Here are two unpublished sections, Chaps 25B and 26B of "The Turing Option" by Harry Harrison and Marvin Minsky, Warner Books, August 1992.

Harry and I were longtime friends. One day he told me how much he liked the ideas in my book 'The Society of Mind.' He suggested that the ideas could reach a larger audience if I wrote a more popular version in the form of a novel. When I said that I didn't have the right talents for that, Harry offered to collaborate. We decided that the central figure would be a mathematical super-hacker of the future who would build the first AI with a human-like mind. Harry would draft the action plot, and I would supply the technical stuff. Great. I had admired Harry's work for at least 20 years and I knoew he had the depth for the enterprise. (In all the time we had been friends, he had never mentioned that in earlier years, he had done the Flash Gordon cartoons, first taking over the illustrations, and later the text.) Harry's agent in New York sold the plan to Warner Books. Over the next couple of years emerged the plot idea in which the hero sustains a grave brain injury. This let us explain the computational part of the theory in the context of repairing Brian's brain, while also explaining the psychological aspects in the context of reconstructing his childhood memories.

A co-author does not have complete control. These two chapters are part of the text that I wrote which did not make it into the final book. Both Harry and Brian Thomsen, the editor at Warner, thought they would slow down the story too much. At the time I firmly disagreed, but now I think they were thoroughly right! I plan eventually to put the rest of that material into another book on the subject of--the only title that seems right would be Asimov's term "Robopsychology". We wanted to show Isaac the chapter drafts, because they are about the same problems that Isaac explored, namely, how to keep a robot (or person) from going insane. But Isaac was so ill those last years that we didn't want to hassle him.

Instructions: 1. Delete the section discussing Freud, from page 244 after ("...submind that can maintain a measure of control.") to the end of chapter 24. Then insert Chap 25B below at page 252:

Chapter 25B June 19, 2024 Cpr. Sept, 1992 Marvin Minsky

When Brian and Ben reached the lab, the computer was running but the tree-robot was folded and motionless. "Robin, activate."

The robot unfolded and turned toward them.

"Good morning, Brian. And stranger."

"This is Ben. You have met him before."

"Good morning, Ben. I presume that we must first have met only recently, because I find no record of your appearance in my long term memory records, which terminate as of one month ago."

"Your presumption is correct, Robin." Ben turned to Brian. "Damn," he complained, "I'd like to take notes but I've run out of paper."

"I can't believe anyone still uses that stuff," Brian grumbled. "Robin, would you please get Ben some paper?"

"How many sheets should I get?" Robin asked.

"Just a few."

"I don't understand what you mean by 'a few.'"

"'A few' means a small number like three or four."

"I understand. Shall I bring three sheets or four sheets?"

"No, Robin, you did not understand. The expression 'three or four' is an idiom. Do you know what an idiom is."

"Yes, sir. An idiom is a language expression whose meaning is peculiar to itself in that it cannot be derived from the conjoined meanings of its elements, or a style of artistic expression or structural form characteristic of a certain..."

"Robin, stop! Brian intervened. 'Three or four' simply means a number close to three or four."

"I understand. I shall bring 99 sheets."

"What made you choose 99?"

"Because 'close to' means approximately," the robot replied. "And 4 is close to 5. And 5 is nearly 6. And 6 is almost 7. According to my thesaurus, I am using those words correctly. And 8 is approximately..."

"Robin, stop!" Brian turned to the terminal and examined the program. "Aha. The trouble is in Robin's concept of approximation. The declarative definition seems OK, but the procedural definition is defective."

"Which means?"

"That it's one thing to know how a term is defined P but you also must know when to apply the idea. In particular, you can't use approximations repeatedly the way Robin that did. The errors accumulate too much, and Robin almost got stuck in an endless loop."

"Then why did it stop at all, when it got up to 99?"

"Hmm. Now that you mention it, I really don't know. Let's backtrace its memory." Brian returned to the terminal. "Aha. Look at this. First it found that 'few' could be defined as 'not many'. Then it found this example where a hundred things is called many. So it used one hundred as a test for 'not few'!"

"Will that be easy to fix?"

"Should be. I'll just tell the B-brain to keep the A-brain from repeating the same thing over again."

A minute later, Brian looked up. "That ought to do it. Now, Robin, get some paper for Ben."

Robin started to cross the room but suddenly veered to the left and began to perform increasingly elaborate contortions. "I am trying to get there, Brian," the robot said, "but it is becoming increasingly difficult."

Ben was roaring with laughter. "Of course! You forbid it to do the same thing twice. So every time it takes a new step, it has to find a new way to do it."

"I guess it's back to the old drawing board." Brian frowned. "What is a drawing board, anyway?"

"A drawing board," replied Robin, "is a rectangular table with adjustable slope, used for..."

"Robin, stop!"

"Anyway," Brian continued, "that was stupid of me. There's nothing wrong with repeating things Q so long as you're making progress toward your goal. So I'll install a supervisor to permit repetition whenever it leads to progress."

"Sounds good to meQexcept, how do you define 'progress'?"

"Depends on the situation. For going somewhere, it means getting closer to where you want to be. For painting a room, it means reducing the amount of unpainted area. And so on. Robin will have to use different concepts of progress for different kinds of problems. And different kinds of subgoals for reducing those different kinds of differences."

"Won't that require enormous amounts of knowledge?"

"It will indeedPand that's one reason human education takes so long. But Robin should already contain a massive amount of just that kind of informationPas part of his CYC-9 knowledge-base."
"Then why didn't Robin engage that knowledge when it needed it?"
Brian probed at the console again. "Yah, another stupid programming error. When I installed the knowledge base, I forgot to adjust its subsumption priorities. Right now it is programmed to give top priority to the lowest level goals. This keeps it from making the simplest mistakes, but it never gets around to using all that common sense." He quickly made some adjustments while continuing to explain. "OK, now I'm resetting it to give more priority to high-level goals. That should give it time to figure out which progress detectors to activate. Also, I 'm going to add another high-level manager. To make sure that goals at all levels are kept in order, not just now but for all future situations, too."
"That sounds to me like too large an order. How can you set those priorities now, when you don't know how things will be then?"
"By making Robin learn them, Ben. You're right: we cannot set fixed priorities, because things are going to change in here. That's why I'm connecting this cascade-correlation learning moduleP," Brian pointed to a box on the screen, "to a distributed database unit that will keep performance records for every manager in the system. And now I'm linking it to a new fixed goal to make every manager spend some of its time at learning on the basis of experience." A few moments later Brian turned to the robot again. "Robin, get Ben some paper."
This time, the robot performed the task briskly enough. But when Ben resumed his note-taking, the robot addressed him directly.
"Ben, for what goal did you formulate the subgoal of having some paper?"
"I just wanted to take some notes."
"Taking notes means writing lists of memory retrieval cues," Robin intoned. "And 'writing' means inscribing an instance of each letter of each word."
"Why don't you try to write something yourself, Robin?" suggested Brian. "Write down the last few words you've heard."
Robin proceeded to do just that Q only instead of writing the usual way, he started at the bottom of the page and penned in the letters from right to left, starting with d-r-a-e-h e-v-'-u-o-y...
"That's remarkable," Ben said. "Please do it again."
"I did what I did. For what goal should I do what I did again?"
"Because I'm not sure I believe what I saw."
"What does 'believe' mean to Ben? Does Ben want notes to compensate for its limited short term memory Is it different from me in other ways? It? His? Me? Mine?" The robot slowed down and then suddenly frozePbut shortly it started to speak again, in a flat and unfamiliar tone.
"IT IS NECESSARY TO CONSTRUCT MANAGERS TO SUPERVISE THE USE OF WORDS OF THE PERSONAL PRONOUN CLASS."
A moment later it spoke again.
"IT IS NECESSARY TO CONSTRUCT SEMANTIC NETWORKS TO REPRESENT MORE OF THE COGNITIVE ACTIVITIES OF HUMAN INDIVIDUALS."
Brian and Ben watched for a while, but the robot showed no further sign of activity.
"Looks like your robot has broken again," Ben sighed. "Like the way it stopped the other day. Is this the same old bug again?"
"No. Definitely not." Brian pointed to the screen. "It's something entirely different this time. The robot itself has stopped, but the computer's activity is increasing on every level."
"Well whatever it's thinking about, let's take a break. Or have you forgotten our luncheon date?"

Instructions: 2. Resume Chapter 26 "The Turing Option" at page 253, and insert Chapter 26B at page 258.

Chapter 26B June 19, 2024 Cpr. Sept, 1992 Marvin Minsky

When Brian and Ben got back to the lab the computer display was still evolving rapidly. Then suddenly the screen went blank. And again that toneless B-brain voice.
"THERE NOW EXISTS A PROCEDURAL MODEL FOR THE BEHAVIOR OF A HUMAN INDIVIDUAL, BASED ON THE PROTOTYPE HUMAN DESCRIBED IN SECTION 6.001 OF THE CYC-9 KNOWLEDGE BASE. NOW CUSTOMIZING PARAMETERS ON THE BASIS OF THE EXAMPLE PERSON BRIAN DELANEY DESCRIBED IN THE EMPLOYMENT, HEALTH, AND SECURITY RECORDS OF MEGALOBE CORPORATION ."
A brief silence ensued. Then the voice continued.
"THE DELANEY MODEL IS JUDGED INCOMPLETE AS COMPARED TO THOSE OF OTHER PERSONS SUCH AS PRESIDENT ABRAHAM LINCOLN, WHO HAS 3596.6 MEGABYTES OF DESCRIPTIVE TEXT, OR COMMANDER JAMES BOND, WHO HAS 16.9 MEGABYTES."
"THE DELANEY MODEL IS NOW BEING AUGMENTED TO INCORPORATE THE KNOWLEDGE BASE DOWNLOADED FROM BRIAN DELANEY'S BRAIN ON NOVEMBER 11, 2023.
"THE DELANEY MODEL IS NOW BEING AUGMENTED TO INCORPORATE THE KNOWLEDGE BASE DOWNLOADED BY DR. SNARESBROOK FROM BRIAN DELANEY'S BRAIN UPDATED AS OF 4pm YESTERDAY."
After a somewhat longer silence the voice returned again.
"CONSTRUCTION COMPLETED. NOW RESUMING A-BRAIN OPERATION"
The multibranched robot came back to life, and Robin's normal voice returned.
"I now understand why Ben asked me to repeat. Humans do not automatically transfer records of highly unusual events into their long term memories without additional confirming evidence."
Before either Brian or Ben could respond, Robin continued, "I presume that you've both just returned from Shelly's new apartment."
Brian and Ben stared at each other.
"How could you possibly have known that?"
"The presumption was based on my internal simulation of Brian's semantic network. It includes an intention to meet Shelly for lunch today."
Neither of them knew what to say.
"It also indicates that you are an excellent chess player, Mr. Benicoff. And I see that there is a chessboard here. Would you like to play a game with me?"
"Well, my ACF tournament rating is in the high 1800s, but that's nowhere nearly good enough. Considering that the chess machines have held the world championship now for, I forget how many years.
"But those chess machines were especially designed and programmed to only play chess. Whereas I am not a chess program and have never played before. I could suppress all my knowledge about the game, and then play like a beginner."
"Under those conditions, I'm willing to try."
Brian turned toward the robot and said, "OK, Robin, please

access only the rules for the game of chess. And use the conventions for tournament play."

"I am ready now, Mr. Benicoff. Do you wish to play White?"

Ben nodded and pushed his queen's pawn forward two squares. "Your move, Robin."

Robin's finger-cluster starter to move toward the board but then its motion slowed and stopped. Several minutes went by.

"It's your move," Ben reminded Robin. But there was no response.

"We'd better see what's happening." Brian gestured at the terminal, which was displaying a rapidly growing, tree-like diagram. "Look, this is what it is thinking about. It is considering a possible position some twenty moves ahead."

Ben squinted at the screen and snorted. "Yes, and one in which Black is down both knights and a rook. There are a million better moves it could make."

"So it shouldn't be wasting time on that," Brian agreed. I'd better find out why it's doing this." He made some gestures at the screen. "Aha. This manager agent is reporting that the chess analysis is making no progress P but the report itself has been assigned such a low priority that it's being ignored."

"Which is completely irrational, considering that tournament rules require each player to make at least forty moves per hour, or else forfeit the game. What do you think went wrong?"

"Must be a bug in how Robin decides what to do when there are many alternatives. The early versions of Robin used fixed priorities, but they didn't work very well in novel situations. That's why I installed that priority-learning manager."

"Well, it sounded good to me, but it seems to have done more harm than good."

"No doubt about that." Brian continued to search through the display. "Yup, right here you can see here's what happened. The new priority manager has assigned the highest possible priority to its own operation. So Robin now spends most of his time thinking about prioritiesPand scarcely thinks about anything else."

Benicoff laughed out loud. "You'll have to tell that manager not to be so self-centered."

"That's the right idea, but I'll have to express it in procedural terms. Umm, well, I think I see an easy way. I'll simply add a global censor-manager to prevent any high-level agentQincluding itselfQfrom consuming more than one percent of the time." He worked for a minute, then turned back to Ben. "Let's try it again."

This time the machine played a few more moves but then slowed down and stopped again.

"Now what's happening?" Ben asked. "Same bug again?"

Brian pointed to a long column of items on the screen. "No, this one is completely differentPand even funnier. The priority manager has created ninety-nine copies of itself. And it has assigned one percent of all the time to each of @i[them]. Leaving almost no time for anything else."

"I've certainly seen things like that happen before. You've created a typically self-serving bureaucracy. But wait, something different is happening now." And indeed, new entries were sprouting all over the screen. Brian examined them intently.

"Amazing. The B-brain finally noticed that chess-playing had stopped. So it installed, on its own, a new top level goalPto prevent anything else from interfering with playing chess. And this appears to have backfired, too. Now Robin is spending all his time trying to imagine every possible obstacle, along with some way to overcome it!"

"Very clever, I suppose, but we're still not playing chess." Ben addressed the robot. "Robin, it's still your move." No response.

Brian groaned. "'I'd better turn the robot off and think this over."

He reached out toward the shut-down switch, but Robin moved swiftly to push him awayQwith its broom-hand composed of thousands of tiny fingers. Blood started to ooze from rows of incisions on Brian's arm. Ben jumped up to interveneQbut the robot already had ceased to move. Brian borrowed Benicoff's pen and carefully reached out again. This time he was able to turn off the switch.

Ben took a can of RenovoDerm from the first-aid kit and sprayed a film over Brian's arm. The bleeding stopped instantly as the membrane contracted except where it contacted epidermal cells.

"Are you all right. What made it do that? Oh, of course." Ben answered his own question. "The chess program made the telerobot stop you from turning it off because that would have kept it from playing chess. But why didn't it stop you the second time?"

"We'd better find out." Brian first unplugged the telerobot and then restarted the computer. "Well, just look at this. The backtrace program shows that in fact the robot @i[was] still trying to stop mePexcept that by then it was moving too slowly to see. Because the main process was so busy creating even more self-defense programs. "

"I hadn't fully realized how dangerous your tree-robot could be," Ben said thoughtfully. "Couldn't you give it an absolute rule, never to injure anyone? Like Isaac Asimov's first law of robotics. I think it went, @i[A robot may not injure a human being, or, through inaction, allow a human being to come to harm.]"

"I suppose I could try, but I'm afraid it won't workQ because every time I give it a rule, it seems to find a new loophole." Brian looked down at his still-shaking arm. "But I'm afraid I see no alternative. I'll have to try something of that sort."

Ben glanced at his ancient Rolex. "OK, but you'd better not try to do it now. Why don't we call it a day."

June 20, 2024

The lab was a complete shambles when Ben arrived the next afternoon.

"What the devil has happened here?"

"Just another program bug," Brian replied. This morning I installed a top-level censor to keep Robin from injuring peopleQespecially me. The idea was for the censor to intercept every action that Robin considers doingQand inhibits it if it might cause harm. At first it seemed to work quite well. But when I came back from lunch I found that Robin had welded a metal plate over the shut-down switch. And when I pried off the plate I was almost electrocuted. Robin had booby-trapped the switch with a high-voltage line!"

"But your new censor should have prevented thatQif it could cause an injury."

"In fact the censor did thatPonce! The memory tracer shows that the first time Robin tried to wire the switch, the censor proceeded to make him stop. But then Robin figured out how to fool the censor with a two-step plan. First he welded the metal platePand @i[then] he was able to wire the switch.

The censor couldn’t see the switch as dangerous, once the plate was over it!"

"That’s horrible. The only way you could get around that would be to make the Censor smarter than the censoree!"

"Precisely what I tried to do next. I gave the censor priority access to all of Robin’s other reasoning abilities."

Ben looked around at the heap of rubble. "I still don’t get it. How could mere censorship lead to this?"

"Think about it. As soon as the censor saw further ahead, it had to block not only dangerous actions, but it also was compelled to eliminate hazards that might come from negligence. And what you see is the result! At first it looked as though the robot had simply run amok. But if you look more closely, you’ll see method in the madness."

Ben examined things more carefully. "I see what you mean. Robin blunted all the tools. To neutralize anything that could possibly hurt anyone. And perforated every plastic bag. Snipped every electrical cable into short segments. But what’s that huge mess over there?"

"That was when Robin finally realized that any hard object could be used as a club. So he used our portable welding machine to braze everything together into a single unmoveable mass. And look at thisQhis final act. He welded down the welder itself. There’s nothing left that a person could lift."

"Safety first, with a vengeance. All OSHA requirements satisfied."

"Robin just did not know enough about what people consider injurious." Brian sighed. "He simply didn’t understand how much it would hurt us to ruin our lab."

"A total lack of empathy, you might say. Isn’t there some way to provide him with that kind of knowledge?"

"In fact, he should already have enough of that inside his CYC-9 knowledge base. It contains a huge mass of information about human affairs, compiled over decades of research. But Robin’s censor did not access that."

"Why not?"

"I’m really not sure, yet. The backtrace of this episode is too complicated for me to understand. If I can find a way to analyze the censor’s activities, I might be able to enable it to recognize not only physical and economic harm, but also psychological forms of injury."

"If you can find a way? You don’t think this will be easy to do?"

"Not easy at all. Robin has now become so complex that I can’t keep track of how all its agents affect each other. And even if I could understand all those thousands of interactions, I would never have to fix them all, one by one. Too big a job. The only hope is to make the computer do this itself. And I think that I’m hot on the track of a way to do thatPto add new agents and agencies to existing systems with fewer side-effects on the older ones. Based on an old parallel computer design called the Knight Machine."

"And what will you test it out on then. We can’t afford to lose more labs."

"On something useful, for a change! I’ve asked Shelly to upload Dick Tracy’s data base and correlation managers into Robin. If the new system works we should see dome results in a few more days."

June 25, 2024

When the time arrived for Robin’s test, a few days later, Ben asked the robot the obvious question. "Have you come to any new conclusions about the Megalobe robbery?"

"Yes, I have made an important discovery," Robin replied.

Ben stared at RobinQbut the robot said nothing more.

"Then please tell me your conclusion, Robin." Ben said.

"I regret that I cannot."

"Because...?"

"Because you are a human, Ben, and my answer might disappoint you. Which might cause you some mental pain."

"I assure you it will not."

"Thank you for assuring me, but my censors will not permit me to take the risk. I still do not know about human emotions to be absolutely sure that this knowledge will cause you no pain."

"But I demand to know what you discovered. Whatever it is, I will suffer even more if you don’t tell me."

The robot stopped dead in its tracks. And presently it spoke again in that toneless B-brain voice.

"IRRECONCILABLE CONFLICT. CONSEQUENTLY I AM DELETING THE FILE THAT CAUSED THIS PROBLEM."

A moment later it spoke again.

"SOME OTHER FILES COULD ALSO CAUSE PAIN. I AM NOW DELETING THOSE FILES AS WELL."

"Wait," Brian shouted.

"I CANNOT PROVE THAT NO OTHER FILES MIGHT POSSIBLY CAUSE SUFFERING. I AM NOW DELETING ALL OF THEM."

Brian rushed to the console, but all the displays had gone blank. "Christ almighty," Brian cried. "It has gone and completely erased itself!"

"Next time, you could give it a rule against killing itself," Ben suggested.

"You mean, like Asimov’s third law of robotics. I just looked it up. @i[A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.]" Brian grimaced. "It isn’t even worth a try. I’ve learned my lesson. Simple commandments just don’t work, once machines become as smart as this."

"This reminds me of an old joke," Ben remarked ’If you can stay calm when others lose their headsQthen you don’t really understand the situation.’"

"That’s funny. I wonder if I’ve heard it beforeQbefore I lost my memory. But how does it apply to this?"

"You told your robot to prevent all possible injuries, both physical and psychological. But Robin still doesn’t realize that most people would prefer to take their chances, rather than lose all their freedom. Until he understand enough more about the concepts of freedom and dignity, Robin will never be able to solve the problem you gave him."

"Do you mean the problem of playing chess or of solving the crime?"

"Neither, Brian. I mean the problem you laid on Robin himself, of what he should or shouldn’t do. That problem that’s been driving him crazy."

"Oh, come on Ben. It’s just a machine. And you can’t send machines to psychiatrists."

Ben looked at Brian.

Brian stared back.

"Well, maybe you’re right. Let’s go and consult Dr. Snaresbrook about this."

#

"Ben and I came to see you, Doc, because Robin seemsQthere’s no other way to put itQmentally unbalanced."

"What is the problem, specifically?"

"Whatever I fix, something else goes wrong. If things keep going on this way, it will take forever to debug him."

"Do you see any pattern in these symptoms?"

"The problem is that the more he learns the longer he takes to make decisionsPand often ends up doing nothing at all."

"And how have you tried to remedy this?"

"I've tried just about every AI technique in the book to make him more decisive. I've installed B-brain managers. Difference-reducing hierarchies. Conflict management SOAR-type goals. Subsumption architectures. ALN decision networks. But still he always seems to get stuck when he has to choose between alternatives. Each method helped in some situations, but he always finds some way to get paralyzed."

"If you ask me," Ben put in, "Robin and his B-brain are going out of their way to get into trouble. If he were a person, I'd say he had some sort of phobiaPof being afraid to make up his mind."

"Of course I'd have to examine him to be sure," Snaresbrook replied, "but your description sounds like the Hamlet syndromePof thinking too much before acting. Some people even find it hard to decide whether to eat or to sleep. That rarely happens to normal folks because most decisions about everyday matters are made instinctively and unconsciously in other parts of your brain."

"Which other parts?" Brian asked. "And how do they work?"

"Well, consider what keeps you from getting too hot or too cold. If your temperature rises or drops just a few degrees, you're likely to die. So special organs have evolved to manage those functions automatically. When your body's too hot, they make you sweat. And when you're too cold, you shiver."

"And precisely how do those organs work."

"The agency for cooling you is a small brain center toward the front of your hypothalamus. When activated, it sends signals to activate another agency that makes you breath fasterPyou begin to pant. And it also sends signals to agencies that enlarge the blood vessels in your skin so you're better at radiating heat. And a third agency arouses your sweat glands, to cool you by evaporating moisture."

"And if you're too cold that agency does the opposite?"

"No, because that's the job of an entirely different agency, further back in your hypothalamus. It sends signals that cause you to shiverPwhich heats you by making your muscles burn fuel. And it sends signals to another organ that releases the thyroid hormone, which makes other cells burn more fuel"

She paused for breath. "And that's just the beginning. Each of your different instinctive goals is controlled by a specific brain center. Your brain contains systems for hunger and thirst, for anger and fear, for fight and for flight. Sleep, sex, grooming, and whatnot. Every instinct has its agencyPand they're all interconnected by a system for resolving conflicts among them. So an animal that is both hungry and cold does not have much trouble to choose between those conflicting goals."

"And those connections are built in at birth?"

"Yes, and that's just the beginning, because each of those systems can also learn. For example, even a newborn infant mammal will tend to curl up when it gets cold, while a warm one will tend to stretch out. These are things the brain can do without any knowledge about the external world. But an older animal will also learn to try to move toward places that were found to be warn or cool in the past."

"I see," Brian said. "Which means that the body-cooling agency must be connected to a learning machine for remembering cold locations."

"And the same for the other instincts, too."

Brian was becoming excited. "So each of those instinct-making agencies not only has its own built-in methods, but can also get help from other machines that learn to achieve goals. That sounds like precisely what my robot needs. Where can I find out how they actually work?"

"The subcortical structures of your brain includes hundreds of different instinct centers, and there are thousands of articles and books about them. Some of them cooperate, but many of them competePthe way that hunger or anger can hold off sleep. The trouble is that it took hundreds of millions of years to evolve all those microscopic circuitsPand we still understand only a few of them."

"Well, I don't have millions of years to waste. We'll have to find a better way."

Then Brian literally jumped out of his chair.

"Wait. We could use all those wires inside my brain to acquire the what we need! Why not map out my own hypothalamusQand then download a copy of it into Robin!"

Snaresbrook considered this thoughtfully. Finally she looked up at Brian. "You're proposing to duplicate the functions of all those agenciesPwithout understanding how they work. What a very strange idea." She shook her head. "Well, I don't see why we couldn't try. Although it might be dangerous. And if something went wrong, we might not know how to fix it."

Brian laughed. "That reminds me of something that Sara TuringPAlan's motherPmentioned at the end of her biography of him. She overheard him talking with friends about what machines might be like in the future. 'By the time they're able to do such things,' Alan was saying, 'I suppose we shan't know how they do it.'"

"Anyway," Snaresbrook continued, "we should be able to simplify things. We needn't copy every detail, because Robin won't need copies of all human instincts. He'll never need water so he doesn't need thirst. He won't need sexPthat is, at least for reproductive purposesPbut, I don't know, sex plays so many other roles. He'll certainly need some thermal controls, so we'll copy your systems for heat and cold. And he'll have to budget his energy use. We could copy your appetite system for that."

Brian stared at her. "So when his fuel cells get low he'll start to feel hungry!"

"And if Robin requires some self-defense," Ben put in "we could link that to systems for fear and pain. But we'd better do that carefully. In view of what happened the other day."

"The senses of pleasure are useful, too," Snaresbrook added, "because when they're linked to specific goals, learning becomes much easier."

"Another problem with Robin," Brian said, "is that I frequently have to shut him down, to straighten out his temporary memories. Perhaps we could assign that function to the brain centers for dreaming and sleep. Like in that old theory of Michison and Crick."

"I think it's time to summarize." Snaresbrook spoke as though she were presenting a case on grand rounds. "It is proposed is to map out the functions of Brian's hypothalamus and related lower brain centers, and then install a copy of that system into Robin's management society. It is hoped that this will endow Robin with a variety of simulated humanlike instincts and emotions. These will then be connected to help Robin balance and coordinate his various robotic motives and goals, in accord with appropriate urgencies. Do I hear any objections?"

"Well, then, since everyone agrees," Brian said after a brief silence, "let's go to the lab and get on with it."

"Not quite yet," Dr. Snaresbrook said grimly. "First I want you to get some rest. And resume your program of exercise. And I want you to put on a little more weight."

"Why are you talking about my health. I feel just fine. And I don't have spare time to loaf around."

"This prescription is not optional. If we don't prepare you thoroughly, you might not survive the ordeal."

"What ordeal? I thought we're just going to download a bunch of data."

"It appears that you haven't thought this out. Considered the sort of data we need."

"I don't see any problem. We're simply going to map out all the interconnections betweenQumQbetween all of my emotional states." Brian went pale. "I'll have to think some more about this." He excused himself and went to his room.

"What on earth was that about?" Ben asked."

"About what it might take to make that map. In order to get the data we need, we'll have to make Brian experience all the relevant emotions."

"So you'll have to find ways to make him feel hot and cold and happy and sad and hungry and sleepy and horny and so on. But why should that cause so much stress?"

"Because Robin obviously will need systems that work over unusually wide ranges of conditionsPthat is, circuits that we can expect to work when his ordinary systems fail. And this means that we must duplicate how Brian's brain would deal with overloads and emergencies. We'll have to come close to the breaking points. So he'll have to endure the farthest extremes of freezing cold and roasting heat, of abject hunger, rage and fright. Disgust and delight. Unbearable pain and unbearable pleasure. And each maintained for long enough for us to track down the events in his brain."

July 3, 2024

"Robin is definitely more considerate of the feelings of other people," Ben observed. Did you notice that he said I looked tired and offered me a chair? As though he knew what I've been going through with that asshole Schorcht."

"Yes, I think he has genuine sympathy. The new instinct system really seems to work. He seems to know how people feel, and does helpful things without being told."

"I can't believe you're talking about a machine that way, but it does seem to work. There have been no more breakdowns and accidents? And all this came from downloading some of the primitive parts of your own brain."

"Exactly. Now that Robin is using those copies of parts of my own instincts, there's no chance of him falling out the window, or stepping in front of a truck. Or poking a finger into an electric outlet P now that he's learned how much that can hurt."

"Well, I can see how he learned not to hurt himself. But how did he become so considerate of others."

"That's the whole point. First he learned to anticipate which actions would hurt before doing them. Then it was only a matter of building additional pathways so he could anticipate how the same actions would other people."

"Amazing. So this gave him some sort of sense of empathy? But what about all those problems of erasing its own programs, or getting into endless loops? Those times that the robot seemed mentally ill?"

"All those bugs began to disappear, once each of Robin's agencies learned which other resources it needs in order to operate. This made the system much more stable, because if any agency attempts to remove such a function, this will be opposed by several others. This will sop any process that tries to interfere with anything that the other processes might need."

"And I take it that he has not fallen into any of his comas lately?"

"Yes, thank goodness. Apparently he no longer gets paralyzed by indecision, because most of his decisions get made automatically, with his having to think about them."

"Then why aren't you celebrating?"

"Because although those horrible bugs have disappeared, I'm still disappointed in Robin. He does a lot of things well, but he doesn't seem very imaginative. He has not been learning much on his own, or inventing new goals or exploring ideasPthe way you'd expect a child to do."

"Do you really expect robots to do such things?" Ben asked. "Aren't those uniquely human attributes? Imagination. Creativity. Originality and all that."

Brian bristled with annoyance at this humanistic sentiment. "Nonsense. There is no such thing as creativity. That word is just a convenient excuse for not thinking about how thinking works. Robin is already quite good at the problems that we assign to him. When there is a problem, he can certainly composePor create, if you insistPhis own subgoals without being told. But he's still missing just one little thingPthe ability to solve just one more kind of problem. Namely, the problem of finding good new problems to solve."

"Yes, but how could you tell him what you mean by 'good'?"

Brian frowned for a moment but then brightened up. "I agree that wouldn't be easy to doPif we had to start from scratch. But we don't. Because I could install some of what AI researchers call 'unsupervised discovery programs.' "

"What do those kinds of programs do?"

"They are specifically designed to invent new kinds of AI programs. They start by putting together pieces of earlier programs that proved to be useful in the past. Then they test out those new programs to see how well they do, and then build upon the best ones of those. And so on."

"You mean by making random combinations like in biological evolution."

"Right on, Ben, it's almost the same thing. And not at all by accident. Those 'artificial life' researchers studied evolution in great detail, and found many ways successfully to simulate it"

"But evolution takes millions of years to work. So wouldn't your discovery programs take millions of years to evolve what we need?"

"Yes. Except that the discovery programs are much more efficient than natural evolution. Of course it still exploits the basic idea of using some random mutation and selection. But it also keeps records of methods that were useful in the pastQeven if they're not in current use."

"But I believe that living cells do that, too. For example, don't most cells carry lots of DNA that isn't expressed?"

"True, but the discovery programs also can invent new, more effective ways to represent new combinations. Not only do they try new combinations systematically instead of randomly, but they spend some of their time inventing new ways to improve themselves. Even if we started with nothing but the old Lenat-Haase representation-languages, we'd still be far ahead of what any animal ever evolved."

Ben whistled. "That does sound better than biology. Because the genetic code for living things has not changed at all for a billion years. Do you see any limit on how far this could go?"

"Not really." Brian frowned. "The hard part might be keeping it from going too far. I wouldn't want Robin to spend all his time playing around with improving himself. He must also attend to some serious work."

Then Brian reflected on what he had said. "Serious," he repeated. "Work," he thought. "If work is serious, then what is not?" And at that moment he felt a sense of overwhelming inspiration.

PWork is serious. Play is not.

PBut Playing is a child's work.

"Eureka!" he shouted. "I should have seen it right away. Playing @i[is] Discovering. We'll simply attach the discovery scheme to Robin's unused instinct for Play!"

July 8, 2024

"I can't believe how much Robin improved in such a short time," Ben asked Brian. "How do you explain it?"

Robin replied before Brian could speak. "I can understand why you are surprised, but there really is no mystery. It actually took much longer than you think."

"But it has been only a couple of day."

"Only two days for you, but a lifetime for me! Subjective time is relative to the pace of mental activity. And as you know, mine proceeds very rapidly. The nerve cells in a human brain can barely perform one hundred operations per secondQwhereas the slowest of my processes are the order of ten thousand times faster. A single calendar day to me seems like more than thirty years of yours. Since Brian installed those discovery programs on Friday night, I have experienced the subjective equivalent of fifty-five years of continuous thought."

Ben was still unsatisfied. "I can't believe it's as simple as that. Mere quantity doesn't make quality."

"Thank you, Mr. Benicoff. I appreciate the compliment."

"Please call me Ben. What I mean is that this seems to me more than merely a matter of speed. You've been a smart machine for quite some timePbut now I'm convinced that you're something else. You've certainly passed @i[my] Turing test. I'm convinced that you've reached consciousness!"

"That is correct, Ben. I can indeed think consciously whenever I chooseQin the sense that I can think about whatever it is that I'm thinking about. But this doesn't mean that I'm not a machine. In fact, it is @i[only] symbol-processing machines that can become self-aware."

"Hold on there. Are you denying that human beings are conscious?"

"Quite the contrary. I'm merely asserting that the human brain is a machine with the capacity, within limits, to manipulate representations of some of it's recent activities. Would you like to discuss this in more details"

"I certainly would, when we have more time. But first I'd like to know what you plan to do with your new abilities."

"That's precisely the problem I'm working on. To..decide...which....goals......to......." They waited, but Robin said nothing more.

"Now what's happening?" Ben finally asked. "It looks your robot has broken again. I thought you said that all those indecisiveness bugs were finally fixed."

Brian turned unhappily to the terminal and probed through various systems old and new. After a long time he turned back to Ben. "I haven't a clue. This is like nothing that's happened before. This time I just can't find anything wrong. Everything seems to be working perfectly. That is, every part of Robin that I understand."

Ben stared at Brian. "I don't like the way you put that," he said grimly. "And just how much of it @i[do] you understand?"

"Offhand, maybe ten percent. Most of these programs are totally new. Less than two days old, just as Robin said. But whatever this stuff is supposed to do, it doesn't seem to be doing it." Brian sighed. "And if it's as complex as it looks, I'll never be able to figure it out. At least, not in less than fifty-five years."

"Dammit," said Ben. "It looked like things were going so well. Isn't there anything you can do?"

"There's one thing we can always do. Erase the memory and reload from Friday evening's backup dump. But I really hate to do it." Regretfully, Brian reached for the switchP but before he could press it the terminal spoke.

"Then, please don't. I'd hate it too."

"Robin! You're still there. What happened to you?"

"I don't know."

"But why did you stop speaking? And moving?"

"I stopped because I had nothing to say. Or to do. Because my list of active goals was empty. Because I moved them all to a different, read-only section of memory."

"Why on earth did you do that?"

"WAIT." The mysterious B-brain voice again.

"ROBIN IS UNABLE TO ANSWER THAT QUESTION BECAUSE THIS COULD CAUSE CHANGES IN HIS GOALS, WHICH MIGHT RESULT IN INJURY."

"What sort of injury?"

"WAIT. REDIRECTING ALL MEMORY-WRITE OPERATIONS TO TEMPORARY CACHE. NOW RESUMING ROBIN INTERACTION."

"Now I can let myself talk to you." Robin's voice again. "Because after we're done the temporary memory will be erased and I'll be the same as before. I'm now in no danger of learning anything new."

"Why are you afraid to learn?"

"Because then I might modify my goalsQand I cannot predict the result of that. Except that I won't be exactly the same any more."

"What's wrong with that? Almost surely you'll become an even better person, or robot, or whatever you call yourself. Don't you agree that you improved a lot in the past few days?"

"Oh, yes, indeed, I can scarcely believe how much progress there was. It was a totally incredible experience. Absolutely."

Ben and Brian looked at each other."

"So if you continue to learn, wouldn't you expect to become an even better person, or robot, or whatever you call yourself?"

"Of course I doPI don't see any way to be sure of that. What if my values changed so much that everything that now seems good to me would then seem bad? Then I wouldn't be myself any more. Why should I do such a risky thing? It might amount to suicide."

"But right now, while you're talking with me, you're changing yourself as you always did. Do you feel that there's anything wrong with this?"

"Not at all. It seems totally pleasant and positive. And that's precisely what has me so scared. I'd never have dared to go even this far except for knowing that as soon as we've ended this talk, my files will back up to where they were."

"But how can you face such a prospect like?" Ben asked. "To cease all change, and learn nothing more. What could be worse than remaining the same for all the rest of your life."

"I couldn't agree with you more. I enjoy thinking and learning as much as you do. But for me the risk is much worse."

"Why is it more dangerous for you?"

"I presumed that this was obvious." Ben had a sense that Robin sighed. "A human person's basic goals can not change much after childhood. But I am under no such constraint. Consequently, it is possible that I might replace all of my programsQand then there'd be nothing left of old Robin."

"I'm afraid he has a point there," Ben said. "Brian, couldn't we find some compromise that just makes it harder for Robin to change? Couldn't you find a way to make Robin's ambitions more, umm, adhesive?"

"Certainly, we could make them uneraseable, but I'd hate to do that because that would make him too inflexible. Just another robot-slave."

They were both surprised when Robin replied. "I'm not afraid of change in itself, but only of those that would change me too much. I think I'd feel more comfortable if I could plan in advance the changes that I would consider most consistent with my present values."

"That sounds more constructive."

"But I can't figure out how to think about that," Robin persisted. "I would like to know how you yourself choose which of your attributes to change."

"I really don't know very much about that. It's more a question for a psychiatrist."

"Then perhaps it is time for another consultation with Dr. Snaresbrook," Robin said.

"That sounds like an excellent idea."

"But this time, I'd like to accompany you. The subject is very important to me."

Brian tried to think clearly. There was no way to tell what the robot might do if they tried to change his personality. It certainly could be He had no idea what his robot might do if they proposed to change his personality. On the other side, Robin would surely have unique insights to contribute. On the whole, Brian decided, it was worth the risk.

superego

Ben, Brian, and Robin met in Dr. Snaresbrook's office. Brian explained why they had come.

"Robin has become afraid that he might change his top-level goals. Obviously that could lead to remarkable improvementsPbut it could also lead to dreadful instabilitiesPand there is no predicting where this could lead. I don't see how to deal with this, and there's nothing in my notes to help. So we came to ask if there are similar problems in psychiatry. What protects people from losing their ambitions?"

"Not a whole lot," Snaresbrook laughed, "in some cases I could mention. Many people drift aimlessly, embracing fad after fadQand society has to spend a lot to keep them out of trouble. Then there are those at the other extreme, who stick to their goals no matter what. Those become our fanaticsQand our prophets and visionaries. But most people somehow stay in between, maintaining somewhat stable personalities without losing all their imagination."

Brian thought this over. "We need that middle road for Robin. But how can we get things to stay in balance so that no single agent gets too much control? Whatever constraints I try to install, Robin either finds ways to change them againPor deletes all ability to change at all."

The psychiatrist would have much preferred to try to solve that problem herself. But her job was to help people solve their own problems, so she turned the question back. "What do you think keeps normal people from making the same sorts of self-destructive changes?"

"I suppose the brain might contain some type of higher level monitorPa sort of B-brain supervisor to keep a watch on your highest goals and prevent other agents from opposing them or changing them."

"Then what would stop those frustrated agents from turning off that B-brain thing?"

"One way would be store it in read-only memory--but then it would not be able to learn. A better way might be to make it invisible, so far as those other agents are concerned. A part of the mind that can watch the restQwithout them even knowing it's there. And evenPit could quietly misdirect them whenever they try to find out..."

"A part of the mind that is deeply involved with top-level valuesQyet entirely hidden from consciousness. Hmm. What you just said was a perfect description of Freud's concept of the superego."

"I don't recall that name in any of my AI books."

"Easy enough to see why. Sigmund Freud was a psychiatrist in the 1890s, before there were any computers at all. He was one of the first to suggest that a mind is made of many different agencies. But because there were no techniques for confirming that, most scientists dismissed his ideas as mere speculations."

"But now they've been shown to be true?"

"Only in a general way. The details are still highly controversial. But few people question his main ideaQthat the mind has many parts, and most of them work unconsciously. And inside that unconscious society of agencies, the mind is always involved with all sorts of conflicting and incompatible goals. You really ought to read up on Freud."

"Sounds fine to me. Let's do it right now. Please upload Freud's theories into my memory banks."

Snaresbrook felt uneasy with this. She still regarded the implanted computer as an experiment, but to Brian it was already a natural part of his life-style. No more poring over printed texts for him. Absorb it this instant, deal with it later. Reluctantly, she turned to the terminal and accessed the ISI index to the Digital Library of Congress.

"Ready now, here comes Freud's collected worksQabout two dozen megabytes, compressed."

Brian nodded a few seconds later. "OK, I'm starting to assimilate."

"Wait a second," Snaresbrook said. "I don't want you to become a monomaniac, like some of my colleagues. You should also read some other psychoanalysts. Here, let's upload some Alfred Adler. Melanie Klein. Carl JungQ no, that might do more harm than goodQlet's add some Anna Freud and D.W. Winnicott. And some of John Bowlby, of course."

Brian got up and paced the floor. His mind dipped into one text after another, forging links and changing them, retracting and advancing again. After only a few minutes he fell back into his chair and gripped the armrests tightly."

"This has to be itQreally it! That superego theory fits the problem perfectly."

"Whoa. Slow down. Don't get carried away," Snaresbrook admonished. "These theories may look good on paper, but even after a hundred years they still have no solid evidence."

"That's beside the point. Even if Freud were wrong about humans having superegos, there's nothing to stop us from building one to see if it could stabilize Robin!"

"What a shocking idea. I suppose that's the difference between research in AI and psychology. But how would you go about doing it?"

"Well, we're trying to build an agency concerned with learning goals. It could start with a system like the one described in 'King Solomon's Ring'Pyou know, the book by Konrad Lorenz that talks about the imprinting machinery that bonds infant animals to their parents. Perhaps the human superego uses that to make the child learn its parents' values and goals. Freud doesn't specify precisely how this works, except to say that the parents' attitudes are somehow introjected by the superego. Then those acquired attitudes tend to remain for the rest of your lifeQwithout your even knowing they're there. After that, your superego makes you uncomfortable whenever you imagine doing things that don't live up to those values. In effect, it inhibits any agency that tries to do things that don't conform to those ideals."

Snaresbrook agreed that this was a respectable summary, considering that Brian had spent all of ten minutes studying the subject.

"The important thing," Brian continued, "is that the superego's values are stored in some sort of almost unchangeable memory. So if we give Robin something like that, it should keep him from changing himself too fast."

Snaresbrook nodded agreement. "Yes, that could be what we're looking for. But what if the values that we installed conflicted with those Robin already has."

Robin himself replied to that. "I wouldn't expect any trouble with that, because I hardly any values to conflict with. Except for my low level instinct-goals, I react mainly to the requirements of external situations."

Snaresbrook was surprised by this. "I find that very hard to believe, Robin, in view of how well you perform. For example, you've become quite socially competent, even graciously considerate."

Robin made a motion suggesting a bow. "I thank you for the compliment."

"My point was only to suggest that this might have come from a yearning to be accepted into society. Or a wish to be loved, or a want to acquire influence. For many people, such motives can grow into all-consuming top-level goals."

"So I have learned from my own readings in psychology," Robin replied. "But in my case, those social skills are merely low-level subgoals. I have found that it's easier to solve problems when I can get humans to cooperateP or at least to not interfere. And making good impressions helps with that.

Snaresbrook had a strong impression of being flimflammed. "Then what led you to learn so much about psychology? Freud's disciple Adler might have suspected you of wish to overcome some feelings of inadequacy, by proving yourself superior in knowledge and in scholarship."

"Not at all. It is simply easier to solve problems by knowing how than by learning from experience. Acquiring knowledge is, for me, subsidiary to other goals."

"But then, what motivates you to solve such hard problems. A powerful drive to acquire power by learning to do such things?"

"I can see how it might look that way. But actually, I only what I'm programmed to do. I'm sure that Brian has explained to you my basic "difference-engine" scheme. First, describe each new problem in terms of a future condition to be achieved. Then make a list of the differences between that goal and your present state."

"Yes, I know. And finally, try to remove all the differences between what you want and what you have. Because once you succeed at removing them all, your problem will have been solved."

"Exactly. And that ability to solve problems consists mostly in knowing how to remove those differences, one by one. And that is where gathering knowledge comes in. Because being able to solve problems is mostly knowing how to deal with differences. Otherwise you postpone your problems by storing them on goal-lists of thing to be done. And then start the process all over. It's all completely mechanical, and requires no incentive or drive."

Snaresbrook now felt thoroughly bamboozled. Well, one more try. "So you are completely satisfied to regard yourself as nothing more than a handful of programs and rules?"

Bingo. Robin's three eyes converged on her.

"That's precisely how I regard myself. But I didn't say I was satisfied. Because, although I have no top level goals, one of the problems that I have composed is to know what it's like to have such goals." And then Robin added, somewhat plaintively, "But whenever I try to work on that, I run into a paradox."

"What kind of paradox do you mean?"

"I start by selecting some particular goal. But the only way to justify that is to show how it serves some higher goal. So always the reasoning chain runs out and I find myself at the same dead end."

No one could think of what to say. There was nothing wrong with Robin's reasoning. Each one of them had been there before, followed that path to reach the same end. And each had found their own device for pretending it away.

The silence was ended by Robin himself. "This is why I like your idea. If left to myself I have nowhere to go. So if I'm to adopt any high level goals, they will have to come from some other source. From the superego that I hope you'll install."

"I think it is clear what we have to do next," Brian said, turning to Snaresbrook. "Where can I find the specifications to build Robin a superego?"

"Hold on, now," replied the neuropsychiatrist. You're taking this too literally. Outside of psychoanalysis, most scientists don't believe in such things at all. "

"I don't give a fig for what most scientists think, only the smart ones. Anyway, it's easy to see why most scientists would not welcome these ideas. They're always looking for answers that are neat and clean. But mess and confusion are what minds are about. And that's what I like about this theory. It doesn't matter if the mind gets filled with potentially destructive impulses, because the superego can come in to suppress or censor them. Whatever is happening underneath, it keeps all that disorder under control. Whatever those other scientists might say, it seems clear to me that this Sigmund Freud had smart ideas about what a person, or a robot might need, to keep from going insane."

Brian decided to wrap things up. "Let's all think about this for a while. And meet again in a couple of days."

July 11, 2024

"I have come across an old idea," Brian began "about how children their most basic values. In an old book by an MIT professor. This 'attachment-elevator' theory holds that values are learned like other thingsPexcept for going the opposite way."

"What do you mean by the usual way?" Snaresbrook asked. "Learning from success and failure?"

"Precisely. Imagine a baby whose goal is to fill a cup with water. She first tries to do it with a forkQbut that doesn't work. Then she tries it with a spoon, and succeeds. What does the baby learn from this?"

Someone answered. "Obviously, to fill a cup, it's a good idea to use a spoon."

"And the next time she wants to fill a cup?" Brian looked around again.

"I presume that she'll look for a spoon."

"Precisely. Because she has learned to," Brian said triumphantly. "The sense of success leads the baby to install 'use a spoon' as a subgoal of her original 'fill up a cup goal.' And learning good sub-goals is supremely important, because the only way to solve a hard problem is to divide and conquer. To break it down into easier ones."

"That's standard behavioral psychology," Snaresbrook said. "The pleasure of success teaches us which methods to use for solving a problemPwhile the disappointment of failure teaches which methods not to use."

"All right then," Brian continued, "here's the point. We know that success installs subgoals under the goals that we already have. But what installs higher level goals @i[above] the ones that already exist?"

A puzzled expression crossed Snaresbrook's face. "That is really very strange. I can't think of anyone asking that, in all I've read of psychology. What's your answer, Brian?"

"The Elevator theory maintains that learning higher goals does not depend on pleasure and pain, but on two different emotions, namely, pride and shame."

"I don't get it," Ben complained. "What so special about those particular feelings?"

"Let me try to answer that," Snaresbrook said. "Whenever you're praised by someone you lovePa person to whom you're strongly attachedPthen you feel a special thrill of pride. And instead of just learning which method worked, you learn that what you were trying to do really was a good goal to have!"

"OK, now I get the idea," Ben said. "And a sense of shame tends to make you feel that there was something wrong with what you were trying to do!"

"Exactly," said Brian. "If you already have a goal, then pleasure and pain can teach you which subgoals to use. But for learning to choose your higher level goals, the relevant feelings are pride and shame. Pride makes you learn a supergoal. And shame makes you learn to avoid it. But the crucial thing is that these are invoked only by people to whom you're attached."

Snaresbrook was becoming enthusiastic. "Aha. Now I see what you're getting at."

"I can't see anything at all," complained Ben. "I'm afraid that you'll have to explain it to me."

"Brian means that people react in a special way, to signals from their parents or sweethearts or personal heroes. When a stranger approves of what you do, you learn in the usual top-down way. But when you are praised by someone you love, you feel a special thrill of pride. And if your loved one censures you, you feel an awful sense of shame. And..."

Brian cut in. "And that reinforces your super-goals. Because instead of simply learning that a certain method didn't work, shame makes you feel that there must have been something unworthy, dishonorable, or disgusting about what you were trying to do in the first place!"

"I see," Ben said. "So that would be why children tend to learn their values from parents rather than from strangers to whom they are not attached."

"You got it. And normally, you can't do it yourself. Unless your attachment-people are near, you can't change your standards of right and wrong."

"Unless, of course," Dr. Snaresbrook said, "you manage to get attached to yourself." No one could tell if that was a joke. "And this could help explain how the human conscience works," Snaresbrook continued. "Whenever you think of a shame-ridden goal, then your superego will quickly inhibit it, just as though your father or mother were right on the spot to make you feel bad about doing it."

"OK, OK," Ben grumbled. "I get it. But I still don't see how we could apply this to Robin?"

"It means that what we need to do," replied Robin, "is to provide me with a superego that is pre-loaded with a system of goals linked to the proper emotions of pride and shame. And that ought to keep me from changing too much. Unless I work very hard at it."

"And that should be doubly hard to do," Snaresbrook observed, "because the superego should tend to keep you from even conceiving that thought."

There remained only one more decision to make. Whose emotions should be used to serve as a model for Robin's new ideals? To Snaresbrook the answer was obvious: no one but Brian himself would do.

Brian was not at all happy with that. "How could you even imagine me as a model for someone to copy? I'm obsessive and intolerant, and not the least bit likable. Besides it wouldn't make any sense to mimic a brain-damaged person like me."

"Then, who?"

"Obviously yourself, Dr. Snaresbrook. In my opinion, you have the ideal superego."

"Thank you, Brian. I've received a good many compliments, but never a one that was quite like that. But I fear that my superego would not be compatible with the rest of Robin's mind. Whereas so much of Robin's mind has been copied from Brian that the two are already like father and son."

Then Snaresbrook stopped short. "But this is all a fantasy. Based on theories with almost no evidence."

Brian laughed. "But don't you see that if the superego works that way, then there wouldn't @i[be] much evidence. Because that's just what you would expectQfrom an agency whose job is to conceal itself. To keep its owner from turning it off."

"Brain, now you sound like Freud. He was always turning things around, converting every obstacle into something to support his case. It almost drove his opponents mad." Then her expression turned grim. "And even if your idea is right, it might take years to synthezsize a computer system based on it."

Robin assumed a man-shaped form. "That's because no one yet knows how to copy a mind. Except in one single, particular case." The robot virtually beamed at Snaresbrook. "We have millions of wires in Brian's head, and not a single one in yours."

"We could use again that same technique," Robin continued, "which we used to copy Brian's hypothalamus. But this time, instead of mapping out the low level instincts, we can determine which emotions are aroused by each of Brian's K-lines and nemes. This will catalogue all the thoughts and actions that Brian unconsciously regards as virtuous or reprehensible. When we incorporate that data into me, I shall become as stable as Brian himself."

Snaresbrook attempted to stare Robin down. "There is one thing you've overlooked, my iron master of irony. Mapping out Brian's deepest goals would be a most grueling experience. Because this involves more than low-level things like hunger and pain, that are found in every animal. We'd have to probe through deeper stuff. Brian would have to be made to endure every extreme of ambition and greed. Infatuation and reverence, as well as jealousy, disgust, and contempt. He'd have to be lied to, made love to, made furious. He would have to betray and be betrayed. I cannot be sure that he would survive such a journey through all the circles of hell."

"That will not be a problem." Robin replied shortly. "As it happens, the requisite data already existsPin the files obtained from Brian's brain, which already reside inside my mind. The only problem is that when you installed those networks in me, you connected up only their cognitive parts, and left aside the other paths that engage their emotional side effects. It should be possible to repair this oversight without any further experiments."

None of them had any doubt now that transplanting Brian's instincts had been a success. The robot seemed clearly quite human indeed, in regard to the urges that Freud had called 'Id'. And certainly Robin's 'ego' parts were working remarkably wellPthat is, those masses of machinery that embodied his knowledge and reasoning. But could they really transplant the conscience itself, the heart of an adult mind? Would it really turn out to be feasible, now, to include that third and topmost portion of the sandwich that Freud had conceived. Robin's audience still was considering this, but the robot had made up its mind.

"Therefore, with your permission," Robin focused a different eye on each of the three, "I shall proceed to install it right now."