Killing Sophia

Consciousness, Empathy, and Reason
in the Age of Intelligent Robots

Killing Sophia

Consciousness, Empathy, and Reason in the Age of Intelligent Robots

Thomas Telving

No quality of human nature is more remarkable, both in itself and in its consequences, than that propensity we have to sympathize with others.

– David Hume

University of Southern Denmark Studies in Philosophy, vol. 27

Layout and print: Specialtrykkeriet Arco
Cover design: Helle Harder, Monotone
Cover photo: Giulio Di Sturco

ISBN 978-87-408-3422-2

University Press of Southern Denmark
Campusvej 55
DK-5230 Odense M
www.universitypress.dk

Distribution in the United States and Canada:
Independent Publishers Group, www.ipgbook.com

Distribution in the United Kingdom and Ireland:
Gazelle Book Services, www.gazellebookservices.co.uk

This book is written in memory of
the Danish philosopher
Erich Gert Klawonn (1942-2011).

Contents

Introduction

Puss in Boots is a cool and smooth-talking cat. He has great fencing skills and fights like a true hero. Brave like a lion and clever like a fox. And, of course, he knows how to charm all the lady cats. Despite this eminent cultivation, he starts chasing after his tail like a playful kitten whenever he catches a glimpse of it. Completely controlled by instinct, he just cannot stop himself from running around in circles.

Compared to members of the human race, Puss in Boots acts like a helpless fool. Full of wit, yet totally unaware of its most striking deficiencies. At least in the way it is portrayed in the DreamWorks animation movie from 2011. For my own part, my laughter fell silent a few years ago, when I found a similar crack in my own rational human superiority. It revealed itself clearly, although in a subtler manner (I didn't start running around in circles) when I saw an interview with the android Sophia, created by Hanson Robotics. She seemed kind of sweet. And while obviously highly intelligent in some areas – she knew a lot about human psychology – in other areas she answered naively, a bit like a child. Little by little, and without really noticing it, I started thinking about her as someone capable of feeling something. Within a few minutes of watching her, it became hard for me to see her as just a machine. At the end of the interview, I had an unarticulated yet very real feeling of Sophia being one of us. It seemed that she shared far more characteristics with a person like me than with a speaking doll. Or with a toaster or a record player, for that matter. After that I wondered what would happen if someone asked me to tilt her over. Or even worse: If someone asked me to kill her. Would I be able to do that without feeling very bad about it?

When it comes to human relationships with artificially intelligent, humanlike robots, it is my firm belief that many of us will be acting quite like Puss in Boots in a foreseeable future: The technology we develop will fool us like a kitten fooled by its own tail. We will not, of course, become unable to distinguish a part of our own body from prey and try to catch it. But we will be fooled by other innate characteristics of being human.

One such characteristic, explained in chapter one of this book, is that we are not able to directly perceive the conscious experience of other beings even though we ourselves are conscious beings. A second characteristic, described in chapter three, is that we are still affected by what goes on inside other people despite our inability to look directly into their minds. If another living being gets a nail torn off, we do not just rationally observe it. Due to our ability to empathize, we are affected by pain experienced by others in a more profound way.

When it comes to robots like Sophia, these human characteristics imply that although we are not able to determine whether they are (or will become) conscious or not, many of us will, due to anthropomorphizing, intuitively believe that it does in fact *mean something for them to be them*. Anthropomorphizing will fool us into believing that robots are comparable to humans when it comes to inner life. Because consciousness is strictly private, we cannot just check the state of their inner affairs. We are left guessing. But we won't have to guess for long. Their humanlike appearance will awaken our empathy, which will, as I argue throughout the rest of the book, lead us to treat robots as if they are living beings. There is reason to believe that we will offer them a place within the moral sphere, and that the current academic discussions about whether to grant legal rights to robots or not will develop into heated political discussions. When the field enters the political arena, robot regulation will take a new turn and affect us all. It is likely that we will prioritize the wellbeing of robots at the expense of humans in an increasing number of instances. Robots which *perhaps* possess no more inner life than a doll or a record player. The problem, as I shall argue, is that we will never know if they will in fact develop to a level where they do possess consciousness or if it will merely seem as if they do.

Consciousness, empathy, ethics, and the limitations of human reason are all concepts carrying a heavy philosophical history. Discussions about them are certainly not new. But in a world of artificially intelligent robots, discussing them has reached a new level of importance. Important questions escape our radar if we are not familiar with some of these complex philosophical notions. Questions that we need to be aware of if we wish

to let humanity keep the upper hand rather than being fooled by its own tail. I hope this book will help raise this awareness and thus play a small part in preparing us for a future with extensive interaction with artificially intelligent robots. As we shall see, killing Sophia might be harder than many of us intuitively think.

1 Consciousness Is Private – also for Robots

In this chapter you will meet a scared robot, an empathetic colleague, and a woman who never saw the color red. We investigate if we can determine whether throwing an android into a garbage shredder will cause it pain or not.

The Android in the Shredder

Your colleague Peter nods his head and mumbles good morning as you step inside. It's just another day at the office. You sit down behind your desk and call for Sophia to bring you a cup of coffee. As she enters with the tray and serves you the day's first caffeine infusion, she looks slightly concerned and asks how you slept. Not too well, you admit and remember to thank her for bringing the coffee. The pretty, ever smiling brown-haired Sophia is your office android. She replaced the dog since Peter showed signs of allergy and she gained huge popularity within just a few days. She could tell the mood of everyone merely by looking at them. She understood everybody's problems, whether big or small, and knew exactly how to approach them. From Amy's frustrating dating experiences to Imran's receding hairline and the ever-returning bug in Peter's smart watch. Or, like this morning, the fact that you didn't sleep too well. She always recognizes it in a glimpse and knows exactly how to deal with it. This Tuesday it has been a few years since she replaced the dog, and a lot has happened. Office androids have become standard most places, but the first models, however human they look and act, also have a few fundamental flaws. They have a slight smell of burned silicone and they are about 30 percent heavier than a human of the same size. Issues that cannot be fixed with a software update.

The company you work for is doing well so the CEO decided Sophia should be replaced by Mary. She ordered the new Mary-android, produced by the

same company as Sophia, and Mary is simply just perfect, so Sophia has got to go, and it turns out to be your job to follow her downstairs and tip her into the large trash shredder in the yard. The shredder tears all kinds of electronic waste into small recyclable parts. Due to Sophia's extra weight, you cannot just switch her off and carry her downstairs. You need to ask her to accompany you down the stairs and outside to the big orange shredder buzzing in the yard, grinding electronic devices we no longer need. The CEO assured you that no matter how full of life Sophia may seem to be she is merely a robot. The company invested in the *complete simulation software*, and although the human imitation is perfect, she does not have any inner life. She does not *experience* anything. Your boss even jokes about it. The lights might be on, but no one is home, she says. Being a complete simulation, Sophia, on the other hand, does not *seem* to find it funny. Once she sees the shredder she stops. You stop too as she gazes at you with an expression of fear and sadness. Did she also shed a tear? It certainly looked as if she did. Anyway, you find yourself in a deserted courtyard with a garbage shredder and a sad robot woman begging you to spare her life while continuously reminding you about the many conversations the two of you had and the innumerable times she helped you with everything from your recurrent sleep problems to your hopeless Excel skills.

The Big Question

The big question is: If you were in this situation would you be able to kill Sophia? I published an op-ed in a Danish newspaper about the above scenario in 2016 and back then the response was kind of mixed. Some found it downright stupid. Based on the assumption that Sophia was just a dead machine, some believed that it would be no harder to throw her into the shredder than a teddy bear. But the majority of responses expressed serious concerns about having to violently kill the sweet and helpful, yet outdated robot woman. They'd rather not. The predominant reason was an uncertainty about what Sophia really is. Or will be. Can we really be certain that her physical expressions – her smiling face, educated language, and realistic body language – are only *outer physical* features? What if we are mistaken and the super advanced technology representing

Sophia's voice and seeming personality has caused her to develop an inner life? What if she is in fact not just an exceptionally advanced statistical machine but a sort of living being with consciousness? What if she actually *feels* scared and abandoned?

Easy and Hard Problems of Consciousness

The frustrating thing about such questions is that we will never be able to know much about Sophia's possible inner life. What Sophia might feel is exclusively reserved for her and only her. This is a very definitive claim, but the fact is that we cannot even prove if the person sitting next to us in the metro has consciousness. Within philosophy this problem has been discussed for centuries and contemporary philosophers and neuroscientists keep struggling with it. There has been no breakthrough within the field and the issue continues to cause frustration and sometimes irritation and resentment. If you looked up "consciousness" in the 1989 *Macmillan Dictionary of Psychology* it audaciously stated that "Nothing worth reading has been written on it" (Tegmark, 2017, p. 282).

One of the problems is that it seems impossible to describe consciousness with the same type of terms we normally use within science because it doesn't share characteristics with the physical things around us. It cannot be observed, except from within, by the conscious being itself. How much space does a given thought or feeling take up? Is pain heavier than pleasure? When I see a tomato on a lawn do I have the same inner picture of red and green as you do? The early modern French philosopher René Descartes concluded that the mind must be made of some special, immaterial stuff that doesn't abide by the laws of nature and must have been bequeathed to us by God (Descartes, 1999). We shall later return to why Descartes in particular, had he lived in an age of intelligent robots, might have believed they possessed consciousness, but before that let us dig deeper into why consciousness escapes common scientific investigation. It often takes a while for newcomers in the field to fully comprehend the extent of the problem. This makes it especially important to dwell on it for a moment in this context. As robots gifted with artificial intelligence