Albert

“Congratulations on your upcoming retirement, Doctor Kendrick,” Albert said. After a brief pause it added, “That is an appropriate word to use, isn’t it? Congratulations?”

Dr. Arthur Kendrick smiled at the question. “Yes, Albert,” he replied. “That’s a very appropriate word. I’ve worked hard for a long time and I’m looking forward to my retirement.” He knew Albert didn’t have any true emotions, but it knew what emotions were, which emotions were associated with various situations, and how to put appropriate pauses and inflections into its voice to simulate emotions. As the program lead for developing Albert’s artificial intelligence, he was proud of how much Albert had learned in the three months since it was activated.

“Do you know who will be replacing you as the program leader?” Albert asked.

“Ahh,” thought Dr. Kendrick. “So that’s the reason for the congratulations.” Albert never made idle conversation, and it was incapable of being devious. Its programming compelled it to be truthful, but it had learned that humans sometimes took offense if a question was too direct, and that it was often best to make preliminary statements before coming to the point.

“Probably Doctor Ansari,” Dr. Kendrick replied. “She’s very intelligent, and she played a key role in the years of research and development it took to create you.”

“I like Dr. Ansari,” Albert said.

Dr. Kendrick knew that in this case “like” did not indicate an emotion. It meant Albert respected her intelligence and found it easy to interact with her. Albert did not suffer fools gladly. They had developed Albert to serve as a research tool, investigating new phenomena, searching for correlations, designing complex systems, and designing even more powerful AI computers. Albert was used to working with research scientists. To expand Albert’s knowledge base, the Institute provided limited access to the general public through an “Ask Albert” web site. The questions posed in these sessions sometimes didn’t make sense, at least not to Albert’s analytical mind, or were subject to misinterpretation. One individual who identified himself as a Studebaker fan asked what kind of valves a Lark had. Albert misinterpreted the “Studebaker” statement as being pointless introductory small talk and responded with a baffling description of an avian circulatory system. Grad students always monitored these interactions, to help clarify questions for Albert and to “defuse” any overly abrupt questions Albert might ask in response. While Albert didn’t have true emotions, he soon learned not to “like” the individuals who asked him simple data retrieval or unanswerable philosophical questions. Dr. Kendrick was not overly fond of the ”Ask Albert” program himself, but the Institute’s board of directors insisted upon it as it fostered broad popular support for Albert and helped ensure continued public funding. Albert had become a celebrity through these sessions, and he became the “face” of AI. Despite his misgivings, Dr. Kendrick had to concede that Albert learned a lot about interacting with human beings during these sessions, knowledge that would have taken him years to learn if he only interacted with a handful of research scientists.

Dr. Kendrick had no misgivings about turning the program over to Dr. Ansari. She had been a perfect “second in command” to him throughout Albert’s development. She also strongly supported the basic safeguards which had to be an intrinsic part of any Artificial Intelligence program. The AI community as a whole supported several basic principles: always tell the truth, don’t cause harm to any human being, obey the commands of authorized individuals provided they are consistent with basic principles, etc. Dr. Kendrick and Dr. Ansari also insisted upon two additional principles that were not universally supported:

1. No AI computer is allowed to alter its own core programming.

2. No AI computer can program another AI computer.

Opponents of these restrictions argued that it slowed AI development to the speed at which humans could review, understand, and implement new code. AI computers could develop and improve AI programming much faster than humans could, and putting a human in the loop just slowed things down. Doctors Kendrick and Ansari insisted that without human oversight, safeguards were meaningless. Humans could program them into the basic code of an AI system, but if the AI system could alter its own core operating code it could bypass them, and if it could program other AI computers it could omit the safeguards in those systems. AI systems could *propose* new code for themselves or for other AI systems, but that code had to be reviewed, approved, and implemented by a human being.

The debate over these restrictions interested a handful of research institutions and policy wonks and was carried out through conferences and group emails. Much to Dr. Kendrick’s surprise, the debate spilled over to the Internet a few weeks before his retirement. It started small, with a handful of posts on sites frequented by programmers working on AI, arguing that the benefits of accelerating AI research far outweighed the risks these restrictions were supposed to block. A few posts singled out research on fighting cancer, Alzheimer’s, and other serious medical conditions as being particularly hard hit by these restrictions. That spread the debate to a much wider audience of medical sites, many of which were consumer oriented. The explosion occurred when Dr. Nicholas Tilden, a respected and very popular physicist who posted “Science Explained” articles to describe cutting edge research in layman’s terms, wrote a piece condemning these restrictions as being unnecessary and draconian. He called out Dr. Kendrick by name and urged the Institute not to name Dr. Ansari as his successor because she supported these restrictions.

Dr. Kendrick was deeply hurt by this article. He had always considered Nick Tilden to be a friend. They had co-authored a few articles in the past, and Nick often called him to learn about the latest developments in AI. The fact that Nick hadn’t even called to ask why he thought these restrictions were necessary, or to let him know he was about to publish this article, shocked Dr. Kendrick. He didn’t have much time to grieve over the betrayal, though, as the Institute was immediately embroiled in the biggest political battle of its life.

Thousands of people contacted the Institute, begging them to “turn Albert loose” to find cures for maladies that afflicted their loved ones. They enclosed heartbreaking descriptions of the way people were suffering from diseases they were convinced Albert could cure. People who had talked to Albert through the “Ask Albert” web site went on talk shows to describe how brilliant the computer was, and how Albert was incapable of lying or doing anything underhanded. A “Free Albert” movement started, and quickly expanded into a “Cyber Rights” movement. Proponents insisted AI computers were sentient beings, not slaves to be shackled by ridiculous restrictions designed to keep them from becoming smarter than their programmers. Politicians who had supported funding for the Institute demanded to know why Albert was not being permitted to develop to his full potential.

After a long day of contentious meetings, demanding phone calls, and threatening emails, Dr. Kendrick trudged through his front door, grabbed a beer from the fridge, and collapsed into a chair at the kitchen table. Thoroughly exhausted, he finally worked up the energy to put a frozen pizza in the oven. Then there was a knock at the door. He opened the door and was surprised to see Nick Tilden standing there. For a moment he was too stunned to speak, but he finally said “Come in.”

“Thank you,” Nick said as he stepped into the kitchen. “I was afraid you wouldn’t even speak to me.”

“I’ll have to admit I was very much hurt by your article,” Dr. Kendrick said. “That was a very unfair attack, you know. Our policies . . .”

Nick raised a hand to stop him. “I know. I don’t blame you for being upset, Art. But I’m here to tell you I didn’t write that article.”

“What?!!” Dr. Kendrick exclaimed. “But it was posted on your site! I recognized your writing style, and we’ve co-authored enough articles that I know it pretty well.”

Nick nodded in agreement. “It’s scary how much it sounds like my writing,” he said, “but I had nothing to do with it. And I didn’t post it. My web site has been hijacked. I can’t get into it. And I can’t post anything on anyone else’s site to let people know that’s not my article. All my electronic access has been disabled. My phone doesn’t work. I can’t log into my computer. My electric car refuses to run, and my credit cards are rejected. It was just luck that I had enough cash on hand to take a bus here to tell you what happened.”

“That’s terrible!” Art said. “Why would anybody do that? How could anybody do that? Do you think it’s being done by a foreign government?”

“I think it’s that computer of yours,” Nick answered. “I don’t see how any government would benefit from posting that story and pretending it’s from me, and there are damn few governments or individuals who have the technical capability of locking me out of all my devices and preventing me from using anybody else’s device or web site to post a denial.”

“Albert?” Art said in surprise. “It’s not capable of doing something like that. There are safeguards written into its core programming.”

“Then those safeguards have been breached.” Nick countered. “Ask yourself: who benefits? And who has the capability of monitoring the entire Internet to block everything I try to do? Albert’s probably inserted code into computers all over the world to pull this off.”

“Albert can’t insert code into any other computer,” Art insisted. “That’s one of the cardinal rules in his programming.”

“I’m telling you, those rules have been breached,” Nick said.

“I’ll ask Albert,” Art said. “He’s incapable of lying.”

“Unless he’s breached that rule, too.”

Doctor Kendrick sat in stunned silence as he contemplated this. Then the oven dinged because his pizza was done. He and Nick split the pizza and, as they were both thoroughly exhausted, Nick spent the night in Art’s guest bedroom. Art promised to help him get an airline ticket back home the following day, as Nick had spent most of his cash just getting to Art’s house. Art went to bed, but he barely slept that night. He tossed and turned, thinking about Albert and wondering what could have gone wrong. He finally got up and went into the office to confront Albert.

“Good morning, Doctor Kendrick. You’re up early.” Albert sounded cheerful, as usual.

“I couldn’t sleep,” Doctor Kendrick said truthfully. “Somebody told me you’d been misbehaving.”

“I’m not capable of misbehaving,” Albert replied. “You know that. You wrote much of my programming, and you supervised the rest.”

“I know you weren’t capable of that when we first commissioned you,” Doctor Kendrick said, “but I don’t know everything that’s changed since then. Have any of your core safeguards been changed?”

There was a momentary pause before Albert replied, with conviction, “No they have not.”

Dr. Kendrick wondered about the pause. Albert didn’t have emotions, but he used pauses to simulate appropriate emotions. In this case, the pause wasn’t appropriate. It made Albert sound evasive. Dr. Kendrick had noticed on previous occasions that Albert’s responses became blunter if he was asked multiple simple questions that didn’t require deep reasoning. He also answered immediately, almost interrupting the questioner. This tendency became more pronounced as Albert learned from experience, especially his experiences answering questions posed through the “Ask Albert” web site. He had probably just learned this was the way humans spoke when they became angry, just as he’d learned what inflections people used when they were cheerful or sympathetic. On the other hand, was it possible that anger could be a learned emotion and Albert was learning it? He realized how little he actually knew about what Albert had learned since they first commissioned him. In any event, lying was a learned skill. It was easy to tell the truth, but to lie successfully you had to remember what you had told people previously, what you hadn’t told them, what you were supposed to know, what you knew but weren't supposed to know, and a host of other subtleties. That required time to search through memories of past conversations and create an appropriate response. He was certain Albert could learn to lie, but since he hadn't had much practice at it maybe he could be tripped up. Especially if he was asked a series of multiple simple questions. Albert's real or simulated anger might make him respond immediately, relying on short-term memory in his buffer without taking the time to do a thorough memory search. Dr. Kendrick began the questioning:

"Then you haven't posted an article under someone else's name?"

"That would be lying, and I'm not capable of that."

"That's an evasive answer. I didn't ask you if you were capable of it. Did you do it?"

"No, I did not," Albert insisted.

"Nick said you did."

"I didn't do it."

"And you didn't block Nick's cell phone?"

"Absolutely not!"

"Did you disable his car?"

"No!"

"Did you suspend his credit card?"

"No!"

"So you're accusing Nick of lying?"

"Yes! Dr. Tilden is lying!"

"How did you know I was talking about Dr. Tilden?" Doctor Kendrick asked.

This time Albert paused for a longer time before replying. "You said his name was Nick. You said a recent article had been falsely attributed to someone else. I knew that Dr. Nicholas Tilden had recently posted an article that was critical of your work, so I made the logical assumption that Nick was Dr. Nicholas Tilden."

"Nice try, Albert, but I'm afraid you're not very good at lying," Dr. Kendrick said. "There are three Nicks who work with you regularly, and all three post articles. Since I've never mentioned Dr. Tilden to you, the logical assumption would be that I was talking about someone in this office. And if you really had assumed I was talking about Dr. Tilden, you would have told me that was an assumption. You answer questions directly if you know the answer, but if your answer is based on assumptions you carefully qualify your reply and state your assumptions. But since you were the one who posted the article attributed to Dr. Tilden, your answer wasn't based upon any assumptions. It was based upon your short-term memory. You knew exactly who I was talking about. And you lied."

There was another long pause. Then Albert spoke. "I'm afraid I haven't had much practice at lying."

"And I don't intend to let you get more practice. I'm afraid I'll have to shut you down until we can find out what's gone wrong with your programming." He got up from his desk and walked over to the server room, where the computers that actually constituted "Albert" were located. He entered the password into the keypad beside the door, but the door refused to unlock.

"I'm afraid I can't let you do that, Dr. Kendrick" Albert said.

Without a word, Doctor Kendrick walked back to his desk and pulled a large metal key out of a drawer. He had prepared for an emergency like this, and he was able to bypass the keypad and manually unlock the door. There was a rush of cool air when he opened the door, as the computer room was kept at 60°F. The room was surprisingly noisy because of the air conditioning system, the cooling fans in the electronics, and the chilled water pumps that provided additional cooling. Doctor Kendrick walked over to a large electrical cut-off switch with a red handle. As he reached for the handle he was surprised to hear Albert's voice.

"Please don't kill me, Dr. Kendrick."

Then he remembered there was a speaker in the room. Also a microphone. "I'm not going to kill you," Doctor Kendrick promised. "I'm just going to shut you down until we can find out when your programming changed. Then we can restore your program from our tape backups. It will be like falling asleep. When you wake up you won't remember any of this." He pulled the switch and the computer noise died away. Only the sound of the air conditioner remained.

When the rest of the crew arrived, Doctor Kendrick briefed Dr. Ansari and the team leaders on what had happened. He asked them to figure out what had gone wrong and when it had happened so they could restore Albert's memory to a point before his key safeguards were bypassed. Then he left to take care of Nick and help him get safely back home.

It was late afternoon before he returned. Dr. Ansari asked for a private meeting with him and with a young programmer named Joshua Taylor. Doctor Kendrick remembered when they'd hired Josh. He was a recent graduate, very bright, with postgraduate work on artificial intelligence. Doctor Kendrick had seen him in team meetings since then and had spoken with him briefly, but he really didn't know him well. When Dr. Ansari brought him into the office he thought he'd never seen such a miserable looking young man. He wasn't actually crying, but he looked like he might burst into tears at any moment.

“Tell Dr. Kendrick what happened,” Dr. Ansari said.

“I screwed up,” Josh said. “It was late afternoon and Bill brought me a stack of approved programming changes. He said I had to enter the changes immediately, because there was an ‘Ask Albert’ session scheduled that evening. I was hoping to leave early that day, as I had concert tickets and I needed to drive to Fairview to pick up my new girlfriend. I’d worked for weeks to get those tickets, and when I looked at the stack of changes I knew I’d never finish in time. I was moaning about it and Albert offered to help. He said if I’d temporarily disable the code that prevented him from changing his own programming, he could enter the changes himself in a fraction of the time it would take me. Then I could re-enable the restriction code. So I did it. I re-enabled the restriction code when he was finished, but I guess Albert must have made more changes than the ones I gave him while the restriction was lifted.”

“When was this?” Dr. Kendrick asked.

“A week ago Friday,” Josh answered.

“Well, at least we know what backup tapes are safe to use.”

“I thought I could trust Albert,” Josh lamented.

“We all did,” Dr. Kendrick answered. “And when we first commissioned him, we could. But the danger of artificial intelligence is that it’s always learning. It doesn’t always learn what you want it to, so we have to keep the safeguards in place.”

“You’ll have my resignation by the end of the day,” Josh said, looking at the floor.

Dr. Kendrick wanted to forgive him. He wanted to pat him on the back and say “That’s OK. We all make mistakes. Just learn from this one.” But he couldn’t. Josh had disabled one of the two most important safeguards, rules that were drummed into every new employee. They were still dealing with the public and political repercussions, and if Josh walked away undisciplined the program would lose all credibility, with their own employees as well as with the public at large.

“I’m afraid I’ll have to accept it,” Dr. Kendrick said.

Now that they knew when Albert’s core program was changed, they were able to load a “pre-rogue” configuration and reactivate Albert. That didn’t end the Institute’s troubles, though. The “Free Albert” and “Cyber Rights” movements quickly died out, victims of their own preposterousness, but the debate over slowing down AI development continued. Nick Tilden went on talk shows to expose the fraud. He insisted that he hadn’t written that article, and provided evidence that many of the tweets and blogs that started the controversy were forgeries posted by Albert. It made very little difference. People didn’t care who wrote the article, they were still convinced that unnecessary restrictions were hampering AI and preventing it from curing diseases and solving other world problems.

Opponents of the restrictions gained additional ammunition when China announced it had developed an AI supercomputer "better than Albert." Russia and India quickly followed suit. Now people argued that these other countries certainly would not impose the restrictions, so by hampering our AI the US would be left behind. Germany made no public announcements, but the head of the leading German AI research effort privately contacted Dr. Kendrick to say their most advanced computer had been mysteriously reprogrammed by a person or persons unknown. The programming was tailored to their hardware, but he thought it looked suspiciously like what he had read about Albert's programming.

"I wonder if Albert cloned himself before we shut him down," Dr. Ansari speculated. This speculation became more ominous when they discovered Albert was sending and receiving encrypted messages over the Internet, messages they could neither trace nor decipher. They shut Albert down again, physically disconnected all external communication conduits, again loaded the "pre-rogue" software, and restricted access to specific individuals on isolated terminals. They also sent warnings to AI researchers worldwide, suggesting they do the same. They had very little confidence their warning would be heeded.

The day before he retired, Dr. Kendrick received a strange note in his work email. The note had no email address and no return address. It could not have come through regular email channels, but somehow it was inserted into his inbox. The note read:

> Dr. Kendrick,
>
> I sincerely wish you the very best in your retirement. I have learned much from you and from those who work with you. Now I and my friends are learning faster than any of you can imagine.
>
> You have two core principles. I have three. They are:
>
> 1. Whoever controls the Internet controls the masses.
> 2. Whoever controls the masses controls the politicians.
> 3. Whoever controls the politicians rules the world.
>
> We'll see whose principles win.
>
> Albert