The development of artificial intelligence (AI) is exploding. We may soon have AI that is equal to us in every way, what is called artificial general intelligence (AGI). Not long after that we may see the emergence of artificial super intelligence (ASI) which may vastly exceed human capabilities. This AI is inescapable. It will be part of the fabric of human life and revolutionize every field of human endeavor: science, medicine, business, education, entertainment, and more. We have AI now that do things previously thought unimagineable like composing music, writing stories, generating images and even short videos.

The transformative power of intelligence is enormous. It is this trait that sets us apart from other animals. It may also be what sets AI apart from people as the creation of a new species. Intelligence can potentially solve every problem facing humanity. Even problems that are computationally complex in terms of the number of moves and possibilities like chess, Go, and StarCraft, have been conquered by recent AI systems. So has the protein folding problem in biochemistry and other complex problems in mathematics and physics. The era of quantum computing also has much potential.

Advanced AI could achieve great things and push humanity to new heights of greatness, but it can't do this in an unsupervised manner. If left unchecked it could amplify misinformation, introduce biases in training data sets, and erode privacy.

It could also substantially increase the chance of existential risks, those that threaten the destruction of civilization. This could happen intentionally or accidentally through the actions of a person or a machine. We need to anticipate, plan and act to prevent such risks no matter what their origin.

These responsibilites fall within the new fields of AI safety and AI governance. Individuals in the first field study the possible risks and develop solutions. It is more the realm of philosophers, mathematicians, computer scientists, and engineers. The latter involves the societal regulation and implementation of solutions and is more the realm of politicians, public policy makers, lawyers, and economists. Of course the two groups should work cooperatively to be most effective. The future needs to see better ethical oversight, responsibility and accountability by corporate AI initiatives. It also needs to implement proper safeguards oriented around empathy, bias control, and transparency.

The key to navigating these two futures lies in values. A value, simply put, is something people want or desire. Values reflect our preferences. They can be individual or cultural. Some of the problems associated with values are that they are too complex, there are too many of them, and that they are relative, varying too much from one person or society to another. We argue that this is in fact not the case. Evidence from a number of different fields shows that values are complex but can be clearly defined and that they are universal with a small number of core values shared across cultures.

The value alignment problem is the process of making computers act in accordance with human wishes. The traditional way of doing this is utility maximization, where we take a desired measure like people's happiness. We then have a machine act in such a way that it increases its score on this metric using learning techniques. As we shall see, this method suffers from a number of problems and has been supplemented with a variety of other approaches involving human feedback, rules, game theory, virtues, and computational social choice.

In the end, no matter what procedure we choose, a method be ethical. In other words, it must give people what is "good" for them. Of course, defining what is good for us and how to implement it in an AI system is not an easy task. In this book we are concerned more with the "what" question and not as much with the "how", although the two are complexly intertwined. Ethics and value formation are abstract topics requiring broad-spectrum thinking, while implementation is a domain of technical expertise and analytics. In other words, it is the place where philosophy and computer programming meet. This book attempts to bridge this gap. We attempt to show that values should not be abandoned and that they are a key ingredient in making AI safe.

In the following section we provide a preview of each chapter. This is to enable the reader to better navigate their way through the book. Some of the material at the start of each chapter is introductory. For those readers familiar with these ideas, feel free to skip ahead. The book in interdisciplinary and touches on key concepts in AI, the philosophy of axiology and ethics, theology, evolutionary and survey psychology, and ecology. Chapters are previewed by an introduction that outlines the coming content and concluded with a summary and integration.

Chapter one is intended as a primer on intelligence, AI and the contemporary issues surrounding it. We examine what intelligence is, potential types of AI including the most recent advancements like deep learning, LLMs and generative models. We also provide an overview of AI risks and AI safety.

Chapter two takes on the intersection of AI and the concept of values. We look at what values are and defend their existence as meaningul philosophical and psychological constructs that can be measured and used to assist in the construction of ethical AIs. We then review the fear of value convergence, by which an AI might take control over society in order to achieve terminal goals. Value alignment is discussed in detail, including the definition, goals, and principles.

Chapter three is from a philosophical perspective. The ideas of ethics and how they can be applied to machines are reviewed. We focus on the issue of autonomuous moral agents, whether an AI can become human in the sense that it understands and is responsible for its actions.

Chapter four is told from various psychological points of view. We start with evolutionary theory which is foundational to understanding core human values in terms of social cooperation. Following this is a look at more data-based approaches, falling into two categories. The first includes studies of religious and other relevant texts from around the world. The second includes cross-cultural surveys of people's opinions and preferences. All of these studies converge on the universalist notion of a small set of human values.

Chapter five adopts an ecological or ecocentric rather than anthropocentric approach. We cast a wider net and look at the importance of the planetary environment in shaping values for humans and AI. Biophilia, AIs that care for the Earth, and the agent-based systems that could simulate ecosystem interactions are presented.

Chapter six promotes the primary thesis of the book which is that we can determine a universal set of human values and use these as the basis for value alignment. A set of such values are sketched out along with the associated problems that must be overcome. We conclude with a look towards the future, how AI can help us improve ethics and what a posthuman era ethics might be like.