Chapter 4 Agreement

The next day, Chelsea got a text from Charlie that the agreement form and link was in her email. She had found Robbie's birth certificate and searching for it had brought her to tears as she saw pictures and drawings from when he was little. Even then, as little as 4 years old, he was putting robots into his pictures. As she had thought about Charlie's gift, she got more excited about Robbie's birthday and how he will respond to a sophisticated robot friend. He loves robots so much and wants to learn about the technology behind them, this toy will be perfect for him to advance with.

Chelsea opened the agreement email. It was long and wordy, but at the top had a link to the online consent form. She did notice this phrase as she quickly browsed the email text.

> *You are embarking on a breakthrough journey in human and robotic relationships. Your Companion Robot will become part of your family, like a fond pet, and will provide joy, comfort, and humor for years to come. We truly wish you all the enjoyment that your Companion Robot can provide. Thank you for choosing WhyRobot.*

As she clicked into the online consent form, the website requested an access code. Charlie had sent that in his text, and she would need that in addition to her email used as the login. Once she was logged in, she realized that this long form had already been filled out by Charlie. At the top of the form, she was prompted to upload a picture of her license or passport. She wasn't ready for those but because she was doing this on her phone, it was simple enough to take those pictures right away. Once uploaded, the website requested an immediate selfie picture. This caught Chelsea off guard. "They are verifying my identity in real time," Chelsea thought, "that is pretty cool technology." The form used electronic signatures, and she was required to sign that all information was correct. Charlie had said he would do most of the details and since her time was short, she was happy that she didn't have to fill in her address and the many other details on the form. As she browsed, she realized that the whole family was being called out, including Sam as deceased and Frank as Robbie's stepdad. Robbie was given the role "Principal Bond" and was identified as a minor of 7 years old. This required her to upload his birth certificate, a current picture, and have her do another electronic signature approving the role. Fortunately, she had just taken a picture of Robbie on his bike the day before. Charlie was given the role Configuration Operator, which she was happy to have him take on as herself or Frank would have never been able to set this up.

Chelsea moved on to the survey and reporting section of the agreement form. For the most part, Charlie's email address was in the key options. Her email address was not included in the daily details or summary but only the quarterly report and survey. Charlie was signed up to receive an instant violation report, a daily detailed report, and the same quarterly report and survey. She was so happy to not have to get more email spam from this robot and have to learn the details. "I'm just going to let Charlie and Robbie nerd out on this thing and let them take care of it", she thought. There was a space for the Principal Bond's email, which was left blank. Robbie didn't use email yet and she didn't want to expose him to all the spam emails that come with an email account. One option did catch her eye, "Ethical and Physical Violations Reporting (i)", which are included in the detailed reports. She paused to think what

this might be, wondering if there was privacy risk. She clicked the little (i) next to the label and a small popup came on her phone:

> *Ethical and Physical Violations Reports detail incidences where the robot's preprogrammed ethics and physical safety has been violated. Examples: (1) When a family member or Principal Bond physically strikes the robot. (2) When an ethical conflict emerges such as asking the robot to hurt someone in violation of its prime directives. Details can be found in section 8.2 of the user's manual here.*

"Oh, that makes sense," thought Chelsea. "If Robbie starts hitting the robot or asking it to do weird things, Charlie will get notified. Perfect. Charlie has such a good relationship with Robbie that it will be much better coming from Charlie than me. Charlie can be the bad guy and I won't have to be."

Chelsea noticed that Sam had been included in the relationships table that Charlie had already filled out and his health status was "Deceased". "Why did WhyRobot need to know that Robbie's real father was dead?" she thought to herself. Looking through the table she noticed Frank having relationship "StepFather". "Does it matter that he's a step father or a real father?" She decided that this was in the details that she didn't want to dive into. She shrugged and said to herself, "let Charlie take care of this" and she moved on.

After a cursory review of the remaining options, everything seemed to be setup correctly. Chelsea used her electronic signature at the end of the agreement and a popup acknowledged her submission. Charlie and Chelsea were emailed a copy of the agreement and the final steps for robot delivery were now set to take place. Estimated delivery to Charlie's house was 5 days. Chelsea took a screen shot of the notice and texted it to Charlie. Little did she know that WhyRobot had already sent him an email and a text message announcing her signature.

###

Charlie, who's work had been interrupted by a WhyRobot text message, quickly went opened his phone and reviewed the agreement options. She hadn't changed anything. Awesome. Charlie was now poised to have this robot engage in Robbie's daily life and give him insight into what is going on there. He paused for a moment to realize that he was now a virtual peeping Tom but mentally justified it because of Sam's death, the nature of it, and how the relationship with Frank did not appear to be healthy. Not to worry, Charlie would engage Chelsea if a real issue came up, so in the end, she will be the one dealing with it.

Neither Chelsea nor Charlie read section 8.2 of the user's manual in detail. Charlie skimmed over it and Chelsea never clicked into it. WhyRobot workers, on the other hand, had spent hundreds of person hours debating the details of those sections and thousands of programming hours to bring the expected functionality to reality. Corporate lawyers, ethics consultants, and even retired police and judges were employed time and again to help refine the details of what a Companion Robot could observe, record, and report. What would a Companion Robot do if it observed abuse to the Principal Bond? What constitutes abuse? What if a law was violated by the family? Has the Companion Robot become a judge, a spy, or just a concerned family friend when something unethical, dishonest, or physically abusive is observed? How much can be disclosed and to whom? These very difficult questions were part of the beta effort for this robot and Chelsea and Charlie had just agreed to be part of that effort. What no one

really knew was what would happen if an artificially intelligent computation system was faced with conflicting violations of its principal directives, its knowledge of law and ethics, and its goal system to serve the Principal Bond. Would it decide to do nothing, which is an easy decision from one perspective. Would it take an action, which is often justified, but now makes the robot and maybe even its creators participants in the situation at hand.

The programmers and validators at WhyRobot had worked hard at making the computation and decision system adaptable, self-learning, progressive, and self-willed all the while keeping the robot and its Principal Bond happy and safe. The complexity of such a system is overwhelming, especially to the validators, because of the huge input space coupled with an intentionally random selection of acceptable decisions and actions. The designers and programmers wanted to mimic spontaneity with preferences towards humor and generating positive human responses. To the validators, the input space was ultimately unpredictable and unbounded. The input also included the saved experiences and training from events in the field, beyond what was initially programmed. If that wasn't enough challenge, the addition of the detection and responses to violations compounded the already huge validation space. Violation detection and subsequent actions were a large focus for the development team. The results were difficult to predict, and everyone was worried about the robot's actions in response to violations. Ultimately, the system architects introduced an independent violations engine. This separate violations engine analyzed both the behavior of humans and the robot. If the robot ever detected a violation in its own proposed actions, the violation engine would identify it and flag it to the decision system to prevent taking that action. For the humans, if the robot identified a violation, it would be logged, sent to the decision system as an additional decision input, and finally, an appropriate action would result depending on how the robot was configured.

The WhyRobot validators had generated over 300 different violation exposure scenarios and tested the robot for its response and conclusions. They also used a scenario generator taking two at a time combinations of the 300 scenarios and tested the robot. An example of this would be the combination of the Principal Bond physically abusing the robot in a time window where the robot was observing a family member abusing the Principal Bond. The robot abuse alone would result in reporting and a likely robot shut down. The Principal Bond abuse alone would result in reporting, potentially to authorities, and the robot remaining a comfort and support to the Principal Bond. The two combined creates a challenging choice for the robot's decision system, shut down or be a comfort to the Principal Bond. The final decision is a gray one as it depends on the severity of both abuses. Because of limitations of the number of WhyRobot validators, they used a method of developing an independent checker that would try and predict what the robot should do and then check if the robot had done the right thing when exposed to these scenarios. The checker was simpler than the robot's computation system, so it would produce many false failures, flagging that the robot had done the wrong thing, when in fact, the robot had done the right thing. This list of false failures had to be manually reviewed and checked off which became incredibly effort intensive. These checker exceptions were then manually moved to an exception file that would prevent the flagging of future false failures of those types. One lead validator believed that two at a time scenarios was not enough to verify safety, but the leaders of WhyRobot decided against three or more scenarios at a time as too complex for the checker or the number of validators available. Instead, two junior validators were tasked with reviewing the list of 300 exposure scenarios and generating manually the best three or more combination scenarios. An example of this would be the parent was abusing the Principal Bond, the Principal Bond was stealing money from the

parent, and the Principal Bond was abusing the robot. Reporting the money stealing to the parent would result in more abuse to the Principal Bond, violating a prime directive for the robot to keep the Principal Bond happy and safe. Unfortunately, this validation task was on top of many other tasks, and only a few complex scenarios would ever get created and tested. The untested combination scenario space of 27 million was enormous and essentially made the robot's response in complex violation scenarios an unknown. One of the major inputs into this decision system is the AI inference engine. This engine uses WhyRobot trained scenarios in addition to the field learned state such as real experiences, decisions, and outcomes. This effectively created an infinite set of inputs into the decision system to which no number of validators at WhyRobot could have ever proved correct. In fact, the programmers and validators had given up on the expectation of correct or predictable behavior of the decision system, they just didn't want the system to ever decide to harm someone or itself. They all relied on the violations engine to prevent that scenario. Unfortunately, the world is analog, dealing with shades of gray and not just black and white decisions, something that makes a complex decision system very hard to predict.

