2

Operationalizing Responsible AI: A Thought Experiment—Robbie the Robot

Before we delve into the details of how to operationalize responsible AI principles, this chapter presents an example, designed to illustrate the complexity of responsible AI, and the broad range of stakeholders that need to be involved in the process. We hope you will find this example both fun and illustrative.

A Thought Experiment—Robbie the Robot

To illustrate just how complex the operationalization of responsible AI principles is—and how many different perspectives need to be taken into account—let's walk through a thought experiment. For this experiment, we use Robbie the Robot, the nonspeaking robot introduced by Dr. Susan Calvin in Isaac Asimov's classic book *I, Robot*.

Robbie is a children's robot, designed to play with and take care of kids. Without the ability to speak, Robbie finds other ways to communicate. As Susan Calvin says in the prelude to the chapter on Robbie: "Robbie had no voice. He was a nonspeaking robot. Robbie was made to take care of children. He was a nanny.…"[1]

1. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 11, 2 December 1950.

The first chapter in *I, Robot* then goes on to tell a story of a young girl, Gloria, and her friendship with Robbie. We first find them playing hide-and-seek in Gloria's garden. Gloria is incredibly fond of Robbie, remarking at one point in the chapter: "He was *not* only a machine. He was my *friend!*"[2] But Gloria's mother, Mrs. Weston, is suspicious of Robbie. Although Robbie has been with the family for two years—and there have been no issues—Mrs. Weston gradually starts to worry that Robbie might do something unexpected, and might even harm Gloria.

> "I don't want a machine to take care of my daughter. Nobody knows what it's thinking." She tells her husband. And then: "I wasn't worried at first. But something might happen and that…that thing will go crazy and…"[3]

In the end, Mrs. Weston sends Robbie back to the manufacturer, US Robots. This action upsets Gloria, who really misses him. To try to show Gloria that Robbie is just "some pieces of metal with electricity," Mr. and Mrs. Weston take Gloria to the factory where Robbie was made and is now being used to manufacture other robots. Things don't go according to plan, however. When Gloria sees Robbie, she runs toward him, not noticing a huge tractor on the factory floor, which would have run her over were it not for Robbie, who, seeing Gloria in danger, rescues her. Mr. and Mrs. Weston are forced to take Robbie back to the house, and Gloria is reunited with her best friend.

Although the story of Robbie was originally published in 1940, and predicted a future where children would have robot nannies, we still don't. And to create one remains fiendishly difficult, both from a technical perspective (we still struggle to get robots to carry out seemingly simple tasks such as playing a game of hide-and-seek) and from a perspective of responsibility (how can Mrs. Weston be confident that Robbie won't go "crazy" and hurt her daughter?). The trope of kids befriending robots has since been explored extensively in popular entertainment, in movies such as *The Iron Giant, Big Hero 6,* and *Earth to Echo*. In many of these stories, the robot AI does indeed go "crazy" and bad things happen; the recent movie *M3GAN* is a good example in the horror genre.

Who Should Be Involved in Building Robbie?

In the remainder of this chapter, we use Robbie the Robot as an example to consider where responsible AI issues come up. Let's consider things from the perspective of US Robots, the company that created Robbie.

As discussed earlier in this chapter, a diverse set of stakeholders need to be involved in building, using, and managing an AI system such as Robbie the Robot. Each stakeholder has knowledge that will contribute to making sure that Robbie is designed responsibly. Table 2.1 lists some of the stakeholders that US Robots should include, as well as the key contributions each of these stakeholders can make when it comes to designing Robbie in a responsible way.

2. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 16, 2 December 1950.

3. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 15, 2 December 1950.

Table 2.1 Stakeholders of Robbie

Stakeholder Type	Role	Responsible AI Contribution	Type of Contribution
US Robots Company Board	To manage the reputation and market risks of developing Robbie	Ensures that a risk management framework is set up and monitored to assess, mitigate, and manage risks associated with Robbie's deployment in family settings	Governance
Government	To ensure the safety of the general public	Enacts laws that regulate how family robots are designed, manufactured, and used	Governance
Industry Bodies	To produce standards for robotics companies to follow	Creates standards for family robots that member organizations agree to follow	Governance
Parent Groups	To advise parents on the use of Robbie and to lobby government on appropriate legislation	Sets up information sharing for parents, e.g., workshops, web portals	Governance
VP Ethics	US Robots executive who ensures the company has a reputation for responsible AI	Defines and rolls out training and practices for responsible AI across the company; may include independent testing of Robbie features before release	Process
COO	US Robots executive responsible for effective and efficient processes within the company	Implements recommendations from the VP Ethics to ensure company practices include responsible AI considerations	Process
Product Manager	Team member who makes decisions as to which features go into Robbie (e.g., what its objectives are, what the constraints are)	Sets up and manages a process to get customer input on desired features and works with technical experts on feasibility	Process
Project Manager	Team member who manages the development of Robbie over time, ensuring delivery of features according to an agreed-upon schedule	Ensures that the project plan includes key check-in points to consider issues related to responsible AI	Process
Technical Manager	Team member who manages the technical teams developing Robbie to deliver agreed-upon features	Ensures best-practice responsible AI guidelines (coding practices, appropriate use of off-the-shelf components) are used	Product
Data Scientist	Team member who manages the data that Robbie is trained on to carry out tasks	Ensures, as far as possible, that training data is representative of the broader population and not biased to one segment of society	Product
AI Expert	Team member who develops AI models to process data	Monitors AI model performance for bias	Product
Software Engineer	Team member who manages the integration of Robbie into US Robots' larger software systems as required	Ensures that best-practice responsible AI software patterns are used	Process
General Public	Users of Robbie	Provides feedback to US Robots to ensure that any responsible AI issues are fixed	Governance
Suppliers	Other manufacturers or AI technology/solution providers	Ensures the supplied product components are without any responsible AI issues	Governance

As Table 2.1 shows, responsible AI is complex: many stakeholders need to be involved. The good news, however, is that this is no different to any complex systems engineering task. Building skyscrapers, flying airplanes, implementing large-scale government information systems—these are all examples of complex engineering projects that society operates routinely today. And, over time, society has agreed upon sets of rigorous processes and methods to ensure that such systems are safe, secure, and operate as expected. The only difference with responsible AI is that AI is a fast-moving technology, so we do not yet have a full set of rigorous practices. (This book, of course, partially fills that gap!)

What Are the Responsible AI Principles for Robbie?

The first step in ensuring that Robbie implements AI responsibly is for US Robots to agree to a high-level set of responsible AI principles. These could be Australia's AI Ethics Principles, as described in Chapter 1, or they could be something company- or context-specific. In his book, Asimov famously captured the operating principles of US Robots as the Three Laws of Robotics, codified in the Handbook of Robotics, 2058 AD:

1. A robot must not harm a human. And it must not allow a human to be harmed.
2. A robot must obey a human's order, unless that order conflicts with the First Law.
3. A robot must protect itself, unless this protection conflicts with the First or Second Laws.[4]

These Robot Laws were encoded in Robbie's positronic brain to ensure that they would be followed. For a modern engineering firm creating a robot like Robbie, these laws could well serve as high-level principles to follow. But to encode them in the design and operation of a robot, they need to be made more concrete (i.e., the laws must be operationalized).

To some extent, Asimov's laws can be related to modern AI ethics principles. Table 2.2, for example, maps them to Australia's AI Ethics Principles. Note that some of Asimov's laws map in a fairly straightforward manner. It becomes quickly clear, however, that Asimov's laws are actually quite narrow. Other than the safety of humans, they say nothing about what is considered societally appropriate behavior by Robbie. For example, one would expect Robbie, as a child's companion, to act and teach in a way that is considered proper. In modern-day AI systems, in contrast, there is a lot of concern about whether AI systems will exhibit behavior that is discriminatory, biased, unfair, or socially unacceptable. None of this concern is captured in Asimov's laws. Arguably, this kind of behavior could be included under the First Law, but this depends on the definition of *harm*, which in Asimov's book is largely focused on physical safety.

4. I. Asimov, *I, Robot* (Gnome Press, 1950), 1st Edition, page 9, 2 December 1950.

Table 2.2 Mapping Asimov's Laws to Australia's AI Ethics Principles

AI Ethics Principle	Description	Sample Problematic Behaviors in Robbie Context	Covered by Asimov's Laws?
Human, societal, and environmental well-being	AI systems should benefit individuals, society, and the environment.	Robbie causes problems with children, such as child safety or psychological dependency.	Partially—First Law covers physical safety but not broader well-being issues.
Human-centered values	AI systems should respect human rights, diversity, and the autonomy of individuals.	Robbie does not encourage children to assert their right of freedom of speech.	No—Laws say nothing about human rights.
Fairness	AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities, or groups.	Robbie treats children from different backgrounds differently.	No—Laws say nothing about diversity in end users.
Privacy protection and security	AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.	Robbie collects data from the child and shares with the company.	No—Laws do not cover privacy.
Reliability and safety	AI systems should reliably operate in accordance with their intended purpose.	Robbie fails to rescue Gloria from the tractor due to a malfunction.	Partially—Second Law somewhat covers "intended purpose" but does not explicitly address malfunctions.
Transparency and explainability	There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.	Robbie fails to tell Mrs. Weston that he has been cheating Gloria at hide-and-seek.	Partially—Second Law guarantees that Robbie explains his actions but only if explicitly asked.
Contestability	When an AI system significantly impacts a person, community, group, or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.	Mrs. Weston is unable to get Robbie to teach Gloria in a way that she wants.	Partially—Second Law guarantees a challenge of Robbie's outcomes but only if explicitly asked.
Accountability	People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.	US Robots fails to put appropriate procedures in place to ensure Robbie follows the three laws.	No.

Robbie and Governance Considerations

Putting aside Asimov's Laws for a moment, as they are clearly incomplete for our purposes, let's move forward assuming the AI ethics principles in Table 2.2 are our driver.

Table 2.1 identifies six stakeholders relevant to Governance. Let's consider just one of these, the company board. Like any board, the main purpose of the board of US Robots is to set the strategic direction of the company and to ensure that the company is operating within all relevant laws, ethically, and in a way that safeguards the reputation and financial sustainability of the company.

Imagine, then, the position of the CEO of US Robots. She's just had a brilliant idea: to create a new robot, which will be called Robbie, that will act as a child's nanny. It could be a big money-spinner for the company and could really place US Robots on the map as a global leader in robotics technologies. The only remaining question is what will the board think? In many ways, the board's main job is to think about what can go wrong and make sure that the CEO has a plan to deal with any potential threats. In the case of Robbie, the board can imagine a *lot* that can go wrong. Robbie could accidentally hurt a child; he's a heavy piece of metal, after all, and could easily put one of his heavy metal feet in the wrong place. Or Robbie could inflict psychological damage on a child by inadvertently creating an emotional dependency. How will Robbie protect children from harm caused by others? Are Robbie's computer vision systems good enough to identify all harmful objects correctly, or will he miss one? Robbie can't speak, so there is less risk that he will fill the child's head with inappropriate thoughts, but there's still a risk of not being inclusive; he'll need to be programmed with all the different customs and traditions of children from different ethnic and religious backgrounds. And what if Robbie breaks the law? Will the company ultimately be responsible? What HR practices should the board ensure are in place to reprimand engineers who build the wrong mechanisms into Robbie?

It isn't the board's job to provide answers to all of these questions. That is the CEO's job. The board, however, needs to make sure that the questions are asked—and that someone has the answers.

Fortunately, the board is a sophisticated one. Board members gather all the relevant experts together and come up with a plan of action. The board directs the CEO to do the following:

- Develop a responsible AI risk assessment (see G.12. RAI Risk Assessment). One way to do this is to start with the AI ethics principles in Table 2.2 and then imagine all the things that can go wrong. Each of them represents a risk; the board agrees to a risk likelihood and impact severity in each case, and considers mitigation actions that can be put in place to reduce the overall risk rating.
- Introduce ethics training across Project Robbie (see G.13. RAI Training). The board is aware that their workforce is diverse. It includes graduates fresh out of college who are up to speed with the latest technological developments but, as primarily technical specialists, may not have any background or training in the social impacts of technology. The company also includes many staff who have worked for the company for years; they have a good sense of the company's core customer needs but may not be up to speed on the latest technological developments and, in particular, the ethical risks associated with them. So, the board decides that everyone working on Project Robbie should undergo mandatory ethics training.

- Set up an ethics committee as a subcommittee of the board (see G.10. RAI Risk Committee). The board realizes that it has too many things to worry about to leave ethics to the board itself. So it delegates responsibility to an ethics committee, whose job is to oversee the implementation of Robbie in a responsible way. But to make sure that the board has visibility and remains accountable, the ethics committee will be composed of a subset of board members and will be chaired by the most relevant board member. It is at this point that the board members realize they do not have enough ethics expertise on the board, so they go back and revise their board skills matrix to include ethics, and the chair goes out to recruit a new board member with the requisite experience who can chair the subcommittee. No work on Project Robbie will commence until this is done.

The CEO explains to the board that Project Robbie is complex, at a scale unlike anything the company has tackled before. "We can't build Robbie by ourselves," the CEO explains, and she goes on to explain that US Robots will need to procure components of Robbie from other providers. The board agrees, but to ensure that Robbie remains an exemplar of responsible AI, the board insists that all acquired components go through a rigorous responsible AI evaluation process before considering their use, including how they will interact with other components (see G.15. RAI Bill of Materials).

The board is happy with its decisions. It's been a busy few weeks for board members, figuring out how all of this is going to work, but they are content with the outcome. They are happy to support this new idea from the CEO, and they agree that it could be a new future for the company. But they are also as confident as they can be that Robbie will be developed in a responsible way and that, in particular, there won't be any adverse events that will come back to haunt the company.

The latest board meeting is about to finish. Everyone is happy. Until, almost as an afterthought, the CEO raises a question.

"Have we done enough?" she asks the chair.

"What do you mean? We're implementing all these measures."

"Yes," continues the CEO. "But are they enough? Is there more we can do?"

"I can't think of anything," says another board member.

The board chair reflects for a moment and then, like the wise experienced executive that she is, she says: "I can't think of anything either. But that doesn't mean there isn't anything. Maybe we are just not seeing it. Let's do two things. First, we'll get an independent review of our plan by experts in the field to make sure it holds water. Second, we'll have a quarterly review at board meetings to make sure it's working and there's nothing we're forgetting."

The board meeting ends, and the exciting work on creating Robbie, the children's nanny robot, begins.

Robbie and Process Considerations

Is the company's work on responsible AI done? After all, the board and the CEO have put in place rigorous mechanisms to assess and track the risks associated with Robbie's development. Things should be fine, right?

Of course, the company's work is far from done. In fact, it is just beginning. Governance considerations have been taken care of, but what about process issues? The CEO summons her VP Ethics and COO.

"I have some exciting news," starts the CEO. "The board has just approved that we can go ahead with Robbie!"

"That's fantastic," says the VP Ethics. "But now we have some *real* work to do."

The CEO, VP Ethics, and COO agree to put a working group together, containing key experts and stakeholders from across the company, to define a process approach to developing Robbie. It takes a few months, and some in the company are frustrated that development on Robbie can't start until the process considerations are resolved, but the CEO is firm: "We must get the processes right before starting."

The working group reports back to the CEO, who takes the recommendations to the board. Recommendations include

- **Verifiable Responsible AI Requirements** (see P.2. Verifiable RAI Requirement): The first issue the working group addresses is that the definition of AI Ethics is too vague to measure. The working group's recommendation is that the business analyst team develop a set of verifiable ethical AI requirements. For example, the group says, the ethical AI principle, transparency, could partially be satisfied by a requirement that Robbie includes a parent app where parents can review all Robbie's interactions with their children.
- **A Rigorous Data Lifecycle** (see P.3. Lifecycle-driven Data Requirement): The working group realizes that Robbie needs to respect different cultural traditions (as captured in one of the verifiable ethical AI requirements!). So the group defines a process for careful management of the data lifecycle—what data is collected, how it is managed, who has access, and so on—so that the data loaded into Robbie initially, as well as the way that Robbie collects additional data through sensing, is diverse and treats people from different cultural backgrounds equally.
- **Responsible Design** (see P.7. RAI Design Modeling): The working group also recommends that the responsible AI requirements are considered throughout the design process. They suggest a suite of processes for designing features that ensures the designers put responsible AI first, not let it be an afterthought.
- **Responsible AI Simulation** (see P.8. System-Level RAI Simulation): The working group strongly recommends that the company's simulation platforms, which it currently uses to simulate robotic interactions before deployment, are updated to build in ethical AI considerations. The working group is excited by the prospects here; they suggest using an AI simulator to run what-if scenarios and measure compliance to the verifiable ethical requirements over as many scenarios as possible. "We're using AI to test AI," they muse.
- **Software Engineering Process** (see P.10. RAI Governance of APIs, P.12. RAI Construction with Reuse, P.16. Extensible, Adaptive, and Dynamic RAI Risk Assessment): The working group takes a good look at the company's existing software engineering processes. Group members quickly realize that responsible AI is not built in. So the working group consults with relevant stakeholders and comes up with adaptations to existing engineering processes to make sure

that responsible AI is the primary consideration. Changes include the reuse of AI assets (to ensure that best-practice responsible AI is reused across the development), AI risk assessment at all levels of development (not just done once and forgotten), and a new process for testing Robbie's APIs to ensure there are no privacy leaks.

The board invites the working group to a special meeting of the board, where it runs a rigorous process to test the assumptions and recommendations of the working group. The careful probing of the board leads to some improvements, but, ultimately, the board members are happy. The board chair, however, wants visibility of the process implementation.

"Let's introduce regular review points," she says. "We'll do this quarterly so we can see how well the new process is working out, and if there need to be any changes."

Robbie and Product Considerations

At this point, many of the developers and AI experts within US Robots are getting very excited. They've been hearing about this new robot project for months. There are rumors, but there never seems to be any indication of a timeline for starting work on the project. Until, one day, the CEO sends an internal communication to the teams:

> Dear Team,
>
> I am very pleased to inform you that the board has now approved a start date for the development of our latest robot, Robbie. Robbie will be a children's companion robot. It will revolutionize the way that families interact with robots. This is an opportunity to change the world! But we must do this responsibly. And so, we have spent the last few months being rigorous about how we will ensure that Robbie does no harm.
>
> We are now ready to embark on this adventure, and I look forward to working with you all on what will be a challenging but exciting initiative.

US Robots is abuzz with enthusiasm.

But the development teams know there is a lot of hard work ahead. They also know that the first, and most important, consideration is to make sure Robbie is developed ethically. The teams have been undergoing mandatory ethics training for many weeks now. There have been constant communications from the executive team about the importance of responsible AI—not just in the Robbie project, but in all projects. And line managers have asked all their staff to write clear objectives in their annual plans about how they will contribute to responsible AI.

The product manager and project manager for Robbie get together to agree on a way forward. They have been briefed on the new process, with responsible AI built in, that they will follow. But many system-level design decisions still need to be made. And the product and project managers are insistent that these also should put responsible AI first. They decide to do the following:

- Ensure responsible AI is built into Robbie's supply chain (see D.1. RAI Bill of Materials Registry). Robbie's development will be highly dependent on external providers, both of hardware and software components. A project as complex as Robbie can't be delivered by a single

company, even one as large as US Robots. "We need to make sure all external components are developed to the same high standards when it comes to responsible AI," says the product manager, sensibly.

- Build in *kill switches* at multiple levels (see D.5. AI Mode Switcher). The project manager is concerned that, even if rigorous responsible AI practices are properly followed, situations outside the team's control may still come up once Robbie is active. "We should build in *kill switches*, both local and remote ones, so that, if anything doesn't look right, we can shut down different parts of the AI before things get out of hand."
- Build redundancy into critical AI systems (see D.6. Multi-Model Decision-Maker). The product manager: "Any time that Robbie could potentially put a child in harm's way—even if that potential is very remote—we should make sure multiple AI models are running in parallel. This will give us confidence that Robbie is only making critical decisions if all the models agree." The project manager: "We could go further than that, and if the models disagree, activate a *kill switch*."
- Quarantine new features (see D.9. RAI Sandbox). The product manager: "We'll need to introduce new features once Robbie is active in society. There's no way around this; at the very least, it will be needed to fix issues without recalling all versions of Robbie. The project manager agrees and replies, "We should quarantine new features when they are rolled out by isolating it from other critical AI components wherever possible—at least until it's fully tested in the field."

Summary

As you can see, when it comes to responsible AI, there is a lot to think about. Responsible AI isn't the job of a single group of people. Rather, it needs to be embedded at all levels across a company. Neither is responsible AI something you do once and then forget. It is a constant challenge to review and re-review the approach. And, of course, there is a tension between the need to be responsible—and therefore, cautious—and the need to get features out the door and into a product. All of these considerations need to be taken seriously.

The example in this chapter is obviously an idealized scenario. There is no mention of the downsides of introducing governance, process, and product measures to ensure responsible AI. In practice, these measures cost money, and these costs may need to be balanced with the need to get a product out to market—although this, in itself, is an important decision to discuss in the context of responsible AI. One might argue that for-profit companies only care about profit, so many of these measures won't be implemented. However, public and government opinion about responsible AI is clearly changing. It is becoming a competitive advantage to be responsible. And we are likely to see companies measured for it in the same way that they are measured—either formally through Environmental, Social, Governance (ESG) metrics or informally through reputation—for impacts on society.

Good luck, Robbie! We hope that US Robots has done a good job in building your AI responsibly.