the team takes on new use cases of the technology, the AI ethics assessment will serve as a valuable asset to exploring how the new use case poses different or similar issues.

One company that has taken the bold step to make the full AI ethics assessment fully public is Technical Safety BC. Find the assessment here.