



*The Ethics of AI and Robotics*

---

A BUDDHIST VIEWPOINT

Soraj Hongladarom

Published by Lexington Books  
An imprint of The Rowman & Littlefield Publishing Group, Inc.  
4501 Forbes Boulevard, Suite 200, Lanham, Maryland 20706  
www.rowman.com

6 Tinworth Street, London SE11 5AL, United Kingdom

Copyright © 2020 by The Rowman & Littlefield Publishing Group, Inc.

*All rights reserved.* No part of this book may be reproduced in any form or by any electronic or mechanical means, including information storage and retrieval systems, without written permission from the publisher, except by a reviewer who may quote passages in a review.

British Library Cataloguing in Publication Information Available

### **Library of Congress Cataloging-in-Publication Data**

Names: Soraj Hongladarom, 1962- author.

Title: The ethics of AI and robotics : a Buddhist viewpoint / Soraj Hongladarom.

Description: Lanham : Lexington Books, [2020] | Includes bibliographical references and index. | Summary: "This book presents a Buddhism-inspired contribution to the ethics of AI and robotics, and the idea that a possible norm for technology must be guided by the standard of "machine enlightenment" informed by a combination of ethical and technical excellences"—Provided by publisher.


Identifiers: LCCN 2020014389 (print) | LCCN 2020014390 (ebook) | ISBN 9781498597296 (cloth) | ISBN 9781498597302 (epub)

Subjects: LCSH: Robots—Moral and ethical aspects. | Artificial intelligence—Moral and ethical aspects. | Robots—Religious aspects—Buddhism. | Artificial intelligence—Religious aspects—Buddhism. | Buddhist precepts.

Classification: LCC TJ211.49 .S68 2020 (print) | LCC TJ211.49 (ebook) | DDC 174/.90063—dc23

LC record available at <https://lcn.loc.gov/2020014389>

LC ebook record available at <https://lcn.loc.gov/2020014390>

 The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI/NISO Z39.48-1992.

# Contents

<b>1</b>	Introduction	1
<b>PART I</b>		
<b>2</b>	A Buddhist View on Nature and Personhood	19
<b>3</b>	Can Robots Be Persons?	43
<b>4</b>	Machine Enlightenment	67
<b>PART II</b>		
<b>5</b>	Autonomous Technology	109
<b>6</b>	Privacy, Machine Learning, and Big Data Analytics	143
<b>PART III</b>		
<b>7</b>	AI for Social Justice and Equality	181
	Bibliography	213
	Index	223
	About the Author	229



## *Chapter One*

# **Introduction**

Today it is scarcely possible to pass a day without coming across news related to artificial intelligence (AI) and its amazing ability to do things that we could never have thought to be possible for machines only a few years before. Ever since DeepMind's AlphaGo defeated the world champion Lee Sedol in 2016, the world has been abuzz with talks and reports about AI as well as its potentials and its threats. AI has promised to accomplish feats such as solving the global climate change problem, driving cars on its own, writing up text summaries, serving as a judge, writing poetry, composing music, designing buildings, becoming friendly companions, becoming sexual partners, acting as pets, serving as bank tellers, diagnosing onsets of cancer, buying and selling stocks—the list goes on and on. Now almost everyone owns a smart phone, and the engine behind many apps on your phone, such as Siri or Google Assistant, and many others is indeed driven by AI. The software can be found not only on the smartphone, however. Coupled with Big Data, a way to collect and manipulate a huge amount of data, these new ways of running artificial intelligence is taking over the world by storm, changing the very face of the world as we know it very rapidly.

However, AI does not come only with the good things that were mentioned above. It poses many threats too, and some of these are so powerful that some of them could before too long affect our very survival as a species. AI has been criticized, apart from destroying humanity, as a threat to our dignity and privacy rights, and many fear, with good reasons, that the technology will result in millions of people losing their jobs, causing untold disruptions in the way people work and live all over the world. As for the existential threat, Elon Musk is well known to be on the record as an advocate to the view that AI is posing a real threat to human beings' own survival. In an interview with Jack Ma, he says that "Humanity is a kind of biological boot loader for

AI.”<sup>1</sup> It is the action of human beings, when they design and employ AI on a large scale, that in effect “boots up” the software which, for Musk, will lead to the end of human beings as AI replaces us in every aspect of our lives, rendering us redundant. It is in fact instructive that the article that reports Musk’s saying that we are AI’s boot loaders in fact reports on a conversation Musk is having with Jack Ma, the famed Chinese entrepreneur. Ma takes a far more optimistic outlook on AI than Musk, and according to him, humans are resourceful enough to find their solutions to whatever situation we find ourselves in. We have faced the threats to our survival many times before in the past, but each time we made it through (otherwise all of us would not be here thinking about AI), so why can’t we make it through this time around? Instead of looking at AI as a threat, we should instead look at it more as a tool that we can use to advance our own agenda and preferences.<sup>2</sup>

It is also instructive to find out that Musk and Ma represent two of the most powerful national forces on AI today. Although Musk was born in South Africa, he works in the United States, and his business corporations operate in the United States. Ma, on the contrary, comes from China, and is among the most well-known figure from China today. And we have just seen that they have diametrically opposite views on AI. This is not to say that everyone from the United States is opposed to AI—at least the people working at Google don’t seem to fear AI that much, and Mark Zuckerberg, the CEO of Facebook, said that Musk’s remarks are “pretty irresponsible”<sup>3</sup>—but it seems to show a level of difference in attitude toward the technology which could perhaps be understood through where they work and where they come from. Musk represents the more cautious and critical stance usually adopted in the West: Individuals are given priority in the sense that their interests come first, and AI could endanger these interests if it is not put on a leash. In China, on the contrary, Jack Ma represents an opposite point of view. Instead of looking at AI as a threat, Ma views it more as a creator of opportunities for the people, downplaying the risks that Musk is so worried about. We seem, then, to have two examples of different attitudes toward AI, which perhaps could be based on different cultures. Musk stands for the more cautious approach of the West, and Ma for the more open one in the East. Of course, this is a very rough generalization, and there are always exceptions. My point here is only that it is interesting to note that Musk and Ma come from two of the world powers of AI at this moment, and they have very different attitudes.

In any case, however, believing blindly in the benefits of AI is perhaps not the best way to approach the technology. Even though Musk’s fear that AI will destroy us may be years away for now, there are still many causes of concern about the current state of AI that we should start thinking seriously about how to respond to their challenges. For example, AI is used in technologies

that collect and manipulate personal data, and these data can be used in such a way that violate the rights to privacy and dignity that should be entitled to everyone. Computers are now able to single out individual faces out of millions; this can result in a surveillance state where the gap in power between the authorities and the people becomes even wider. The huge amount of data that users of smart phones generate everyday can also result in business corporations becoming tremendously powerful, as they can manipulate these data in such a way that they will be able to control what we think or believe. These are scary situations, but they are real, and more importantly they do not lie in the future as Musk's superintelligent beings are. These risks posed by AI prompt many in the past few years to think very seriously about its ethics. What should be the ethics of artificial intelligence? If we were to draw up a set of guidelines or regulations that AI developers and everyone involved must abide by, what should such guidelines look like? How can we find a guideline that everyone in the world agrees upon? The last question poses a serious challenge because guidelines are governed by ethical norms, and the latter depend essentially on philosophical theories and assumptions, which are mostly based in long history and particular cultural traditions.

In fact the serious threat of AI has actually resulted in many groups around the world coming up with their own sets and theories, so much so that the website [Algorithmwatch.org](http://Algorithmwatch.org) lists more than eighty AI ethical guidelines from around the world,<sup>4</sup> and it would not be surprising if the list continues to grow as the reader checks out this website. Having many guidelines is not a bad thing; instead it shows how much the global community is concerned with the need for some kind of ethical guidelines and regulation of the technology. However, glancing over these guidelines, one is struck by the fact that so few of them focus in any substantial way on the intellectual traditions coming from elsewhere other than the West. In the list prepared by [Algorithmwatch.org](http://Algorithmwatch.org), most guidelines come from Western countries or international organizations, and there are only two non-Western countries which have published their own guidelines, namely China and Japan. Nonetheless, the documents prepared by these two countries do not mention anything in terms of their own intellectual and religious traditions. The documents coming from both China and Japan talk about well-known concepts such as privacy, inclusivity, fairness, and justice, but the way they argue why these concepts are important do not mention their own intellectual resources at all. This is surprising, and shows a wide lacuna in our, global, deliberation on AI ethics that is ethically and theoretically informed.

In the past few months, I have had the good fortune to work with the wonderful group of people such as John Havens, Jared Bielby, Rachel Fischer, and others who came up with the first edition of *Ethically Aligned Design*

(<https://ethicsinaction.ieee.org/>), which is part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.<sup>5</sup> The part I was involved with was called “Classical Ethics,” and it was a part where we looked at theories from various cultural and philosophical traditions to find bases for the IEEE guideline. What is most interesting in *Ethically Aligned Design* is that it is perhaps the only guideline available today that has parts dealing specifically with ethical traditions which are not based in the West, such as Confucianism, Ubuntu, and Buddhism. This is doubly interesting considering that guidelines coming from the East themselves, as I have mentioned, do not seem to include resources from their own intellectual traditions.

The upshot, then, is that from among the more than eighty guidelines on AI only one has anything substantial to say based on intellectual traditions of the non-West. This is a gap in our understanding that I would like to fill in. What I plan to do in this book—that is, what I plan to do to fill in the gap—is to present a Buddhist view on the ethics of artificial intelligence which is more substantial than is possible in *Ethically Aligned Design*. Furthermore, my purpose is not only to present that Buddhism has something interesting to say on AI ethics, but I would like to show that Buddhism offers a theory, a way of thinking about AI ethics, about how to come up with workable guidelines, which is more tenable, philosophically speaking, than the competing theories that are already available. The idea is not just to show that there are a number of theories lying around that those in the AI business can pick up and choose, but that Buddhism offers a robust theory that can provide a way of understanding how to formulate an ethics of something such as AI in a coherent and effective manner. Spelling out how to achieve this will comprise the bulk of the book that you are about to read.

Offering the Buddhist viewpoint, then, is more a way of contributing to the ongoing discussion on which theory should be the best one for AI ethics, and not only to show that there is another one available in the market.<sup>6</sup> So the question is: What is there in Buddhism to justify such a claim? What is so unique in Buddhism that, as I argue, proposes an effective solution to the problem of how to think about AI ethics most effectively? I can only offer a brief sketch of an answer in this introductory chapter.

What is unique in Buddhism is that it offers a way of thinking about the ethics of AI in a way that combines technical excellence and ethical excellence together, as well as providing a model of ethical perfection which can be used as a guidepost for formulating ethical guidelines which are closely aligned to the natural world. Both of these claims require a lot of unpacking, and it is in fact the job of this book to do that. In summary, however, the idea of combining technical and ethical excellences is based on an analysis of the word “good” in English (and indeed in many other languages). When



function well in the end, and it can become very, very dangerous too when the AI they develop become able to think on their own and become, as Nick Bostrom says, “superintelligent.”<sup>78</sup> Without proper guidance from the ethical development, the superintelligent AI will act as a very powerful, but blind, agent capable of inflicting untold harm not only on the environment, but also on themselves. This shows that they are not actually superintelligent after all because no intelligent creature acts in a way that harms themselves.

I also argue that machines, or AI, that achieve the state of ethical perfection has achieved the state I call “machine enlightenment.” In Buddhism, Enlightenment is the highest state that a human being can achieve, and in principle everyone can achieve it through practice and knowledge. Ethical considerations are all geared toward helping the practitioner achieve this goal eventually. Thus, an action is good just in case doing it contributes to the doer getting closer to the goal, and is bad just in case otherwise. This is the state where one is totally free from all the defilements (*kilesa*) that tie one down, preventing one from attaining the goal. The three main defilements are greed (*lobha*), anger (*dosa*), and delusion (*moha*). The idea is that this state of perfection is the same for everybody because being totally free from all the defilements is the only way to achieve true and lasting happiness, something that every sentient being aspires to, or something that lies ultimately in the interests of any sentient being. It might be completely incongruous to say that machines can become enlightened in this way, but if we consider that machines in the future are supposed to surpass human intelligence, then it should not come as a surprise to find machines which realize at least some of the components of enlightenment, for a main component of the Buddhist Enlightenment is full understanding of how things are. Since the AGI machines are supposed to know more and think better than we do, these beings should be able to achieve Enlightenment more easily, and they should also be able to get rid of the defilements better than us too. Moreover, in a case of artificial specialized intelligence (ASI), which has not achieved the state of full parity with human intelligence, it would be ethical just in case it is progressing along the path toward Enlightenment, just like an ordinary human practitioner who becomes ethical when she is cultivating herself and is practicing the virtues along her path.

All this pertains also to the less developed AI that we have today, namely blind algorithms that operate along machine learning or deep learning programs. Although these narrower machines are not conscious, they can become ethical in the sense described earlier with the ethical car. To say that ASI is ethical is only a way of speaking, a *façon de parler*. If someone insists, we can say that an ASI device, strictly speaking, is neither ethical nor unethical because it does not know what it is doing.

Only its programmer or manufacturer does. But that is all right. We can say either the ASI, or its manufacturer being ethical or unethical, and when I say that an ASI is ethical, then this should be taken as a shorthand for saying that its creator or designer is.<sup>9</sup> In the case of the ASI, we can also say that it does progress along that path leading to Machine Enlightenment or not, and this is useful as a benchmark against which behaviors of AI can be assessed as to whether they are ethical or not. This way of looking at things has its benefits in that it helps narrow down the possible number of what is considered to be an ethical action for an AI device. Without the Buddhist theory I am proposing here, there will be a plethora of views concerning what action should be considered ethical, a situation that we are having today with very many AI ethics guidelines and theories. What makes the Buddhist theory unique, then, is that it is based on what is natural for both human beings and AI; this is based on the identification of ethical and technical excellence described earlier. Furthermore, Buddhism has a detailed view on how to practice and cultivate oneself so as to achieve the final goal, and this can be adapted to AI ethics guidelines so that we know what should be an appropriate course to take. Details and elaborations of this argument will be offered in the book.

At this stage the reader might wonder that the Buddhist ethical theory I am proposing here sounds very much like many ancient Greek ethical theories. In fact this is indeed the case. I would also like to show that the Buddhist ethical theory has many affinities with many Hellenistic, post-Aristotelian theories, most notably Stoicism. It does have some similarities with Aristotle's virtue ethics too, but we shall see that the Buddhist theory is closer to Stoicism than to Aristotle's theory. This is because of the important role that moral luck plays in Aristotle, but not in either Stoicism or in Buddhism. Nonetheless, this book is an exploration in Buddhist ethical theory for AI, and not a scholarly investigation in the detailed similarities and differences as they exist between Buddhism and Stoicism; hence, the discussion on the comparison can only be superficial at best. Nonetheless, a very brief sketch of some similarities can be given here. Both Buddhism and Stoicism share the same kind of overriding goal: for Buddhism it is to become an *arahant*, someone who is totally free from the defilements, and for Stoicism, it is to become the Stoic stage, that is, someone who is "never impeded, who is infallible, who is more powerful than everyone else, richer, stronger, freer, happier and the only person truly deserving the title 'king.'"<sup>10</sup> The ideas are roughly the same, and what is important is also that in both traditions it is extremely difficult, if not all together impossible, to become either an *arahant* or a Stoic sage. In our case it is more useful to say that the *arahant* or the Stoic sage represent an ideal, a guidepost, whose presence tells the practitioner where to go and how to practice, and which represent a yardstick with which to tell which action is

right or wrong. As for the dissimilarities, Buddhism has a much more detailed prescription for the practitioner than Stoicism ever does. This might be due to the fact that many Stoic texts have been lost through time, while Buddhism has maintained its status as a living and thriving tradition ever since the demise of its founder more than 2,400 years ago. Another important difference is that Buddhism, unique among the world's religions and philosophical traditions, famously teaches that the self is only a construction, in a way that a rainbow is a construction that the eyes see out of reflections of sunlight and droplets of water. This is known as the Doctrine of Emptiness, and there is nothing like it in Stoicism nor in any other philosophical tradition in the West (except perhaps in Heraclitus, but then his works only survive in fragments). Nevertheless, I intend this book to be rather practical, offering a guideline to those who are involved in the business of governance of AI, so we cannot go into any further detail in the book on these very interesting philosophical issues in Buddhism, Stoicism, or in any other tradition.

The ideas presented in this book find their affinities with a number of works on AI and society. John C. Havens, in *Heartificial Intelligence*,<sup>11</sup> calls for AI that promotes well-being rather than only economic growth, thus chiming with the Buddhist ideal of cultivating oneself so as to avoid the defilement of greed (*lobha*). The idea that ethics should be integrated at the designing stage of technology is also a rather well-worn idea; what I am doing in this book is to put it on a more secure theoretical foundation derived from the Buddhist theory. This idea is given an elaboration by Michael Kearns and Aaron Roth in *The Ethical Algorithm*,<sup>12</sup> who argue and suggest a detailed way in which ethics can be encoded into AI algorithm. What is naturally missing in such an account is an account of what kind of ethics should be encoded, and I suggest in this book that the kind of ethics informed by Buddhism should be able to accomplish the task as it is based directly on the natural condition of both machines and humans. In other words, what is good for machines and humans is what lies in accordance with the nature of the two. The overall aim of an AI algorithm, that it cares for the interests and well-being of others more than its own, should thus be its primary ethical aim because that is the only way to promote the interests and well-being of everyone. Thus, my view is that, for an algorithm to exhibit the ethical ideal, it needs to be related to its social and cultural contexts. This is part of the Buddhist view that all things are interdependent, and solving the ethical problem of AI algorithms cannot be solely a technical matter.<sup>13</sup> Moreover, the idea that ethics should be part of the programming of AI from the beginning is also present in Stuart Russell, one of the pioneers of AI since its inception period. In *Human Compatible*,<sup>14</sup> Russell calls for what he calls "beneficial machines," and the understanding behind such machines bear a resemblance to what I propose as enlightened

machines in the book. The difference, however, lies in Russell's insistence that the robots always maximize the preferences of human beings and attach no importance to its own internal value or well-being at all,<sup>15</sup> while I argue that enlightened machines are modelled upon the ideal of ethical perfection not only for human beings, but for any type of beings capable of rational thoughts and emotional feelings.

The book consists of three parts. The first part deals with more theoretical and metaphysical topics, the second part with more practical ones, and the third part on how AI could contribute to social justice and equality. After this introductory chapter, the next chapter provides a basic background for those who come to this book but might be new to Buddhism. Those who are already well-versed in Buddhist thought may skip this chapter, though I think there are some elements in the chapter that could merit further thought and discussion among Buddhist scholars and practitioners. I would like to note here that the content of the Buddhist teaching discussed in the chapter, and indeed throughout the book, is the common and basic teaching that can be found in all the traditions. These are the teachings that are essential to Buddhism no matter what school a follower belongs to. Thus, there is no argument here whether the content of the teaching here belongs to any school in particular. As a consequence, I am using Pali words throughout, as Pali represents the older tradition of Buddhism whose teaching form the core which can be found in all the later developments of the religion. I will use Sanskrit only to refer to Mahāyāna sutras and their tradition, including the writings of Nāgārjuna. What I would like to ask the reader to bear in mind is that the Buddhism I refer to in this book represents the original teachings of the Buddha himself. This is an important point that I have consciously maintained throughout the book. The teachings covered in the chapter are the Three Practices, the Doctrine of Emptiness, and also my discussion on Buddhist wisdom and its role in elimination of suffering and another section on Buddhism and modern science, something which is particularly relevant for the book.

Chapter 3 deals with the question whether robots can be persons. The discussion here follows from the one on the Buddhist analysis of the self in chapter 2. Owing to the Buddhist analysis of the self in the previous chapter, I develop an argument to the effect that robots that can interact with us fully (such as C3PO in *Star Wars* or Data in *Star Trek*, who are characters in their respective stories in their own right) are in fact persons. However, there are conditions. They must be accepted by their own community, that is, the group of people (and robots) in which they find themselves interacting and in effect living with, as *being one of us*. What is interesting in this conception is that it is an external conception. That is, the criteria by which something or someone

is judged to be a person is its relations with something (or someone) else. It does not make sense to set someone or something alone and asks whether that one is a person or not. Furthermore, as for the narrower AI of today, the case is a bit more difficult to establish. What I am arguing is that they, the narrow artificial specialized intelligence agents, should not be considered persons at the moment, for the reason, essentially, that we, the community who are using and interacting with these robots, are not considering them as being one of us just yet.

Chapter 4 is the most important chapter of the book and is the linchpin that ties all the chapters together. The idea of combining technical and ethical excellence mentioned earlier is derived from the Buddhist thought because being good is a skill that one has to practice since being good is conducive to attaining the final goal. This is an ancient idea and can be found in ancient Greek ethics also. The differences lie in the details of each particular theory. Applied to AI, this means that an AI, in order to achieve its perfection, has to incorporate ethical elements from the beginning since doing so is beneficial to its own interests (as well as those of the manufacturers and programmers) in the long run. The chapter introduces and argues for this notion, as well as that of machine enlightenment, in detail. Ethical ideal for machines and AI is the state where they care for the well-being and interests of others more than their own and the realization that all things are interconnected and interdependent. In this sense, a *really* superintelligent being, if there ever can be one, needs to be an ethical being also. This means that the terrifying scenarios that have been described in the literature, such as superintelligent beings wiping out humankind or putting humans in zoos, need not happen if we take care to instill an ethical sense in the AI from the beginning to the extent that they realize that ethics is not something imposed on them from the outside, but emerges naturally from their cognitive and affective capabilities as a kind of truly intelligent beings. They cannot be truly superintelligent without having wisdom.

Part II of the book is to apply the argument presented in part I in concrete situations. Chapter 5 deals with two autonomous technologies that have been making headlines, namely autonomous cars, autonomous weapon systems, and elderly care robots. Basically, these applications need to exhibit machine enlightenment in their own way. For the autonomous cars, this means that the cars meet high safety standards, always follow traffic rules, and are always courteous to other drives, and so on. As for the famous trolley problem, I argue that the ethical standard for autonomous cars cannot provide a perfect answer because the trolley problem presents a dilemma, and the cars will incur negative consequences any way they choose to act. Thus, the proposal is that we should focus on the factors that we can control, which is designing a good autonomous car, rather than on the dilemma in the trolley problem.

2. Ricki Harris, “Elon Musk: Humanity Is a Kind of ‘Biological Boot Loader’ for AI.”

3. Catherine Clifford, “Facebook CEO Mark Zuckerberg: Elon Musk’s doomsday AI predictions are ‘pretty irresponsible,’” *Cnbc.com*, July 24, 2017, available at <https://www.cnbc.com/2017/07/24/mark-zuckerberg-elon-musk-s-doomsday-ai-predictions-are-irresponsible.html>, retrieved December 26, 2019.

4. “AI Ethics Guidelines Global Inventory,” available at <https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/>, retrieved December 26, 2019. See also, Thilo Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Arxiv.org*, available at <https://arxiv.org/abs/1903.03425>, retrieved December 31, 2019.

5. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems, Version 2* (IEEE, 2017), available at [http://standards.ieee.org/develop/indconn/ec/autonomous\\_systems.html](http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html).

6. For a discussion of the metaphysical issues surrounding AI from the Buddhist perspective, see Somparn Promta and Kenneth Eimar Himma, “Artificial Intelligence in Buddhist Perspective,” *Journal of Information, Communication and Ethics in Society* 6.2(2008): 172–187, <https://doi.org/10.1108/14779960810888374>.

7. In this book I distinguish two kinds of AI, namely artificial general intelligence (AGI) and artificial specialized intelligence (ASI). AGI is the kind of AI that is the holy grail of AI researchers since the time of John McCarthy and Marvin Minsky; that is, a machine that is capable of fully imitating human thought. The goal, however, has not realized as of now, but is believed by many to be achieved in the next few decades. ASI, on the contrary, is the kind of AI that is in use today. It can perform a narrow range of tasks, and only that. DeepMind’s AlphaGo, impressive as it is, is still an ASI. The two kinds are also known as strong AI and weak AI, respectively. For an introduction to the field as a whole in plain language, see Melanie Mitchell, *Artificial Intelligence: A Guide for Thinking Humans* (New York: Farrar, Strauss, and Giroux, 2019), Kindle.

8. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

9. See also Virginia Dignum et al., “Ethics by Design: Necessity or Curse?,” AIES’18, February 2–3, 2018, New Orleans, LA, available at <https://dl.acm.org/doi/10.1145/3278721.3278745>, retrieved January 2, 2020. Dignum et al. ask this important question: “Can we, and should we, build ethically-aware agents?” My answer in the book is that we must, because the stake is too high to do otherwise, and also because we actually *can*, as we shall see in the book.

10. John Sellars, *Stoicism* (Chesham: Acumen, 2006), p. 36.

11. John C. Havens, *Heartificial Intelligence: Embracing Humanity to Maximize Machines* (New York: Penguin Random House, 2016).

12. Michael Kearns and Aaron Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford: Oxford University Press, 2019).

13. Annett Zimmermann, Elena di Rosa, and Hochan Kim, “Technology Can’t Fix Algorithmic Injustice,” *Boston Review: A Political and Literary Forum*, January

9, 2020, available at <http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic?fbclid=IwAR2Uo9LVPfZ8md1Iwu9sgylqqBNDDTbXGmvZjLHqzknvn3g7pIpRKUEHVgM>, retrieved January 10, 2020.

14. Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Penguin Random House, 2019), Kindle.

15. Stuart Russell, *Human Compatible*, chap. 7, Kindle.

16. Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: PublicAffairs, 2019).

17. James Hughes, "Compassionate AI and Selfless Robots: A Buddhist Approach," in Patrick Lin, Keith Abney, and George A. Bekey (eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (Cambridge, MA: MIT Press, 2012).





## *Part I*



were clearly symbolic and were perhaps too good to be true. They showed the reason why the Buddha became bored with mundane life and sought to find a release from these conditions. According to his life story, after Prince Siddhartha saw these signs, he became dissatisfied with his princely life and sneaked out of the palace with his trusted friend, cut off his hair and became a wondering ascetic, emulating the fourth sign that he saw.

Siddhartha spent the total of six years wandering around, practicing meditation assiduously, and in the end he achieved what is known as the Enlightenment when he was thirty-five. The nature of the Enlightenment is precisely the reason why there is Buddhism in the first place. He is supposed to be fully released from the chain of *samsāra* which binds him to the cycle of birth, death, and rebirth in many life forms with no end. We will have ample opportunity to discuss what the Enlightenment actually means for us in the twenty-first century later on. After Siddhartha attained Enlightenment, he was then known as “the Buddha,” literally “the Awakened One.” He gathered a number of followers and started a movement which carried on until today.

### THE THREE PRACTICES

According to the Buddha, the reason why he sought Enlightenment was to become released from the cycle of *samsāra*. That is what the tradition says. What this actually means is that he found that the life he had been having up to the point where he renounced it was profoundly unsatisfactory. We can imagine he was talking to himself: “This is not it. This is not the kind of life I want. What is the point of living, having fun with all the pleasures in the palace, and then dying?” In other words, the Buddha sought after the ultimate meaning of life, a kind of life that goes farther and deeper than merely spending time, having pleasures and passing away. I think we all have the same feeling from time to time. But the Buddha sought to seek a *permanent* way out, a life that guarantees once and for all that it is going to be meaningful. The core of his teaching is that the permanent way can be achieved so that those who attain it do not have to swing back and forth from attaining what she takes to be her way of meaningful life and then forgetting about it. The task for a Buddhist, a follower of the Buddha, is then to study his teachings and practice accordingly, with the aim of achieving this permanent way out. To put it very briefly, one achieves this state when one follows the Path, which consists of rightful behavior (*sīla*), meditation or concentration (*samādhi*), and wisdom (*paññā*). These might look to be quite difficult at first, but let us look at each of them in more detail, since these three practices

comprise the heart of Buddhist teaching and are those that should be practiced regularly by all followers of the Buddha.

These three practices comprise all there is for a Buddhist eventually to achieve the final goal. *Sīla*, rightful behavior, is sometimes translated as morality, but as I will discuss in the section on Buddhist ethics, to translate *sīla* as “morality” is quite misleading, because morality refers to a set of rules that determine whether an act is right or not. In a way *sīla* also refers to rules, but the emphasis with *sīla* is that the rules are there not for the purpose of merely specifying right action in itself. Instead *sīla* refers to a set of rules which is needed in order to engage in a particular practice. We can compare it with the practice of driving. In driving a car, we need to follow certain rules—these are not traffic rules, but rules which are internal to the act of driving itself. For example, in today’s modern car, in order to start the engine, one has to press down on the brake pedal and put the transmission at the P position before one can push the red button to start the engine. If both the brake pedal and the transmission are not in place, then the car will not start. This is a rule set by the designer of the car which is integral to the act of driving the car. Stepping down on the brake pedal and putting the transmission at P is a right action only because it is necessary for the car to start. It is necessary for achieving the goal of getting the engine to work. In the same way, the Buddhist practitioner engages in a set of rules specified in the *sīla* in order to finally achieve the end of Liberation or Enlightenment. The difference with starting the engine is only that stepping on the brake pedal and moving the transmission to parking position is sufficient for the car to start when the red button is pushed; whereas following the *sīla* rules is only a necessary condition. In order to be sufficient for Enlightenment, the practitioner also needs to follow two other practices, namely meditation and wisdom. An important point here is that stepping on the brake and moving the transmission to P position is neither moral or immoral in itself; in other words, the value of stepping on the brake and moving the transmission to P is only dependent on its being necessary for starting the engine. It is only when we *value* starting the engine (because we want the car to move) that stepping on the brake and putting the transmission to P are valuable on their own. In the same vein, *sīla* species a number of acts that one needs to follow if one wants to become enlightened in the end. *Sīla* is neither moral nor immoral in itself; this is an important point that I will discuss further in more detail in the section on Buddhist ethics.

*Sīla* rules vary according to the nature of the practitioners. For example, a layperson is expected to follow the most basic rules, the Five Precepts. A monk, who has dedicated time and effort to achieve Enlightenment directly, is expected to follow much more. In the end the reason why there are the *sīla* rules is because they are necessary for meditation, which is the next step

in the Three Practices. There are a lot of books on how to practice Buddhist meditation already, so we don't have to spend too much time on the topic here. The main purpose of meditation according to the traditional teaching is that meditation is necessary for achieving the last state of the Three Practices, namely wisdom. The idea is that wisdom cannot be achieved without meditation. When the mind is in the meditative state, so goes to traditional teaching, the mind is stilled and free from all distractions, so it is enabled to see the real truth that is hidden behind phenomena. An analogy is that of clear water which lets one to see through it. The water is clear by being still so that all the sediments fall down. In such a state, the mind is able to see nature as it really is. This is why Buddhist masters always emphasize that meditation consists of two parts. The first part is for stilling the mind, and there are many ways to achieve that. The second part is for seeing things as they are, which cannot be done without the mind being still in the first place. So the first part is necessary for the second, and when the mind sees things as they are, they achieve the last stage, namely wisdom.

Wisdom here is the result of seeing; it is like someone knows that the tree is green by looking at it. When the mind in the meditative state sees things as they really are the mind comes to understand the true nature of reality. This is not a result of ratiocination, but a direct perception of the basic nature of reality which is usually hidden from view because the mind in normal, unmeditative state is so clouded with distracting thoughts and ideas that it is unable to see the true nature. Or at least this is what the traditional Buddhist teaching says. As for what the true nature of reality according to Buddhism is, that will be the subject matter of the next section in this chapter. Here the focus is on the Three Practices which comprise the only path leading to the supreme goal for someone's becoming a Buddhist. When wisdom is achieved, that will be the state of Enlightenment. An analogy is that of a lotus flower, which is born from the mud under water, but is undefiled by the mud itself. The unspoiled and undefiled lotus is thus an analogy for the mind of those who have attained wisdom and thus Enlightenment.

## DOCTRINE OF EMPTINESS

So what is the content of the Buddhist wisdom which can only be obtained through meditation? The content is known as the Doctrine of Emptiness, which lies at the heart of Buddhist philosophy. There are several conceptions of the doctrine, but one of the clearest and perhaps the easiest way to put it is through the Three Characteristics. According to Buddhist philosophy, all things have at least one the following three characteristics, namely they are

always changing (*aniccatā*), liable to change (*dukkhatā*) and lack any essential property (*anattatā*). Suppose there is such a thing that it is not changing at all—that is, there is no internal dynamism inside of it at all. We can imagine something that is not changing at all and stays the same for all eternity. Let us imagine that such a thing exists (there might not be such a thing at all in nature). If such an unchanging thing does exist, it is very likely that it is liable to change. That is, it is possible for that thing to change. If we imagine an unchanging and very solid rock, then it is liable to change when something hits it and shatters it. In this case the Buddhist would say that the unchanging rock here is liable to change. So even an unchanging thing (if such a thing does exist) is very likely to be liable to change because we can imagine something hitting it hard enough to change that thing from the outside. However, if we imagine something that is not liable to change (whose existence, if there is any, is even rarer than an unchanging thing), that thing in any case does not have any essential property that identifies it as the thing it is. What this means is that anything whatsoever lacks what Aristotle calls “the what-it-is-to-be” of that thing. For example, a rock for Aristotle is a rock because it has its essence, or its “what-it-is-to-be,” by virtue of which it is a rock and not, say, a piece of wood. An essence of a thing, such as a rock, is that which, when contained within the rock, makes it a rock and not a piece of wood or any other thing. It is precisely this essence that the Buddhist denies when they claim that things have this last characteristic of *anattatā*. We are now imagining something that is not liable to change. Even if we could do so, then this thing does not have any essence or essential property that identifies it as the thing it is and not another. This does not have to mean that a thing can be another thing with no fixing of its identity, but it does mean that there is nothing such that it is what it is and such that its identity is fixed. Thus, the hypothetical thing which is not liable to change, even if such a thing exists, does not have any essential property that identifies it as the thing it is taken to be. It could be characterized as another thing, because there is no fixing of its identity. According to Buddhism, this is the final characteristic and there is indeed no exception here.

It is exactly this point that illustrates the Doctrine of Emptiness. According to this Doctrine, things are devoid of their inherent existence. In other words, things have no essence or essential property (the two are the same, only conceived differently) in such a way that it is always the thing it is. Nature is empty; in other words, nature is populated by things whose identity is not fixed through objective means. That things around us are known as rocks, trees or houses depend on our *conceiving* them to be that way. Taken completely in themselves, they are neither rocks nor trees nor anything at all. This is one of the most difficult points in Buddhist philosophy and lies at the

very heart of the whole teaching. This view should not be taken to be an idealistic one where the mind makes up the existence of rocks and trees, for that view presupposes that the mind has a prior existence over material objects. The Buddhist view, on the contrary, also holds that the mind itself is also dependent on material objects too. Without perceived objects such as rocks and trees the mind would have no material to perceive and thus would have no existence either. This is so because the mind (in this case the *viññāṇa*) is but a conception of external objects when they are being perceived or thought of. Without these objects at all, the mind would have nothing to hold on to, no material upon which it works. We only recognize that the mind is there and is working only by reflecting upon it and hence on whatever that it is reflecting upon, but if there are no things that it reflects upon then there is no mark with which the existence of the mind can be grasped. My position on the Buddhist philosophy here is thus different from certain traditions such as the *Vijñānavāda*, which holds that the existence of the mind is primary. Fortunately, we do not have to enter this traditional debate among schools of Buddhist philosophy. As a very brief introduction to Buddhism for further discussion on its role for today's AI, we can, I think, safely leave this issue aside and conclude that there is an ongoing debate. On one side, the side which I think is more tenable, the Doctrine of Emptiness has no exception; even the mind itself is empty in the sense that it lacks inherent existence—its existence depends on external factors such as things perceived and cognized. On the other hand, there are certain Buddhist traditions, such as the *Vijñānavāda* already mentioned, which holds that the existence of the mind is more primary in that external objects depend on the mind, but not vice versa.

In order to illustrate this, let us look at one of the key passages in the Buddhist scripture on the issue:

O monks! No matter if there is a Buddha, or if there is not a Buddha, but all things will remain the same and stay the same according to their nature. The Buddha has realized that all constructed entities are impermanent. Then he announces, declares, and explains that all constructed entities are impermanent. O monks! No matter if there is a Buddha, or if there is not a Buddha, but all things will remain the same and stay the same according to their nature. The Buddha has realized that all constructed entities are liable to change. Then he announces, declares, and explains that all constructed entities are liable to change. O monks! No matter if there is a Buddha, or if there is not a Buddha, but all things will remain the same and stay the same according to their nature. The Buddha has realized that all things at all lack any essential property. Then he announces, declares, and explains that all things at all lack any essential property.<sup>5</sup>

What this very important piece of the Scripture means is that, whether there is a Buddha or not, things will possess their inherent nature. The Buddha

## BUDDHIST ANALYSIS OF THE SELF

The Doctrine of Emptiness discussed in the previous section leads to one of the most famous teachings of Buddhist philosophy, that of the No-Self (*anattā*, or *anātman* in Sanskrit). In fact the original Buddhist term for the Doctrine, *anattā*, is just the same word for the last of the Three Characteristics. In the previous section I translated *anattā* as “lacking essential property” because in that context the Three Characteristics are supposed to apply to all things whatsoever. In this sense the Three Characteristics also apply to the person and the self. In other words, what we recognize as *us*, our very own selves, are also changing, liable to change and lack any essence. It is this last Characteristic that is the subject matter of the Doctrine of No-Self. The self and the person lack any essence or essential property that identifies it objectively as this particular self or person and none other. This is a poignantly startling doctrine and one that is unique to Buddhist philosophy. To state it more precisely, the Doctrine of No-Self states that the referent of the first-person pronoun is subject to the Third Characteristic; that is, it lacks any essential property that makes it a *this* as opposed to any other entity in the world. We normally assume that we are what we are because there is an essence to our own identity. For example, when someone looks at an old picture of themselves when they were two years old, they normally believe that there is something that remains the same from the time when the picture was taken until the time when they are looking at the picture. That something is the core of the person of the one who is looking at their own picture. This is the normal belief that everyone appears to have. However, what Buddhist philosophy argues is that this belief in the existence of the something here that remains the same throughout is actually an illusion. Almost nothing remained from the time the person is two years old until the time they are grown up and reflect on their early childhood. Suppose I am looking at my old picture taken (I think) by my mother when I was two. I believe that I am the young boy in the picture, and normally I would believe that there is something, that is, me, that endures from the time when the picture was taken and the time when I am looking at the picture now. But according to the Three Characteristics, my self changes and moreover my self does not possess any enduring substance. It is true that brain cells are not replaced throughout the entire lifetime, but brain cells are not me because I consist of trillions and trillions of other cells apart from the neurons inside my brain, and I don't think I am reducible to only a bunch of neurons. The neurons may act as my essence, so goes the essentialist argument, and if so then they are essentially me. The point, however, is that I could have been somebody else even though the very same group of unchanging neurons remain intact. Suppose I was raised in a very different environment, having different



parents; suppose that, for the sake of thought experiment, somebody was born of a different couple altogether, in a different country and speaking a different language, but having exactly the same neurons that I am having right now, then there would be no indication that that person would be the same one as me. Even though that person would presumably look the same and have the same genetic material as me (I don't know much about whether having the same neurons mean you have the same genetic material and would look the same or not, but let's assume that for argument), it is hard to believe that that person would in fact be me because he was born in a different country, having had completely different experiences, and so on.<sup>6</sup>

The upshot of the Doctrine of No-Self, then, is that what we normally take to be our own self—the supposed one who is thinking, believing, feeling, planning, and so on—does not in fact have an essence to it. The Buddha apparently intended this to be a direct refusal of the belief in the existence of immaterial soul that was prevalent in India during his time. According to the belief, our body is animated by the existence of our own soul which is immaterial and continues to exist after we die. This is an ancient doctrine in India and elsewhere. It is our soul that either goes to heaven or hell after we die, depending on our own past actions when we live. It is the soul that does the thinking and feeling and is the one who is conscious. According to this ancient belief, the body is just a lump of matter which will remain lifeless until the soul comes to it and gives it life. It is clear then that the soul is directly in contrast with the Doctrine of No-Self that the Buddha expounded. However, a very important point is that by expounding the Doctrine, the Buddha does not thereby claim that there is no soul; in other words, the Buddha does not say that the Doctrine of No-Self implies that there is no one who goes to heaven or hell after she dies. Someone will still go to heaven or hell and wander in *samsāra* even though the self does not possess any essence that makes it the self of this particular person and no one else. This is a very difficult position to maintain, and Buddhist philosophers have had a very hard time doing so throughout the centuries. Let us then try to unpack this seemingly contradictory position in the space available in this chapter.

Buddhist philosophy revels in contradiction. This has given rise to many debates and schools within Buddhism which we do not have time to go into in any detail here. However, in this particular topic, the contradiction is between the Doctrine of No-Self on the one hand, and the belief that there must be something that wanders in *samsāra* on the other. Justification for the latter belief is clear enough. The Law of Cause and Effect is central to Buddhist thought. There is nothing that can arise on its own without being caused by another thing. Thus, for someone to be born in heaven this must be the result of that particular person having done good deeds in his or her past life; how

else could someone be born in heaven? But if there is no self, then how could someone reap the reward of her past good deeds and be born again in heaven? In fact we don't have to go as far as the supramundane heavenly realm. Someone receives an award for a heroic deed that she did a few months ago. If there is no self—no essence to the person in question, then should we be forced to admit that the person who receives the award is different from the person who did the heroic deed? The point of the Buddha's teaching lies precisely *between* the two extremes. On the one hand, the Doctrine of No-Self is inviolate. There is indeed no personal soul that endures through time. On the other hand, Buddhist philosophy confirms that the person in this case who gets the award is the same as the person who did the deed. The key to this lies in the Doctrine of Cause and Effect. Since there is no enduring self, what we have is a series of momentary selves. We can see this clearly if we start to attend to what we are doing at the moment. For myself, I am typing this chapter on a tablet while waiting for my car to be fixed at a Honda service center close to my home. I am sitting in a large waiting area with other customers also waiting and doing their things. A few moments ago, I remember that I drove the car to the center and was talking to the staff there. The point is that there is only this thread of momentary selves occurring one after another. The awareness of the selves arises when I attend to what I am doing and thinking. What I find whenever I attend to what I am doing and feeling is only these momentary selves which arise and cease each moment I am attending to them. I cannot find anything that endures and functions in the way I normally believe, that is, as the permanent seat of selfhood and identity. If I look for that, what I find is another momentary self, that is, one or more particular, specific episodes of thinking and being conscious. If there is such a thing as the enduring soul, it should be able to be discovered, especially when this kind of soul is supposed to be so close to the person as to be *identical* to the person themselves. Nonetheless, there being no enduring self does not entail that the person receiving the award is a different person from the one who did the heroic deed. Here we have to see more clearly what the talk about the same or a different person actually means. In normal parlance, I am the same person as I was, say, two years ago when I went to Indiana University to do research. I know this because I can remember myself walking around the campus, going to the library and so on. But when analyzed more clearly, there are many differences between my self two years ago and my self as of now sitting and waiting for my car. I have grown (a bit) older; there are more grey hairs on my head; my life situation has changed; situations around me also changed, and so on. Still I believe that I am the same person because the convention of assuming a person to be the same does apply to my case. That is, I count on my bodily continuity; I do remember what I did; there are

pictures taken when I was in Bloomington, Indiana, which much resemble what I look like right now, and so on. However, these are only conventions, that is, what people normally take to be the measures of personal identity. By themselves they are not objective indicators of personal identity. For it is always possible that someone else might pass these criteria but he is not me. It is difficult to imagine, of course, but not logically impossible.

Here the Buddha argues that the person who receives the award is the same person as one who did the deed in the *conventional* sense mentioned above. So there is no question of somebody receiving the award without merit. Being the same person as another one in the conventional sense does not mean that it is not there in nature. I am still the same person as the one who went to Indiana University in November 2017; this is obviously beyond any reasonable doubt. Any experiences I gained there still remain with me and have become part of my self and my overall experience as a scholar and also as a person. But there is nothing that endures. There is a series of causes and effects arising from the time when I was at Indiana and right now when I am sitting in the service center in early 2019. This string of causes and effects do not absolutely and sufficiently constitute who I am; nothing does, but it represents a convenient way of singling out, so to speak, who I am for the purpose of living in the world. In the same way the person who receives the award can rest assured that she really deserves it if all the conventional and normal means of identifying herself indicate that she is actually the same person as the one who did the heroic deed.

In short, the Buddhist position is that even though there is absolutely speaking no means by which we can identify the identity of a person, the law of cause and effect applying to a person still works because the momentary episodes that constitute the so-called person (absolutely conceived of now as an event) are all connected to one another in a series of causal relations. In this sense one is still responsible for one's own action (or more accurately the action of the one earlier in the chain of causal events and momentary episodes). One cannot claim that the action was done by another, unrelated person because there is a direct chain of causal relations leading up to oneself.

### **The Five Aggregates**

Another way of arguing for the notion of No-Self is through analyzing what is understood to be the self into five components, which are traditionally called the Five Aggregates. These are form (*rūpa*), sensation (*vedanā*), perception (*saññā*), thought-construction (*saṃkhāra*) and consciousness (*viññāṇa*). In short, these are a mixture of body and mind that together comprise the self. Form is the material stuff that makes up our body. Sensation is the feeling

that arises when there is a contact between an external object and a sensory apparatus. For example, when light meets the eye, there is sensation. At this stage it is only a bare sensation that occurs when an outside object, in this case light, meets the eyes only. The third Aggregate is perception or recognition, which occurs when the perceived object falls under a familiar mental category, such as when light is reflected from a chair and meets the eyes, the subject then recognizes the object to be that of a chair. The fourth Aggregate then takes up the perceived object, in this case the chair, and forms other thoughts about it, such as this chair is brown or is old and so on. In the final stage, the subject then reflects back upon the whole process and becomes aware that she is seeing the chair and is recognizing it to be a chair. This whole process then constitutes her being conscious of looking and perceiving it to be a chair. We may also summarize the whole Five Aggregates as a conglomeration of body and mind, where the first Aggregate is the body and the other four represents different episodes of the mind. Together they make up the self of a person and there is nothing else to the self except these five. When one recalls something to her memory, for example, the body and the sensation are not at work because one is not using her outer senses, but her third Aggregate recognizes the memory to be, say, that of her trip to the countryside. The fourth Aggregate then furthers up the information presented to her, elaborating it and presenting it to her as a pleasant trip with many further memorable episodes. Then at the final stage she becomes conscious that she is having the memory, becoming self-aware in other words. The point of the Five Aggregates is that this is argued to be an exhaustive list of the self. These Five Aggregates are what the self must be able to be analyzed into and there are no more aggregates other than these five. Thus, if all of these five components can be shown to possess or not to possess any property, then it can be concluded that the self does possess or does not possess the property too. The Buddha's next strategy is to show that each of these Five Aggregates do not possess any essence. Thus, the self does not possess any essence. Furthermore, as the self can be analyzed exhaustively to only these Five Aggregates, none of these Aggregates can be a candidate for the soul since all of them are always changing from moment to moment and are thus insubstantial.

## **WISDOM AND THE ELIMINATION OF SUFFERING**

We have seen that Buddhism has strong views on a number of topics. All things are either always changing, liable to change, or lack an essence. Moreover, the self also falls under this law as we have seen. An upshot is that what we normally conceive as the self is only a construction, something we our-

or Nāgārjuna, in other words having faith that what either tradition says is exclusively true, will automatically lead one to the Goal. The content is only useful when one understands and follows it, using it as a base upon which one practices. Here the practices prescribed by both traditions are similar. One has to follow the *sīla* rules and also meditation. The finer points that underlie the difference between the two traditions are only details that are suited to different individuals. One individual may find the teaching of the Abhidharma to their liking and they are free to follow it; another may be inclined more toward Nāgārjuna and the Mahāyāna. But both of them are equally capable of attaining the Goal. In Buddhist terms the Abhidharma and the teachings of Nāgārjuna are *upāyas* or skillful means that one can take up as a means by which one attains *nibbāna*. The exact content of the teaching of each tradition—whether things are analyzable to indivisible atoms or can be analyzed forever—is in the end, in context of the practice, not as important as whether one attains the Goal or not. There is even a saying in Tibetan Buddhism that if one believes that a dog's tooth is a tooth relic of the Buddha and, having complete faith in the Buddha and his teachings as a result, devotes herself completely and practices the teachings wholeheartedly, then it does not matter whether the tooth is that of a dog or a real Buddha. The tooth itself does not have any special power, but the faith and the inspiration that the follower has arising from what she believes to be the Buddha's tooth relic (which makes her feel closer to the Buddha) is more important.

So does this mean that Buddhism does not pay enough attention to truth? If this is true, then what does the Buddhist wisdom consist of? The disputes among Buddhist schools here does not imply that Buddhism does not have any firm view on the nature of reality. The Three Characteristics are the firm view of Buddhism, but it is how the Three Characteristics are interpreted that can vary within the boundary set by whether the varying views do in fact result in the desired goal of the teaching. A follower of the Abhidharma, believing in the permanence of atoms, is still able to become released from attachment to the material world because by believing in atoms she sees that ordinary things are not what they seem and thus are pointless to hold them fast which is the cause of suffering. Here the atoms function as an anchor for objectivity; without them there would be no way of ensuring that things are objective, and this is one of the main criticisms that followers of the Abhidharma have against views such as Nāgārjuna's. However, this debate takes place within the confines already established by the broader view that does not accept the apparent appearance of entities as being permanent and inherently objective. Within limits, different truths are skillful means for achieving the same Goal.

A text that illustrates this point is Nāgārjuna's final stanza of his *Fundamental Wisdom of the Middle Way*, which says in effect that the true doctrine

is one which “leads to the relinquishing of all views.”<sup>77</sup> It is paradoxical to claim that a true doctrine is such that all views are relinquished. For the doctrine here is also a view and it is also relinquished. This final sentence of Nāgārjuna’s masterpiece has been a subject of countless studies and interpretations. However, what we can see here in our brief overview of Buddhist philosophy is that truths are only *upāyas* or skillful means by which we eventually achieve the main aim of becoming a Buddhist in the first place, that is elimination of suffering and achieving *nibbāna*. A famous analogy here is that of a ladder. When we climb up the ladder and manage to get to the higher plane, we don’t proceed by continuing to carry the ladder. Instead we leave the ladder behind. In the same way, when we study the content of the teaching, understanding its truth, and manage to get to the higher plane (i.e., achieving a certain level of accomplishment laid out in the teaching), we don’t proceed by continuing to emphasize and repeat that truth. Instead we leave the truth behind, that is, we relinquish it as Nāgārjuna suggests in the quote above. Furthermore, not only is this idea of leaving truth behind available in Nāgārjuna, who belongs to the Mahāyāna tradition, but this is also present in the Theravāda too. The Buddha is well known to tailor his teaching to the abilities, preferences and needs of those whom he is teaching. For example, when he was teaching a disciple who used to be a musician before coming to the order the Buddha employed an analogy of the lyre. The lyre makes the most beautiful sound when the strings are neither too lax nor too tight<sup>8</sup>; in the same vein, the practice should not be too lax nor too tight either. Willis Stoesz claims that according to the Buddha in one of the key passages in the Theravāda canon, the *Kassapa-Sihanada Sutta*, comparing his way of teaching and that of others, “[t]he comparison to be made is focused on the way teachings are held and given rather than on their content. Teachings, even those of the Buddha’s path (*magga*), can be given or received in a way that might entrap the would-be seeker of enlightenment.”<sup>79</sup> The idea is that it is not the content of the teaching itself that is of primary importance, but the way in which the teaching is delivered, that is, in such a way that helps the learner genuinely enter the correct Path. In other words, truths themselves, within limits, are not as important as the mindset and way of practice that genuinely enables one to realize the Goal. This point has a strong relation toward Buddhist ethics that we will discuss in full later on in the book.

## BUDDHISM AND MODERN SCIENCE

So far I have given a very brief outline of the basics of Buddhist philosophy. Many of the points are easy enough to understand, such as all things are al-

ways changing, or every event has a cause. These points seem to show that Buddhist tenets are compatible with modern science.<sup>10</sup> In fact many leading Buddhists, such as the Dalai Lama, have organized a series of meetings, known as the Mind and Life meetings, where leading scientists and Buddhist monks are engaged in dialogs, and the results usually confirm that Buddhist teachings are indeed compatible with modern science.<sup>11</sup> Furthermore, recent studies on the effect of meditation on the brain by Richard Davidson and others<sup>12</sup> also serve to confirm that how one practices the teaching has tangible and measurable impact on the material body, which then shows that the teachings themselves have direct relevance on the world. For example, Davidson has found that monks who meditate a lot have measurable changes in the structure of their brains as well as how their brains work in a positive way.<sup>13</sup> These findings and series of talks contribute significantly to the surge of interest in Buddhist and mindfulness meditation in the West.

Nevertheless, there is an aspect quite central to the Buddhist teaching that does not go well at all with modern science. For example, there are numerous places in the Buddhist scripture that talk about hell beings in many gruesome forms, or hungry ghosts wandering around being always awfully hungry but cannot eat anything, and so on. Buddhism is full of these stories and the conventional wisdom in the religion believes that they are central to the teaching. In other words, without beings such as hungry ghosts and the like the teaching does not work. Hungry ghosts belong to *samsāra*, as do heavenly gods, human beings, non-human animals, hell beings, and so on. They are important because they are part of *samsāra* and because they show what will happen to us human beings here and now what could happen to us if we are not diligent in our practice to become released from it. Without these wandering beings in *samsāra*, then, it is as if the whole point of Buddhism is taken away.

However, beings such as hungry ghosts are hard to justify in today's scientific world. Philosophers who have come to become interested in Buddhism, such as Owen Flanagan and others,<sup>14</sup> make it clear that modern scientific worldview just cannot accept that hungry ghosts or heavenly gods exist, and they try to promulgate Buddhism in such a way that these beings do not play a central role. This is a rather difficult trick to pull off, because *samsāra* itself is populated by these beings. Among the six realms of *samsāra*, that is, those of gods, *asūras* (a kind of lesser gods), humans, animals, hungry ghosts and hell beings, only two realms can be verified by modern science, namely human beings and non-human animals. But how can the belief in *samsāra* be maintained without the other four? And without *samsāra*, what then would be the point of assiduously practicing in order that one become released from it? Even among the two verifiable realms, the teaching is such that when one dies it is possible for one to be reborn in the other realm. A non-human animal

can be reborn as a human being and vice versa. But where is the scientific demonstration of that? In order to reconcile the modern scientific worldview to the teaching of Buddhism, it seems that a number of tenets central to the teaching itself have to be abandoned.

According to Flanagan, however, this can be done and in such a way that the objective of Buddhism itself is not impaired. The neat trick I mentioned earlier is that Flanagan argues that the point of Buddhism is literally speaking not that one become released from *samsāra*, but to become released from the bond of suffering in this very lifetime. In a way this can be regarded as being in line with the thrust of Buddhism. I have pointed out earlier that truth takes a back seat to attaining the goal, and this principle can be applied here. According to the tradition, *samsāra* is necessary for attaining the goal of elimination of suffering; cessation of suffering just consists in not being reborn in *samsāra* any longer. This is one version of the truth. But according to the scientific worldview advanced by Flanagan, *samsāra* is not necessary; in fact the belief in *samsāra* is actually a fairy tale belief on par with the belief in Santa Claus. In his version of the truth a practitioner can eliminate her suffering in her lifetime without worrying about what will happen after she dies. In this sense not being reborn in *samsāra* is not a necessary part (or not a part at all) of the elimination of suffering. Is Flanagan's version here compatible with the basic tenet of Buddhism? As I said, truth takes a back seat to attaining the goal. Thus, if it is possible (and indeed it is) for one to eliminate her sufferings without believing in *samsāra*, then she can dispense with the latter, believing that it is a fairy tale, and carry on with her practice until she attains the Goal.

But then one might wonder: Does *samsāra* actually exist? Do heavens and hells actually exist? What I can answer is that such questions do not go along with the spirit of Buddhism. The point of Buddhism—whether one believes in *samsāra* or not—is not to ask whether *samsāra*, or God, or unicorns, exist, but what can be done right now in order for one to be free from suffering. Truth is very important of course, but it can only function in this context as a helper, a means by which one attains the goal of total elimination of suffering. The point is that if there are different versions of truth, and if each of these different versions can bring a practitioner believing it to attain the Goal, then the truth versions here are all acceptable.

I am quite sure that Flanagan himself will disagree with my argument here, for he apparently believes exclusively in the scientific worldview whereas I am taking a more agnostic attitude. This does not mean that I don't believe in science, but that the attitude of Buddhism is such that truth is only a means by which one achieves the final goal of practice or state of realization, and it is not an end in itself. The point I am making here also has support from



one of the key passages in the Scripture. Apart from Nāgārjuna's concluding stanza in the *Fundamental Wisdom of the Middle Way*, there is a famous passage in the original teaching of the Buddha as follows. A man has been shot by a poisoned arrow.<sup>15</sup> The urgent task then would be to pull the arrow out and treat the wound. But if the man or the people surrounding him keep asking questions such as who shot the arrow, what the arrow is made of, under what circumstances was he shot, why was he shot, where was the poison made and how was it put on the arrow, and so on, it will not be long before the man dies of the wound. In the same vein, one should not concern oneself too much with questions such as "Is the cosmos eternal or not?" or "Is the cosmos finite or not?" because there is the more urgent task of getting rid of the defilements and sufferings that come from them. What this parable, known as the Parable of the Poisoned Arrow, shows is that one should, in Nagarjuna's words, "relinquish all views" because all views lead one astray, away from being focused on getting rid of the urgent problem facing oneself. To relinquish all views does not mean that one does not have a view. A Buddhist who follows the Buddha's and Nāgārjuna's advice here does not need to become a total skeptic who does not believe anything, but he believes that all these philosophical speculations and theories do not directly lead one to achieve the Goal, and that achieving the Goal is the more urgent thing to do.

So I don't believe that this attitude is contrary to the spirit of modern science. In fact the spirit of science precisely consists in this attitude of skepticism, of not accepting anything that is not opened up to scrutiny by the community. However, the difference lies in the fact that perhaps Buddhism and science do have different goals. I think philosophers who discuss the relation between Buddhism and science have largely neglected this point. Flanagan tries to make Buddhism more palatable to the modern audience who are brought up within the scientific worldview, so it is not quite surprising for him to present Buddhism "naturalized," shorn of the things that are not scientifically verifiable. But that is different from pronouncing that Buddhism does subscribe to whatever is presented by modern science, as if Buddhism itself is a branch of science. After all, the question that Mālunkaya asks of the Buddha, whether the cosmos is eternal or not, is still being debated by cosmologists at this moment and no consensus has emerged yet. (At present the question becomes "Does the universe exist before the Big Bang?" and "What will happen to our universe in the very distant future?") The scientists and cosmologists, intent on finding the answer in order to advance human knowledge, are focused on finding the best answer to this question, but the Buddhists, focusing on something else, accept whatever is given to them by the scientists and use that, when an occasion arises, to aid them in their quest for realizing the Goal.



## *Chapter Three*

# **Can Robots Be Persons?**

Hollywood movies love robots. They range from the friendly but mute R2D2 to his loquacious partner C3PO; from the menacing and decidedly nonhuman-looking HAL, to the very human child robot David in *AI* (2001) and the all-encompassing and extinction-threatening AI in *The Matrix* (1999). Usually what is there in the movies is a result of the imagination of their screenwriters and producers, which reflect the general trend of the age, the *Zeitgeist*. Artists usually reflect on what is going on around them and try to find deeper meanings behind it. The imagination of artists, poets and screenwriters sometimes precedes that of the scientists and engineers who are engaged in the task of turning imagination into reality. Jules Verne's depiction of humans going to the moon in the late nineteenth century perhaps reflected the technological advances and belief of people at that time that there were other worlds out there that humans could aspire to travel to. In the story humans used cannons to fire "astronauts" to the moon. My hunch is that few in Verne's time actually believed that it was possible for human beings to travel to the moon, and fewer would have thought that a little more than a century later humans would actually travel to the moon and back many times.

Stories about automata who do the hard work for us have, in fact, been around for quite some time. A famous story in Chinese literature, *Romance of the Three Kingdoms*, tells of a general, Zhuge Liang, who invented mechanical oxen made of wood which could travel on their own power and were able to carry heavy loads of rice, much to the surprise of the general's enemies. Not only that, but we also have stories about robots yearning to become humans. The heart-warming story of Pinocchio and its modern version in the movie *AI* shows a character who yearns to become a real human being, capable of loving and being loved. Pinocchio wants to have a soul, something breathed into him so that he can become a living, breathing person. The robot

boy David also wants the same. He wants to be with his mother, living the life of a real, full-blooded human boy, cuddling in bed with his mother. And when in the movie real people singled out these human-like robots for destruction, believing that they were a threat to flesh-and-blood humans, we the viewers could not help but root for the robots. They are so lifelike that we forget they are in fact lifeless, made of plastic and silicon and metal, not flesh and blood.

Perhaps there is something deeply meaningful in these stories. They seem to show that we can empathize with robots. Suppose the story in the movie *AI* was real, and there were certain groups of people intent on catching and destroying the humanoids. We would then feel that we needed to protect them, for our own *humanity* instructs us to do so. In this case it does not matter whether someone is made of flesh and blood or of metal and plastic (or indeed any other type of material); what does matter is that we feel that they are one of us. It might be the magic of the movie that makes us feel that way, but we can imagine that, if the scenario in the movie were actually happening, and more and more humanoids were being targeted for destruction to whet the anger of certain groups of people who hated them, then we could not help but feel sorry for them. In this case we come to believe that these robots (such as those in the movie *AI*) are *persons*. However, the people in the movie who are trying to whip up the emotion of the people, inciting them to hate the humanoids more and more, obviously do not view them as persons. This is in fact normal in the sense that when certain groups want to target another group as objects of hatred, they are not viewing members of that other group as persons, or even as human beings. Certainly, the humanoids are not human beings, but in the movie they exhibit all the characteristics that we associate with humans. Perhaps it's part of our human nature to do so. We even attribute human-like traits to dogs and cats, and feel no qualms about doing so.

In this chapter I discuss the notion of robots being persons according to Buddhist philosophy. This topic is an important one not only because Buddhism has a lot to say about what makes for personhood, but also because a clearer understanding of what it means to be a person goes quite a long way toward establishing some clear conclusions about AI in general. Asking in what sense robots can be persons pertains not only to humanoid robots, those that look like us, but also to AI in general because AI today is in fact exhibiting more and more human-like characteristics. These machines are capable of remembering faces, engaging in conversation, writing news articles, becoming television anchor persons,<sup>1</sup> and much, much more. What I am trying to establish in this chapter is that according to Buddhism, robots can well be persons, and the conditions for their being so are more relaxed than in other religions, especially the monotheistic ones predicated on the belief that human beings are created in the image of the Creator. Robots can be persons

because AI in general exhibits human-like traits. Here Spielberg has it right. In the movie we empathize with the humanoids who have been arrested and subjected to brutal shows to satisfy the blood lust of the audience. We believe that our own humanity makes us feel that way, and by believing so we are extending our notion of humanity to include these humanoids too. Buddhism, being a non-theistic religion, does not have a creation story, and thus has no myth implying that entities have to be created in God's own image in order to qualify for our compassion.

All this, of course, depends on whether humanoid robots can really become movie characters in the same way as those in the movie *AI* or like C3PO. Here scientists and philosophers disagree vehemently. Some, such as Ray Kurzweil, believe that computers will achieve a level of general intelligence, where in only a few decades from now computers will be able to think exactly the same way as a human being does. Some believe that it will take much longer, perhaps a century. Still others think that the time lies so far ahead in the future that it does not matter to us in the twenty-first century at all. Luciano Floridi, for example, has said that we should not occupy ourselves with the possibility that robots will become conscious.<sup>2</sup> That time lies in the distant future, if it were to happen at all. For Floridi we should be more concerned with the problem that robots and AI pose here and now, which is at the level of specialized intelligence, where robots and AI function only within a very specific domain such as playing chess or writing news articles or predicting movements in the stock market. But as Jules Verne's example shows, we should not underestimate the ability of our immediate descendants to do things that we today deem impossible. The Wright brothers were ridiculed by those around them when they tried to realize their vision of powered flight, and we all know what happened after that.

Therefore, my position on this issue is that we should not be pessimistic about the possibility of robots becoming fully conscious, or AI achieving what Kurzweil calls singularity<sup>3</sup> or what Nick Bostrom calls superintelligence.<sup>4</sup> For my argument here to work, it requires only that it be possible for robots to achieve general intelligence. Floridi's point that we should concern ourselves with artificial specialized intelligence is well taken: As long as generally intelligent robots are not imminent, we should indeed be more worried about what is happening at our doorstep. However, not paying attention at all to the potential scenario that robots could become generally intelligent could mean that we are unprepared when the time actually comes, sooner rather than later, and if that is the case, then things could become much worse than if we actually made some necessary preparations.

In any case, the topic of this chapter is the question: Can robots be persons? So far, we seem to take it for granted, intuitively, that if we feel empathy

toward robots, such as those in the movie, then we are treating them as persons. But what are the conditions for someone, or something, to be a *person*? As previously mentioned, this is an important question that has significant implications for our discussion later in the book. So we are now turning toward this topic.

### CONDITIONS FOR BEING A PERSON

In “Conditions of Personhood,” Daniel Dennett writes that there are six conditions for being a person, and that some of them are logically connected to the others.<sup>5</sup> These are (1) being rational, (2) being capable of being attributed consciousness, (3) being capable of having others adopting an intentional stance toward it, (4) being capable of returning the intentional stance adopted by others, (5) being capable of verbal communication, and finally (6) being conscious. For Dennett the first three conditions are mutually interdependent, but they are necessary, though not sufficient, for the fourth condition. Being rational means one can predict the being’s future behavior; it does not mean, for Dennett, that the being is conscious and capable of talking and reflecting to itself. That would be condition six. Here condition two and six are also different. A being can act as if it were conscious, but it is not. Suppose we actually have a talkative robot like C3PO with us; he certainly talks a lot, so he acts as if he were conscious, and in this case Dennett would say that C3PO is capable of being attributed consciousness, even though he might not be conscious, for all we know. The third condition is an interesting one. It means that for a being to be considered a person, it has to exhibit the characteristic, among others, that makes it possible for others to treat it as if it has intentionality. For example, a robot acts as if it were a real human being; it responds to commands and answers questions meaningfully and so on. In this case we normally would adopt what Dennett calls “an intentional stance” toward it, namely we are treating the robot as if it has beliefs and desires. R2D2 acts in a way that we are certain that he wants to go out of the room, and to say that R2D2 *wants* to go out of the room is to adopt an intentional stance toward it. The fourth condition goes one step further; in addition to allowing us to adopt an intentional stance toward it, condition four requires that the being return the intentional stance toward us too. R2D2 acts in such a way that we believe that it wants us to have a certain attitude, such as wanting to go out with it, too. The fourth condition is then necessary but not sufficient for condition five, which in turn is necessary but not sufficient for the last condition, which is a state where the being is really conscious.<sup>6</sup> According to Dennett, the concept of “person” is a slippery one

Dennett's conditions of personhood. These monsters may look like humans; they may indeed *be* humans, but they are (or have become) monsters, and hence non-persons because of their beliefs and behaviors. Of course, they don't have to turn into monsters literally. They can look like humans, but it is their behavior that, figuratively speaking, turns them into monsters. It is enough, however, for them to turn into non-persons in this sense. Here the ontological and normative dimensions of the concept of personhood merge, which is in accordance with our intuitive understanding of the issue. (This is something we shall explore further in more detail later on as we discuss the moral dimensions of robots according to Buddhist ethical theory.)

My point is that the same is also true for robots. The Buddhist position of "lacking an essence" (*anattā*) is clearly pertinent in this discussion. Because human beings and other entities are all lacking an essence of their own, what marks them out as persons or not depends also on what they do rather than what they objectively and essentially are. It is what they do—their outward behavior—that is going to be judged by the community of language users, who then come to an agreement that what these entities—humans, robots, and others—do deserves to be worthy of a person or not, and "worthy" of course is a valued term. In this sense, then, some, but not all, robots and AI in the movies qualify as persons. The reason why I talk about movies is that we still do not have these robots in real life, though of course there is an ongoing debate as to when such generally intelligent robots will come on the scene. We have seen that the humanoids in AI do qualify as persons because we empathize with and feel for them. The lovely and entertaining R2D2 and C3PO obviously qualify, although R2D2 does not talk, which violates one of Dennett's necessary conditions. This seems to be because R2D2 is too lovely to be excluded. In the movies he has shown enough behavior, I think, to qualify him as an independently thinking figure. On the contrary, HAL in *2001: A Space Odyssey* seems to be a non-person in the sense, not that it does not look like a human, but that it exhibits inhumane behavior. Thinking that it is going to be shut off, the computer makes a pre-emptive strike and locks the human astronauts outside of the ship, causing them to die. Our intuition seems to tell us that in this case HAL is not a person: He may be very clever, but he seems to lack the very quality that would make him a person. He operates on the directive given to it by his programmer, and he follows this directive to the letter. In this case he is the epitome of the rational computer, a giant machine capable of performing a tremendous amount of calculations in a fraction of a second, but with no human-like feelings whatsoever. We can see this better by comparing him with R2D2. Even though the latter cannot talk, we do not hesitate in believing that R2D2 is a person because he is so lovable. Of course being lovable is not a sufficient condition for being a person; otherwise

teddy bears would be persons. But it is the combination of the character, its lovability, and behaviors that together make R2D2 appear, at least to many of us, to be a person. Here being a person in the ordinary sense means someone we can trust, someone who is “one of us,” so to speak. The key point, which I will argue for in more detail in the next section, is that being “one of us” is something that cannot be found inside the robot or the human being that we are considering to be a person. R2D2’s behaviors lead us to consider him to be a person; the point is that there are at least two factors: R2D2’s own behavior and *our* regarding him as a person on account of this behavior. And the second of these conditions—our regarding him as a person—is external to R2D2 himself. As for HAL, the fact that he is a totally cool and calculating robot, lacking in human emotion to the extent that he decides to kill in cold blood the human beings he perceives to be a threat, does not seem to qualify him to be a person. My point is that our intuitive reluctance to regard HAL as a person stems from his inhuman behavior. In contrast to R2D2 HAL does not generate any feeling of being “one of us.” On the contrary he evokes the feeling of being remote from us, definitely not one of us in any sense. I cannot help but feel that being a person depends on being one of us in an intimate way. Newborn infants cannot talk; they are not rational. In fact, they lack almost all of Dennett’s conditions for personhood; yet we do not hesitate to think of them as persons, as being full members of the community of human beings. The same is also the case for elderly patients suffering from severe dementia. Deep down we feel that infants and the demented elderly are one of us, and we use this basic feeling as a stepping stone toward realizing that those who lack full consciousness can be persons too.

Furthermore, I elaborated the second problem, personal identity through time, quite extensively in an earlier book.<sup>8</sup> Here I would like only to briefly rehearse the argument presented there, with some added clarification. The idea there is that there are no internal criteria by which one can establish identity through time. By “internal criteria” I mean features belonging to the entity itself that purportedly serve to identify the very same entity through time. John Locke famously argues that one’s own memory is such a criterion: One recognizes that a person existing in the past is identical to the person here now because one has a memory of the former as being the same person as the person recollecting right now. A problem with this kind of account is that it does not allow for lapses in memory or false memories. Suppose we normally have lapses in memory—suppose, that is, that when we recollect a scene from our lives in the past, we are prone to errors such that we don’t really know if the recollected scenes are what actually happened, or something we are creating right now and we mistakenly believe that to be what actually happened. The problem is that, as long as we rely on internal criteria, namely



what we can think and remember, we are inside a loop with no way out. In fact this problem was pointed out to John Locke himself when he wrote about the topic. Bishop Butler argued that Locke's account suffers from a vicious regress because in order to verify that one's memory episodes are veridical one must rely on one's own memory, but that is precisely what is being questioned.<sup>9</sup> The way out, I suggested in my work, is that one has to rely on external criteria.<sup>10</sup> Suppose you are looking at a picture that you believe to be that of yourself when you were two years old. Obviously, you do not look like the boy in the picture, but then there might be some writing in the picture. Perhaps your mother wrote on the picture when she got it that it was of you, taken on such and such a date. Then you come to believe that this is indeed a picture of you. But then your mother's writing is an external criterion; it does not come from within your memory. The point is that without such an external criterion it is very difficult to be certain that our memory episodes are in fact our own.

One might object to this line of argument asking how we can know that the writing on the picture really shows that the person in the picture is the same one as the person looking at the picture many years later. Here we have to accept that no set of criteria is ever sufficient to indicate a person in all cases. This is so because not only is "person" a fluid concept, but also because, as Dennett points out, an exhaustive list of conditions of personhood are found to be wanting in the sense that even though something satisfies one or several of them, there are still cases where that thing is not a person. Even though an entity is self-conscious and can talk, this does not automatically mean that the entity is a person. We could well think of HAL as being self-conscious; he obviously can talk, but as we have seen many of us would balk at accepting that he is a person. In this case, though there is writing on my picture which I recognize to be my mother's stating that this is a picture of me, aged two years old, this does not necessarily prove that the picture is really *my* picture. It could have been the case that the picture was taken of another toddler who looked exactly like me when I was two years old, and, unbeknownst to my mother, was given to her, perhaps by the photo developer, under the context that it was my picture. This kind of scenario is clearly unlikely, but it is not impossible. And it shows that no criterion for identifying persons through time is ever exact or sufficient. We can imagine further an unlikely scenario where my mother somehow comes to wonder whether this picture is actually a picture of me. She then goes to the photo developer and asks how the picture was processed and under what circumstances and so on.

Someone who has been following my argument so far will raise another objection, saying that I have been talking about robots in movies or science fiction, such as HAL or C3PO, and so on, but what about the robots or algo-

rhythms that exist in the real world right now? Are they persons? And is there any connection between the two? Let us imagine that there exists in the world today a child robot such as David in the movie *AI* who talks and behaves just like a ten-year-old child. The fact that there is no robot like David should not deter us from having any feeling that we have when we watch the movie. We can certainly imagine that a robot like David exists and we naturally feel a lot of empathy with him. However, actual robots are being designed and developed that are capable of feeling pain. A team of researchers from Leibniz University of Hannover have developed robots with an artificial nervous system that are capable of exhibiting behavior and responses that mimic the human response to pain.<sup>11</sup> The robot arm winces at the touch of a cup of hot water, for example. An objection to this kind of robot is whether it actually feels pain, in contrast to behaving as if in pain. But that is a long-standing philosophical dispute. We obviously do not know that other persons actually feel pain when they are in a situation where we would feel pain. But if we are holding a cup of hot water perhaps we would exhibit the same behavioral responses as the robot developed by the German team, and perhaps the reason why we don't seem to believe that the robot is really feeling pain is because it does not look enough like us. In any case, we can also imagine that if the robot were around longer and we experienced its behavior and action for an extended period of time, we would tend to believe intuitively that the robot was feeling pain. This is because it is a natural tendency in us to anthropomorphize. This is why we love R2D2 and C3PO so much but feel threatened by HAL. But is the robot developed by the German team here a person? Perhaps not yet. But this is only because we are not very familiar with the robot and the robot is now only at an early stage of development. The point is that there is nothing in principle that would prevent it from eventually becoming a person. And now I will elaborate on the important point I discussed a brief while ago; that is, our feeling of robots being one of us as a criterion for their being considered a person. My contention is that if we feel that a robot is one of us, in other words, that it belongs to the community of thinking and feeling beings (what Kant calls "Kingdom of Ends"), then there is no reason at all why the robot should not be considered a person, which means that this condition (being one of us) should be strong enough to become sufficient for personhood in both the metaphysical and moral sense. According to Dennett, the metaphysical sense of the person has to do with someone who is rational, capable of understanding, and so on, and the moral sense is where the person is accountable for his or her actions.<sup>12</sup>

The objections presented so far have actually been minor ones. The externalist criteria I am proposing faces a much more serious objection when it comes to a completely different conception of the person that does not rely on

externality or relationality at all. In *What Is A Person?* Christian Smith argues for a view of personhood that harks back to the traditional notion found in Aristotle and Kant. According to Smith, human personhood exists on a higher plane of existence, so to speak, enabled by the unique capability of humans to engage in meaning creation and reception. In Smith's words:

I argue that human beings as they exist in the world embody a particular constitution—they have a human nature rooted in nature more broadly. Human bodies interacting with their environments give rise through emergence to a constellation of powerful physical and mental capacities. Those capacities endow humans as real causal agents capable of intentionally affecting outcomes in the world. Those causal capacities interact in complex ways to give rise through emergence again to the “higher” level reality of human personhood. Personal being subsists irreducibly at a level of existence that transcends the lower level elements that sustain it, being characterized by properties, abilities, and qualities unique to human personhood proper. In short, human persons are actual, new realities existent in the world and universe, what we might even think of—if we were not so allergic to the term—as embodied soul-like realities, emergent from the material world from which they arose.<sup>13</sup>

There is nothing that is social or relational in this view. Other human beings are not needed at all for an individual human to become a person. The human person “subsists irreducibly” and “transcends the lower level elements that sustain it.” A human being is entitled to the status of person, not through being recognized as such by her peer, but through properties that she possesses by virtue of being herself. This view of the human person is still the dominant one in the literature. Characteristics such as being “real causal agents,” “new realities in the world,” or “embodied soul-like realities,” are those that are meant to show that human persons are separate, over and above all other “existents in the universe.” These qualities are being evoked when Smith talks about the “properties, abilities, and qualities unique to human personhood proper.” Presumably such properties and abilities include those of language use, conscious thought and the like, namely those qualities that Dennett discusses as the necessary, but not sufficient, conditions of personhood we have discussed earlier. However, language use and conscious thought are actually not possible outside of the human community. Language would be meaningless if there had been one person alone from the beginning with no one else to talk to. If there is no one to talk to from the beginning then even talking to oneself is not possible because language ability requires communication, which itself requires a community. Wittgenstein argues that the idea of a private language—a kind of language which only one person can know because it refers to their own private sensation, for example—is incoherent because there need to be consistent rules for the use of any language and these rules

might be only that, a moving shadow. The objection is that the robot cannot be a person because it does not have an inner life, but it is very difficult for us to *prove* that other humans have an inner life. Still, we keep on believing in it anyway, so why can't we do the same with robots if they can exhibit the same kind of outwardly observable behavior?

In *The Conscious Mind*, David Chalmers argues that consciousness must be a further fact distinguishable from all the physical facts of the universe.<sup>18</sup> He proposes the now famous zombies argument where one is asked to imagine a possible world which is exactly similar to the actual world down to the last detail, but instead of being inhabited by conscious human beings, this other world is inhabited by zombies which are physically like us in every detail except that they lack consciousness. Chalmers argues that the very possibility that there is such a parallel, zombie world shows that consciousness must be irreducible to physical facts, which is what is denied by physicalism. In other words, if physicalism is true, then the parallel world where there are unconscious zombies must not be possible, but for Chalmers it is possible, and therefore physicalism is false. However, we have to see whether it is really possible to imagine a zombie world in this sense. The idea, also stipulated in Buddhism, is that consciousness is a physical fact such that if physical things are arranged in such and such a way, then such an arrangement will give rise to a phenomenon that we recognize to be consciousness. If this idea is true, then the parallel world where every microphysical facts are exactly the same as ours cannot fail to contain *conscious* human beings because consciousness here is just what arises from the constitution of the human brain. The living, normally working human brain always gives rise to human consciousness. This is just what the brain does. Otherwise the hypothesis that the two worlds are the same in all microphysical details would not be tenable. This idea is not a new one at all, but in fact has been around for at least two millennia. It is present in both the Greek and the Indian traditions.

In the Greek tradition we have Aristotle's *On the Soul*, where he holds the hylomorphic view of the relation between body and soul.<sup>19</sup> According to Aristotle, "It is not necessary to ask whether soul and body are one, just as it is not necessary to ask whether the wax and its shape are one, nor generally whether the matter of each thing and that of which it is the matter are one. For even if one and being are spoken of in several ways, what is properly so spoken of is the actuality" (*De Anima* ii 1, 412b6–9). The idea is that the question whether the wax (matter) and its shape (form) are one or not should not arise for it appears to be a trivial one. On the one hand, they are one because that is what the wax shape (suppose it's a figure of a god) actually is; it is not that the wax and the god-shape exists separately in different places. On the other hand, they are distinct because the wax is the matter having been moulded

into the form of the god, and certainly the wax can be melted and moulded again in another, and another piece of wax can be produced having the same form as the earlier one. As long as we are clear about this, then Aristotle seems to advise us not to get embroiled in this kind of dispute. In the human being, then, the matter is of course the body and the mind is the form, and it does not make much sense, in the same manner, to argue whether the two are the same or distinct. In the Buddhist tradition, the Pali text (which forms the basis for all the subsequent Buddhist traditions) is also quite clear on this. In the *Sabba Sutta*, the Buddha teaches that all things, generally speaking, consist of matter and mind together:

O Monks! I will show you all things. Listen to what I am saying. O monks! What are all things? They are eyes and forms, ears and sounds, noses and smells, tongues and tastes, bodies and tactile sensations, minds and objects of consciousness. These, I say, are all the things. O monks! If anybody is to say the following: "I am denying all these things [as the Buddha says], and I am saying that others comprise all the things. These words may belong to sacred objects and are god-like." But when asked, those who propose these words may not be able to open their mouths, and they will be unable to utter any words. Why is that? It is because what they say is not the way things are.<sup>20</sup>

The main point of the Buddha is that he is presenting a view that all things in the total scheme of things consist only of the following and none other, namely "eyes and forms, ears and sounds, noses and smells, tongues and tastes, bodies and tactile sensations, minds and objects of consciousness." In other words, all things are nothing more nor less than our sense perception—our eyes, ears and so on—and whatever is perceived by the senses. This, of course, does not mean that the eyes are a part of the natural world as a sense organ, but it means that the perception and awareness of things perceived by the eyes forms an indelible part of reality, the totality of all things. We are more familiar with the conception that the totality of all things consists only of things on their own, and whether those things are perceived or not is irrelevant. But for the Buddhist this is a mistake. Things on their own without being perceived or without being perceptible at all are nothing to us. It is as if they do not exist. What the Buddha means when he teaches this *Sutta* is that all things depend on our perception of them. It is the pair consisting of the eyes and the objects seen that comprises all things. The eyes alone are not sufficient because the eyes would then see nothing. This makes the eyes useless so it would be as if there are no eyes in this case. And the same applies for the other sense modalities. An interesting case is the pair of the mind and the object of thought, or of consciousness. Here the Buddha is talking in the empirical sense of momentary episodes of thought, such as when someone

is thinking of something. In this case, the person is thinking, for example, of an object. The thinking cannot go on if there is no object, that which is being thought. The two have to exist together in order for the thinking, understanding or awareness to function. More interestingly, the Buddha claims that these two, the mind and object of thought, also comprise all the things that populate the universe. I think we can perhaps understand this better if we realize that when I, for example, am thinking of something, both I and that of which I am thinking both exist as components of the totality of all there is. This is a crucial point. We are very familiar with the Cartesian world picture where there is a radical separation between thought and world, but that is not the Buddhist picture. The Buddhist picture, according to the Buddha here, is one where the thing being thought of and the act of thinking belongs together in the same sphere, namely they both belong to the universe of things, so to speak. In this sense the thing being thought of by me has the same ontological status as the keyboard I am typing on at this moment. They both belong to the universe of what the Buddha calls “all things” (*sabba*). Body and mind, in other words, do not belong to separate realms, but they belong to one and the same realm, on the same ontological level with each other.<sup>21</sup>

## ROBOT PERSONHOOD: THE BUDDHIST PERSPECTIVE

So according to Buddhism, can a robot be a person? The answer is that it depends. Firstly, if the robot is of the thinking, sentient and conscious kind, then there is nothing in Buddhism that would prevent it from becoming a person for the reasons that we have seen in this chapter. Thus, an AGI robot is a person by virtue of its possession of subjectivity (or all the behavior that shows that it has an inner life), sentience, and rationality, all the conditions that are necessary for being a person. Secondly, if we are talking about the robots we have at this moment, ones that operate under what is known as artificial specialized intelligence (ASI), then it's a more open question. Basically, though, one would not consider ASI robots to be persons because they are not enough like us to merit our including them into the community of persons. This may seem *ad hoc* and arbitrary, but consideration of the concept of personhood appears to include a significant amount of fuzziness, so much that depends on our own intuitive perception and judgment, that there does not seem to be any hard and fast rule. Many philosophers such as Tom Regan and David Gunkel have argued, respectively, that we should give rights to animals and robots respectively. According to Regan, higher mammals such as elephants and dolphins are, in his words, “subjects-of-a-life,” meaning roughly that they have an inner life and are aware of their experiences, thus they have an

intrinsic value and are entitled to moral rights.<sup>22</sup> Robots, according to David Gunkel, are also entitled to rights because this is what we should do.<sup>23</sup> In other words, our interaction with social robots is such that it has become natural to treat them as *social* beings, and this entails that they should be entitled to a system of rights and personhood that typically accompany a social being.<sup>24</sup> To put it simply, Gunkel argues that robots (at least the social kind) *do* have rights because they should, and they should because it is the fact of the matter that we are treating them as such. And it is a simple logical matter that when they have rights, they must have personhood, too. In this case, Regan's conception of higher animals as being subjects-of-a-life would be applicable to robots if they are the same, but this would qualify them to be thinking and sentient beings, a kind of AGI robots already. For Regan, if, presumably, a robot lacks a kind of inner representation that would qualify it to be a subject-of-a-life, then such a robot would not have the same rights as those that do. Here Gunkel's approach is broader: No matter if a (social) robot does or does not have an inner, subjective life, if we treat it as a social being, then it has rights and is thus a person. Here the more difficult problem is, of course, the ASI robot, one that does not have an inner life yet.

So, according to Buddhism, is an ASI robot a person and thus entitled to moral rights? There is no direct discussion in Buddhist philosophy about rights or personhood *per se*. The Doctrine of No-Self stipulates that the self is an illusion or a construction, so it should be safe enough to conclude that for the Buddhist, the person is also an illusion or a construction, because the concept of the person is based on that of the self. Be that as it may, for the AGI robot, the question should still be answered "yes." An AGI robot should be considered a person in Buddhism because it possesses all the necessary qualities—sentience, rationality, subjectivity, and so on. Even if it is the case that the self is a construct, this does not detract from the AGI robot being considered a person because the AGI robot will have to have a self, albeit a constructed one. It is the same as each and every one of us humans, who has a self, even though what we take to be our own individual self is a construct. However, an ASI robot is a more difficult case. Gunkel argues in no uncertain terms that at least a class of ASI robots, the social kind such as the robot dolls or robot companions for the elderly and perhaps those that work as a sexual companion, should be considered persons having a certain set of rights. In Buddhism the person should be considered in relation to the type of beings that wander around in *samsāra*. Even though there is no self-subsisting soul, the Buddhist still believes that what one has done in the present life has an influence over the condition of existence of another being in the next life. When someone has died having done a certain set of deeds in her life, another being (not necessarily a human being) will be born bearing the karmic fruit

of the one who has already died even though the two are in all likelihood different persons. This is roughly comparable to the situation where I inherit certain traits, such as having curly hair, from my grandmother, and it is obvious that my grandmother and I are different persons. Certain traits belonging to my grandmother, having curly hair, cause certain traits in myself to happen through genetic transmission. All schools of Buddhism believe that what one has done in one's present life will have an influence over the condition of life of another person who will be born and who receives the influence generated by the earlier person. This sounds like a mumbo jumbo, but, in the spirit of scientific Buddhism, we do not have to accept this teaching. We only need to recognize that this topic of karmic influence is a possibility, and that it is a possibility that underlies the story of *samsāra*. The chain of karmic influences makes it appear *as if* there were a person carrying the influence around when the person enters the body-mind complex of another sentient being. The point is that it is the person who functions in talks about karmic influences and *samsāra*, making it appear as if there is someone who gets reincarnated, and so on, while in fact there is no such person, only a chain of karmic effects. So if the robot is to be considered a person, Buddhist philosophy has to accept that the robot in question is capable of sending out karmic effects when it dies as well as receiving influences from others when it is born.

This of course sounds almost nonsensical. Even to a committed Buddhist, the story of a robot wandering around in *samsāra* sounds very much like a complete distortion or fabrication of the Buddha's teaching. However, in the spirit of scientific-minded Buddhism advocated by Flanagan and others, there is nothing wrong with this. We should not regard *samsāra* as a kind of extrawordly realm where the spirits of the dead move around waiting to be reborn, but *samsāra* is a reflection of the quality of action that we human beings (and beings like us such as robots in the future) do in this life. There are six levels of beings in *samsāra*; these are hell beings, hungry ghosts (*pretas*), nonhuman animals, human beings, demigods (*asūras*) and gods (*devas*). Hell beings are consumed with anger; hungry ghosts live with constant hunger; animals lack a clear mind and understanding; demigods are consumed with jealousy of the gods, and the gods always enjoy constant sensual pleasure. The human being lives in a central position in *samsāra* in that they can become any of the beings depending on the state of mind she is in and the action that she undertakes as a result of the state of mind. In order to realize the main goal of Buddhism, one does not have to believe that *samsāra* physically or objectively exists. One only has to believe that it is a reflection of the quality of action that one is doing at any moment. Thus, someone who is consumed with anger is in a sense living in hell at the moment, and when



animals, as well as gods and hungry ghosts. What can be said for sure, what is shared by almost all schools of Buddhist philosophy, is that there is no soul such that it takes off when the body dies and seeks another body to reside in. The Buddha is very clear on this. But that does not mean that the living body is a zombie and lacks consciousness, because being sentient is just the property that we encounter all the time when the animal we are encountering is alive, and being sentient is a necessary condition for consciousness. You have to be able to feel in order to know that you are feeling. Thus, if robots can show to our satisfaction that they are capable of feeling and have an inner life in the same way that we can infer from observing our friends that they have an inner life, then these robots are persons. All this is a serious matter; in fact, the issue is of central importance to Buddhist philosophy. We will take all this up again in more detail when we discuss in the next chapter whether thinking and sentient robots can attain the highest stage of *human* perfection in Buddhism, that of being enlightened.

## NOTES

1. Alex Linder, "Xinhua Shows Off World's First Female AI News Anchor," *Shanghaiist*, February 21, 2019, available at <https://shanghai.ist/2019/02/21/xinhua-shows-off-worlds-first-female-ai-news-anchor/>, retrieved January 2, 2020.
2. Luciano Floridi, "What the Near Future of Artificial Intelligence Could Be," *Philosophy & Technology* 32.1(2019): 1–15.
3. Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Viking, 2005).
4. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
5. Daniel Dennett, "Conditions of Personhood," in Michael F. Goodman (ed.), *What Is a Person?*, 145–167 (Totowa, NJ: Humana Press, 1988).
6. Daniel Dennett, "Conditions of Personhood," 178–179.
7. Daniel Dennett, "Conditions of Personhood," 193–194.
8. Soraj Hongladarom, *The Online Self: Externalism, Friendship and Games* (Springer, 2016).
9. See Joseph Butler, "Of Personal Identity," in John Perry (ed.), *Personal Identity*, 2nd ed., 99–106 (Berkeley: University of California Press, 2008).
10. Soraj Hongladarom, *The Online Self*, 51–82.
11. Evan Ackermann, "Researchers Teaching Robots to Feel and React to Pain," *IEEE Spectrum*, May 26, 2016, available at <https://spectrum.ieee.org/automaton/robotics/robotics-software/researchers-teaching-robots-to-feel-and-react-to-pain>, retrieved February 24, 2019.
12. Daniel Dennett, "Conditions of Personhood," 176.

13. Christian Smith, *What Is A Person?* (Chicago: University of Chicago Press, 2010), 15–16.

14. Karl F. Macdorman and Stephen J. Cowley, “Long-Term Relationships as a Benchmark for Robot Personhood,” in *ROMAN 2006—The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 378–383 (Hatfield, 2006).

15. Rom Harré, *The Singular Self* (London: Sage, 1998), quoted in *Raya Jones, Personhood and Social Robotics: A Psychological Condition* (London: Routledge, 2016), 8.

16. Amelie Oksenberg Rorty, “A Literary Postscript: Characters, Persons, Selves, Individuals,” in A. O. Rorty (ed.), *The Identity of Persons*, 301–324 (Berkeley: University of California Press, 1976), 322.

17. F. Patrick Hubbard, “Do Androids Dream: Personhood and Intelligent Artifacts,” *Temple Law Review* 83(2011): 405–474, p. 419.

18. David Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (Oxford: Oxford University Press, 1998).

19. Aristotle, *De Anima (On the Soul)*, Hugh Lawson-Tancred, trans. (Penguin Classics, 1987).

20. Samyutta Nikāya, Salāyatana Varga 18/24/19, quoted in Prayut Prayutto, *Buddhadharma: Expanded Edition*, printed as a Memorial in the Funeral of Mr. Thalerg Laojinda, B.E. 2558 (2015), 51 [in Thai]. Text available online at [http://www.watnya.naves.net/uploads/File/books/pdf/buddhadhamma\\_extended\\_edition.pdf](http://www.watnya.naves.net/uploads/File/books/pdf/buddhadhamma_extended_edition.pdf).

21. The idea here is not unique to Buddhism. In fact, the idea that body and mind belong to one and the same entity is also found in Spinoza’s philosophy. According to Spinoza, there is only one thing, essentially speaking, which can be called either “Nature” or “God.” This one thing, one Substance, then contains infinitely many Attributes, but only two of the Attributes can be conceived by the human mind, namely body and mind. The point of similarity with the Buddhist thought here is that body and mind—the thinking and the thought—are ultimately one and the same in the sense that they do not belong to separate categories as in Descartes. In Buddhist philosophy, both body and mind are entities which become entities only because of their relations with other entities; thus, both of them can be regarded as being empty of inherent characteristics. In Spinoza, body and mind are conceived as pertaining to the essence of Substance, which roughly means that they both express the very basic reality of the thing itself. In both Buddhism and Spinoza, then, body and mind—the thought and the thinking, or the subject and the object—are essentially one and the same thing. Buddhism only goes further when it insists that the thing which is conceived of as either body and mind is in fact empty. That is, it is devoid of any inherent, self-subsisting characteristics at all. For a preliminary treatment of the similarities and differences between Buddhist philosophy and Spinoza, see Soraj Hongladarom, “Spinoza & Buddhism on the Self,” *The Oxford Philosopher*, available at <https://theoxfordphilosopher.com/2015/07/29/spinoza-buddhism-on-the-self/>, retrieved December 11, 2019.

22. Tom Regan, *The Case for Animal Rights*, updated with a new preface (Berkeley: University of California Press, 2004).

23. David Gunkel, *Robot Rights* (Cambridge, MA: MIT Press, 2018), Section 6.2. See also David Gunkel, “The Other Question: Can and Should Robots Have Rights?” *Ethics and Information Technology* 20(2018): 87–99. <https://doi.org/10.1007/s10676-017-9442-4>

24. David Gunkel, *Robot Rights*. Section 6.2

25. Tom Regan is well known for his staunch defense of the rights of animals. See Tom Regan, *Defending Animal Rights* (Urbana: University of Illinois Press, 2006); *The Case for Animals Rights* (Berkeley: University of California Press, 2004); Tom Regan and Peter Singer, (eds.), *Animal Rights and Human Obligations* (Englewood Cliffs, NJ: Prentice Hall, 1976). See also Lidia Cano Pecharroman, “Rights of Nature: Rivers that Can Stand in Court,” *Resources* 7.1(2018), <https://doi.org/10.3390/resources7010013>.