Foresight into AI Ethics (FAIE):

A toolkit for creating an ethics roadmap for your AI project

By Open Roboethics Institute

Version 1.0

October 2019

Brought to life by

AJung Moon, Shalaleh Rismani, Jason Millar, Terralynn Forsyth, Jordan Eshpeter, Muhammad Jaffar, Anh Phan

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. 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 non-profit 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 and we plan to share outcome of our research in AI ethics openly.

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. We invite you to try FAIE and let us know what works and doesn't work by submitting a case study. These case studies will be made public for everyone to benefit from.

Why 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 and anticipate stakeholder and public perceptions. It allows for a smooth rollout and engender trust in the technology.

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, a 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. Further, AI policies, processes, and products that are values-aligned will be better positioned to earn and retain consumer trust. Finally, 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 ten 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 your technology to become knowledgeable stewards. You will acquire the capability to map foreseeable ethics issues specific to your project in design, business, or communication decisions 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 reflect. In order to better demonstrate the process, we walk you through a fictional example of a company that is creating a new AI division to innovate for their product and operations.

How to use this toolkit

To help you walk through the ten steps of FAIE, we include a fictional use case to illustrate the how it can be applied to your project. In this use case involving a fake company, Kids & Adults Inc., we will assume that we are third party consultants hired to create an AI ethics roadmap for their AI project. The complete example from this case study is available on our website if you'd like a more detailed look.

A case study: Kids & Adults Inc.

Kids & Adults Inc. is a nanny agency with around 200 employees based in Vancouver, Canada. Since the company was founded about twenty years ago, its largest source of revenue has been the service they offer in connecting parents with babysitters, professional nannies, and daycares. People wanting to work as nannies or babysitters would file an application, and a Kids & Adults' customer service representative (CSR) would match them to parents who ask for babysitting or nanny services. The process has traditionally been performed manually, where a parent would call the company and a CSR would search through a database of nanny profiles to manually match and schedule the babysitter, nanny, or daycare service.

The new CEO of Kids & Adults just established a data science department to create a software application to automate parts of the process. The CEO hired a lead data scientist to create machine learning algorithms to better analyze profiles of nannies in their database and to continually improve how they provide tailored nanny/babysitter-to-parent matches. She also wants the machine learning algorithm to eventually be integrated into a mobile application, the NannyNow app. This app would be available for download by both parents and nannies/babysitters to book services more efficiently and without needing to go through a CSR.

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 of the same technology or dataset. However, each use of the same technology can have its own 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 cases (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 project.

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 this 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?

Kids & Adults Example

Use Case 1 (primary): The machine learning algorithm will be used by the customer service representatives to better predict which profiles of nannies will most satisfy the parents after a successful nanny booking.
Use Case 2: The machine learning algorithm will be used by both parents and nannies through a smartphone app without needing to go through an employee at Kids & Adults Inc.
Use Case 3: The data science team will use the technology to identify the profile of nannies that tend to lead to dissatisfied parents after a nanny service, and reject applications from future nannies in advance.

In this case, the CEO and the lead data scientist (our reference stakeholder) want to take an incremental approach to the overall AI strategy. The CEO decides that Use Case #1 is the foundation for extracting value out of the data her company already has.

Step 2. Identify stakeholder groups

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

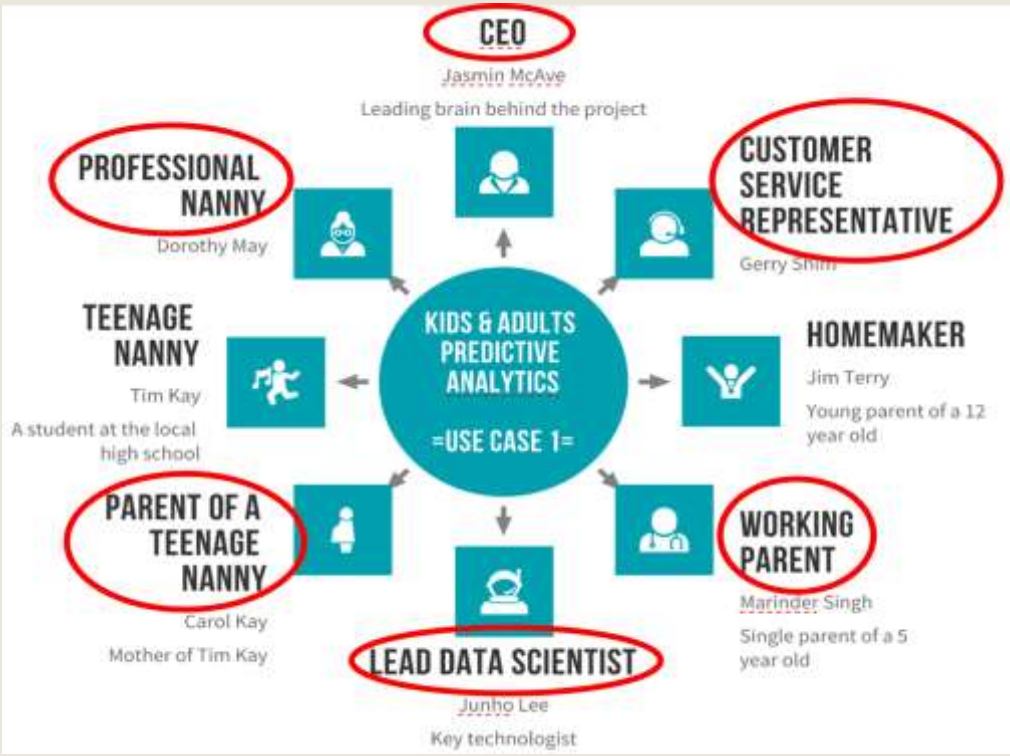
Stakeholders refer to people or organizations who are directly or indirectly impacted by the technology (e.g., customers, vendors, communities), or are closely related to its design and deployment in some way (e.g., data scientists, other practitioners). List as many **stakeholder groups** (i.e., the natural grouping of individual stakeholders who belong in the same category) as possible, as it helps 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 and impactful 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 based on their understanding of who should consulted in this process.

Kids & Adults Example

We identified six stakeholder groups as being key (circled in red). This included the decision to keep Parents of a Teenage Nanny as a key stakeholder as teenage nannies will require permission and parents will likely have stronger opinions.



Phase 2: Discover ethics risks

Now that we know which use case and 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 the technology. This analysis leads us to identify what kind of risks are applicable to the primacy use case, such that we can address or manage them. The following steps are involved:

3. Listen to the key stakeholders
4. Map personas and identify values
5. Discover tensions in people's values
6. Discover tensions in people's activities
7. Discover tensions 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., 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 you can use: secondary research, interviews, and day-in-a-life. Use any combination of the three techniques to find your key stakeholders' values and learn about who they are.

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 ascribes 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 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 goals and values to guide your interviews and analysis
- **Organizational chart** – this chart can help you understand how various stakeholders are related to each other
- **Corporate policies** – existing policies in dealing 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 attain an in-depth perspective of 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.

1-on-1 vs group interviews

Group interviews could be considered; however, make sure to consider the following:

- All the interviewees need to be comfortable to share their experience with each other.
- It is important that the power dynamics in the group does 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 the key stakeholders, making sure to interview at least one person from each key stakeholder group. 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 questions indirectly. Encourage them to share stories and anecdotes in these scenarios as they can be powerful tools for you to better understand their perspectives. Conduct the interviews with the following themes of questions (below).

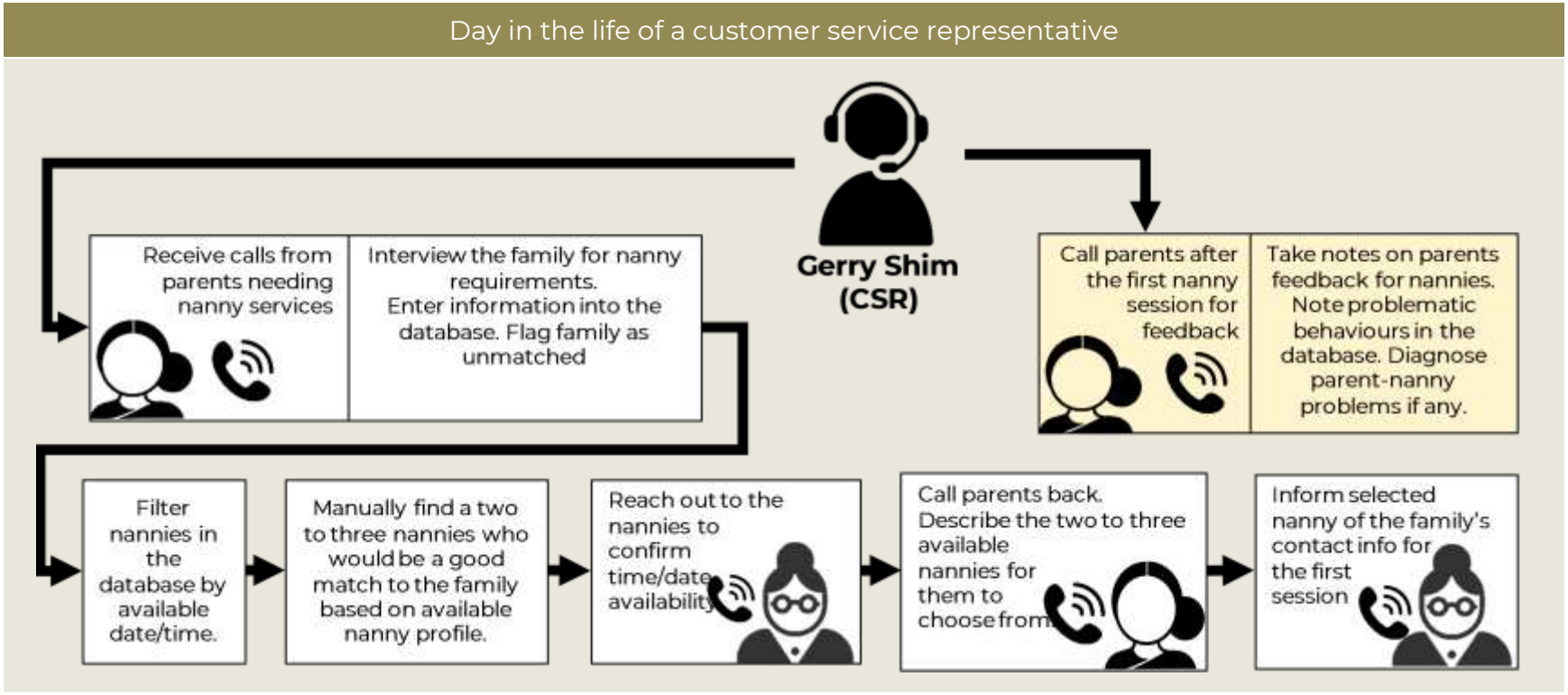
Interview Themes:

Social Ask questions about the stakeholder's history with the organization, what their typical workday looks like, and what their role within the organization and the use case is. This will give us an understanding about the social context within which the technology is to be deployed.	**Value discovery** 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.
Functional If the new AI system is meant to replace another technology (and/or human function), ask questions to reveal the stakeholder's knowledge and use of the affected technology, people, processes, and policies. This will give us an understanding about the stakeholder's relationship with existing technology solution and its relevant processes and 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 societal values such as transparency, trust, fairness, accountability, human rights, human autonomy, and diversity are considered in the existing technology and processes.
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.

Day in a 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 tasks and how personal and organizational values 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.



Other techniques

Group interviews (A.K.A focus groups), supplementary surveys, additional 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.

Selecting stakeholders of day in a 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 providers of data to 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.

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-a-life, and secondary research. It not only highlights the details of what each stakeholder does under their particular constraints, but it also helps to extract the values that are important for the stakeholders. To complete a persona, answer the questions in the stakeholder persona template.

Once you've answered the questions, review your answers to identify which societal values are implicitly or explicitly highlighted by the stakeholder. Societal 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 and their context. 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 provide an example persona of a CSR from Kids & Adults Inc.

Stakeholder Persona Template

Name	**Gerry Shim**, Customer Service Representative	What is this person's primary goal?	To make good matches between families and nannies to reduce unhappy clients & children
What do they like about their job and their interaction with the company?	• Progressive nature of the company • Genuine care for children's well-being • The salary is competitive • Human connection with nannies and families in the network	What other stakeholders support them in achieving his/her goal?	• Other CSRs help Gerry find alternative matches if he has a hard time digging through the database. • The IT team works with Gerry for computer support • Data science team is building a nanny-matching algorithm to support CSRs • Nannies and families give feedback on their latest experience to help improve subsequent matches • The CSR Manager coaches him with regular performance reviews
What do they dislike about their job and their interaction with the company?	• Not enough time to go through good candidate nannies and find an optimal match • Getting blamed for making unhappy family-nanny matches and getting complaint calls from parents • Feeling uncomfortable about matching families with new nannies in the system	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?	• Right now, the machine learning nanny-matching algorithm has yet to be launched. Today, Gerry uses existing database instead. He filters for nannies who are available for the specified date/time first, and then manually goes through the list of nannies to find a match. • The upcoming nanny-matching algorithm is supposed to process nanny profile data and provide CSRs with a list of nannies likely to be a good match to the families.
What are the company/group values?	Public reputation, Client trust, Efficiency, Children's well-being	What policies, regulations, social norms, or technical constraints do they need to work with?	• Parents' description of their preference for nannies are often qualitative. CSRs provide a text summary of their preference in the database, but it is rarely enough. They are also not necessarily the items used to train the algorithm. • Family's privacy and cultural norms need to be respected. • Company HR and privacy policy exists.
Stakeholder values based on the persona	Efficiency, Children's well-being, trust, public reputation, innovation	Societal values extracted from the persona	Privacy, transparency, trust

Step 5. Discover value tensions

Values are the glue that hold 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 some values conflict with other values. These are what we call **value tensions**. These tensions can exist between specific stakeholder values and broader societal values. 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. job security
Tension description: How are the values in tension with one another?	The data science team wants to streamline the nanny matching process for maximum efficiency using the algorithm. Customer service representatives feel that they might lose their jobs if the algorithm performs well and ultimately replaces them.

Value Tension #2

Value tension: Which values are in tension with one another?	Fairness vs. fairness
Tension description: How are the values in tension with one another?	The lead data scientist believes that fairness means all parents have equal chances at getting the best rated nannies for their children. This conflicts with the notion of fairness by customer service representatives, who believe that even unrated nannies (e.g., new nannies) should get equal chances of being matched with easy-going families.

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 peoples' activities.

Let's take a look at the persona of a CSR at Kids & Adults Inc. for an example. In the CSR persona, we find that the feedback parents provide to CSRs about nannies or the description of their preferences are often captured as qualitative text. These text inputs are not used by the algorithm, although it contains rich information that can help find the right parent-family match. This is in tension with the value of efficiency when it comes to use of the existing information to determine a good match.

Step 7. Discover tensions in technology

We now need to understand how the technology is related to the people and our societal values. Your 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**, **how they are related to the stakeholders** (e.g., who is providing the data? who is receiving the output?), **and which societal values may be affected**.

A. Understand the technical components

Consult with your 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.

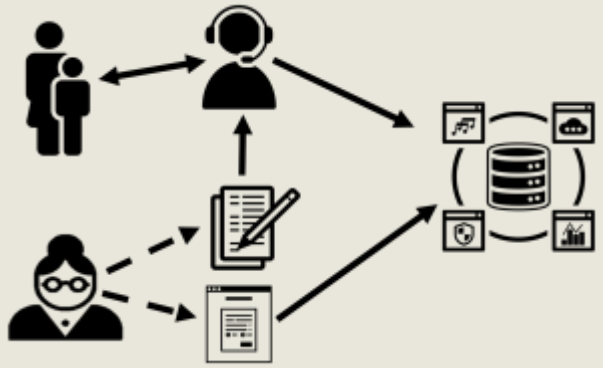
Input

To understand what the **input** of the technology is, you and your reference stakeholder need to consider your use cases 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 Interface: What interface, if any, do people use to provide the data?
- Format: In what format is data stored (e.g., numerical, text, photo, date)?
- Nature of Data: What is the nature of 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 stakeholder's performance or other qualities?

Kids & Adults Example

Kids & Adults Inc., gathers data from multiple sources. However, for the particular use case, the data used for the project comes from nanny applicants when they sign up to be part of the company's group of nannies and CSRs. CSRs also evaluate historical data.



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

The model and output

To understand how the input is processed and what the output is expected to be, answer the following questions:

- Processing: What are the parameters that are within designer/technologist 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?
- Output: What are some core attributes of the technology? Does it make judgements or decisions? Are its outputs predictive? How is it connected to other technical systems or products? Does it have physical actuation capabilities (e.g., smart systems or robots)?

Output and interaction

To understand the **output** of the technology (e.g., a numeric score or a category) and how it's related to the stakeholders, answer the following questions. Note that 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

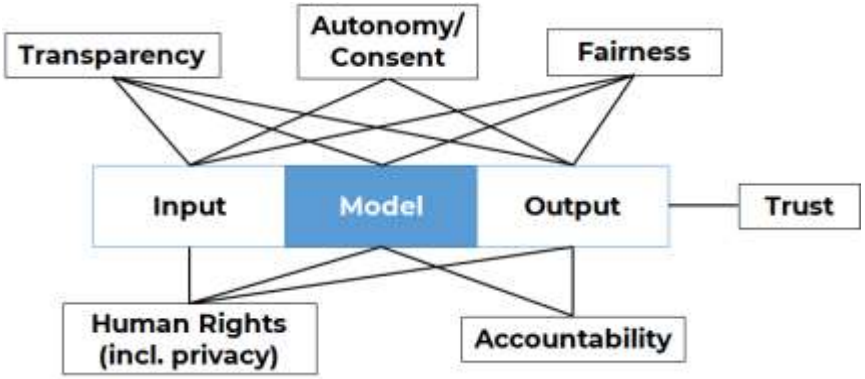
We want to determine how the values are related to the input, model, and output of the system and evaluate where value tensions may arise. Not all values are equally important across the three parts of the technology. The figure below illustrates a mapping of which societal values are typically most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

[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.

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 Kids & Adults Inc. example along with the value questions below.

Notice that, depending on how far along the design of the technology is, the value questions can be modified to ask about plans for the future, rather than present, state. Answers for the Kids & Adults Inc. example is provided below. For our example, we use hashtags #risks, and #need4action to mark where the ethics risks are and where actions are needed to address challenges.

[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	Kids & Adults Inc. Answers
	Input	
Transparency	Do the relevant stakeholders know how/when the information is collected/changed/used?	*No*
	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?	*Not yet in the first version of the system. However, analyzing information from nanny's social media accounts to filter them for any troubling patterns is an attractive idea for Kids & Adults data science team. If nannies are informed of this, we can see how this feature would be undesirable for them and would no longer want to work with the company. #risks*
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?	*Not yet. Kids & Adults will need to put a modified informed consent in place to reflect this new use of the data nanny applicants provide to the company. #needAaction*
	Can the stakeholders decide not to have their data used for the algorithmic system?	*Not yet. #needAaction*
	Are there any elements in the data collection process (e.g., user interface used for inputting data) that could result in unintended outcomes?	*Yes. The web application form nannies use is to enroll can easily be updated with a new informed consent process. However, for nannies who fill out the old paper application form, collecting consent from them afterward is a hassle for CSRs. Kids & Adults may decide to simply exclude them from being matched using the algorithm. This may not work out in the nanny's favour. #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 observations and data entries.	*CSRs collect nanny ratings from parents through phone calls. The majority CSRs at Kids & Adults Inc. are typically female in their 40's to 50's who tend to be more conservative than younger CSRs. Many of them reportedly favour matching families with female over male nannies, and heterosexual nannies over homosexuals if the information is made available in the nanny bios, even if the requesting family does not indicate such a preference. #risks*
	Are certain group of stakeholders' information collected disproportionally more than others? If so, does this fact support or conflict with the societal and stakeholder values?	*Kids & Adults have a disproportionately large number of nannies in their late teen and early 20's living in the cities, although families needing nannies are distributed across cities and suburbs. Therefore, Kids & Adults have more ratings on nannies who are willing to travel to the suburbs to babysit, because they often take on more nanny work than those only willing to work within the populated areas. #risks*
Human Rights	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?	*Yes*

Model		
Transparency	Is the model and its performance understandable to and monitored by those training it?	*Yes*
	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*
Accountability	How often is the model updated/re-trained and is the frequency adequate for the use case?	*Yes. It is re-trained every night.*
	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 and yes, he is able to detect new problems and act upon them*
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 marginalised groups?	*Yes. Since there are richer and poorer neighbourhoods, nanny's address information, for example, could be used to discriminate against nannies from different socioeconomic class. Using data on gender, religion, race, and age can also be lead to discriminatory practices. #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 marginalised groups?	*Yes. All nanny bios written in a non-English language are ignored, since the algorithm only handles text data in English. This may have negative implications for non-English speaking nannies or nannies who speak a foreign language and want to attract families who speak a particular language. #risks*
Human Rights	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*
Output		
Transparency	Is the output from the algorithm presented in such a way that is understandable to its audience?	*Not yet #need4action*
	Is the output presented to the stakeholders in a way that allows them to understand how/why the system has produced the specific output? Is it important for them to understand this?	*Not yet. It will be important for CSRs to know how the system produces a particular output when the output is surprising. #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?	*Not yet. Kids & Adults is considering different options at the moment. #need4action*

	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. There is a possibility that female nannies, who are more often assigned to families due to existing biases in the CSRs, may be recommended to families frequently than male nannies. #risks*
	If someone were to take the outputs from the system and generalise it to other use cases, is it reasonable to foresee problematic interpretations or increase in distrust among stakeholders?	*Yes. Any existing bias that is perpetuated by the algorithm may be wrongly interpreted as facts (e.g., "female nannies are better nannies"). #risks*
Accountability	Who is responsible for acting on the output, and does this stakeholder group have ways to remedy or override erroneous or questionable output?	*CSRs are responsible for using the output appropriately. They have yet to design a way for the CSRs to handle erroneous recommendations. #need4action*
	Is there a communicated and unobstructed means for different stakeholder groups to raise an alarm on possibly dangerous usage of the technology?	*No. #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?	*Yes. If a nanny who is a good match scores low in the nanny-family match score, then the nanny will not be matched. The risk of this is minimal. If a nanny who is a terrible match scores highly, on the other hand, Kids & Adults will likely lose its clients. Execs and the lead data scientist are aware of the balance between these risks.*
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?	*Not yet #risks*
	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?	*It's meant to assist human decisions.*
Fairness	Are the primary users of the technology aware of the potential biases that may have contributed to the output?	*Not yet. #need4action*
	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 does produce the same result for all CSRs.*
	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 the technology development. #need4action*

Human Rights	Does the technology suppress or protect fundamental human rights, such as right to life, liberty, security, freedom of movement and of expression, among others?	*If the output from the algorithm is used to create an overall profile of nannies as being good or bad in a way that is accessible to a third party, this can result in a reputational harm. #risks*

Step 8. Synthesize into ethics challenges

Take a look at the value tensions, risks and needs for action you've noted from the 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 of what are the most pressing challenges for the use case.

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 hard to be generalized. 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.

Kids & Adults Inc. example list of ethics challenges

Transparency: The output is not yet presented in a way that is understandable to the CSRs.
Autonomy and consent: Acquiring consent from older nannies can be a challenge and this may in turn affect the nannies who are part of the predictive algorithm system.
Fairness and diversity:

- CSRs at Kids & Adults Inc. have their own preferences on matching the nannies with families and over time this can affect the matches in enough scale that future matches predicted by the machine will be heavily biased towards the historical opinion of CSRs
- The input data is skewed towards a certain population of nannies just because there is more demand for nannies who are willing to travel to suburbs.
- Using location data, gender, race, age can lead to discriminatory practices
- The algorithm can only take in and respond to bios that are written in English and therefore could be discriminatory to other languages.
- Certain profiles of nannies are more trusted and preferred by the CSRs. However, improving trust by favoring these nannies can be unfair for other nannies
- Various stakeholders have different definitions of fairness. Data scientists think that fairness is giving equal access to families.

Human rights: The nannies can opt out of giving their data but at the same time they benefit directly from giving their data as the families know more about them and it is easier to create better matches for them – there human right to not give the data is given to them but they lose advantages when they do not give their data.
Trust:

- Categorization of the output can be quite subjective when a probability score is matched to a category – in case something is wrong then it will be hard for the user of the output to know how to fix the problem
- Defining trust within the context of nannies – who should you trust – what does trust mean

Human connection: There is less chance for human connection versus efficient processes – risk of losing human connection
Autonomy and job security: The people might lose their job and ultimately their livelihood if there is no plan to remedy that

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 **business**, **design**, or **communication** decisions could be made to help align the values and interests.

Examples of **business decisions** include:

- considering different business models,
- enabling the internal auditing team of the company to keep people accountable for the AI ethics issues and solutions identified,
- creating new corporate policy or modifying existing ones, and
- identifying individuals or groups to take on a specific role or task.

Examples of **design decisions** include:

- design of interfaces to collect specific feedback from a stakeholder group,
- implementing technical solutions to de-bias a known bias from a dataset, and
- modifying how, when, or where different information about the system is presented to whom.

Examples of **communication decisions** include:

- changing the key message used to market the technology,
- holding training sessions with key stakeholder groups to demystify the ins-and-outs of the technology, and
- presenting the findings from the AI ethics assessment at the annual general meeting of the company.

Focusing on business, design, and communication decisions as your solution space helps to frame solutions that are **practical**, **immediately implementable**, **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, the better (as long as it is appropriate, and you have everyone's permission to do so). Sharing 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.