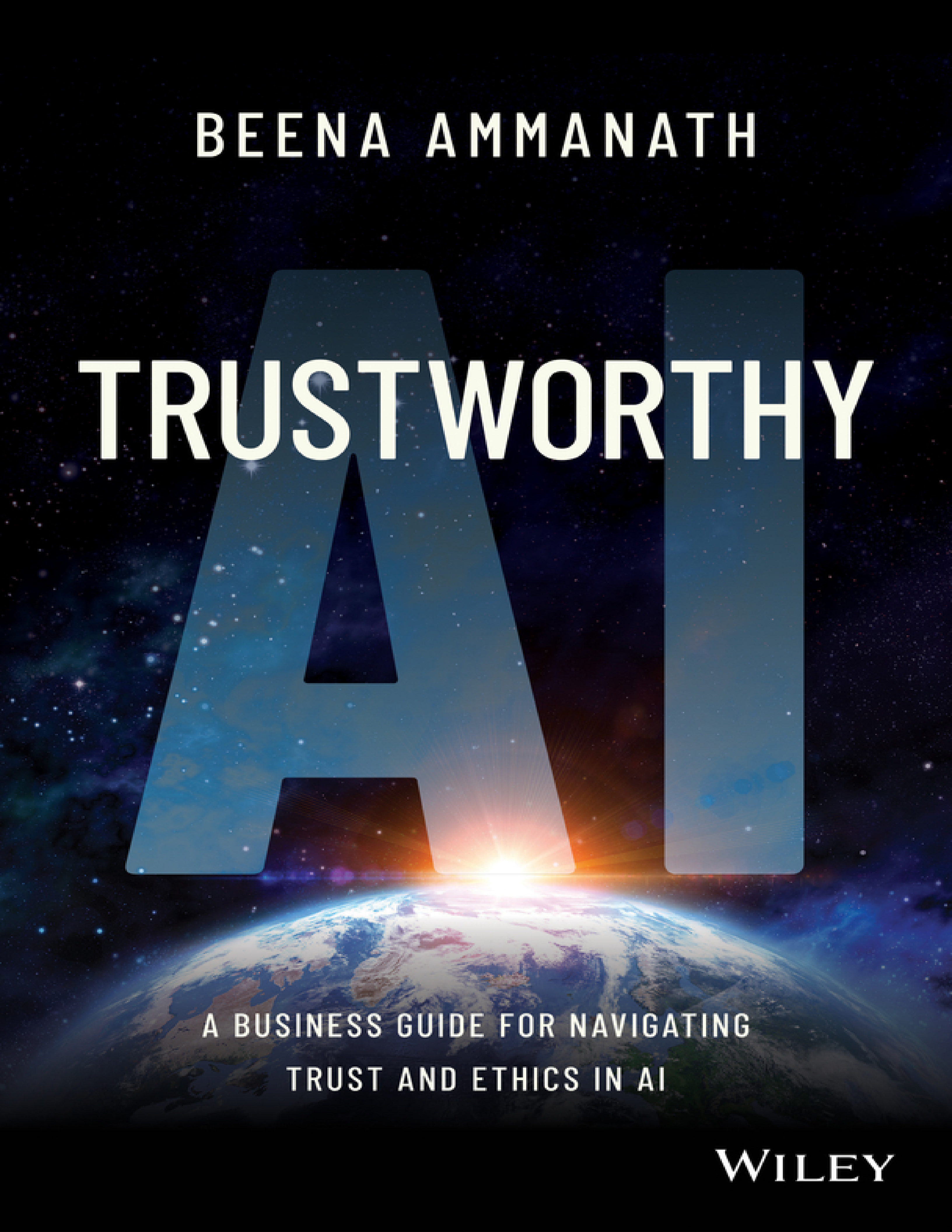
BEENA AMMANATH

TRUSTWORTHY AI

A BUSINESS GUIDE FOR NAVIGATING TRUST AND ETHICS IN AI



WILEY

Chapter 1
A Primer on Modern AI

As a prelude to investigating the dimensions of AI trustworthiness, it is important to clarify what AI is – and what it is not. In popular press and general discussion, AI is often ascribed a sense of self, as if it "thinks" and even has a measure of "volition." The nouns we use to describe AI projects include "training" and "learning," concepts typically reserved for thinking creatures. Thus, it is perhaps unsurprising that common ways of talking about AI drift toward conceiving it as a true intellect.

This is quite far from reality. In truth, AI does not "think." AI tools are in fact highly complex mathematical calculations that have been constructed such that the solution to the calculations accurately describes something in the real world. More than this, AI as a concept is not just a discrete thing. It is as much a collection of model types performing different functions as it is a professional practice area for data scientists and the technology ecosystem that permits their work and the AI they develop.

To better define what we mean when we discuss AI, consider how the field developed to its current maturity and the kinds of AI models in use today.

The Road to Machine Intelligence

Humans have long imagined mechanical tools that act in a seemingly intelligent way. These ideas permeate the stories we tell, from those going back thousands of years to the ones we enjoy today. The automata created by Hephaestus in Greek myths, the mechanical beings in the Mahabharata and other ancient Hindu texts, the genre-defining fiction of Isaac Asimov and other writers – humans have always wondered about how inanimate machines might be given independent will that serves (and sometimes threatens) its creators.

When we discuss AI, one important step is delving into just what we mean by intelligence. Human beings have a clear sense of self and a rich internal world. Our decision making and knowledge owes to a pantheon of experience, intuition, superstition, emotion, and all the things that make us thinking creatures. AI as it is today is much narrower in its cognitive potential, accomplishing only what it is designed to do.

AI as a modern scientific field of study and practice emerged in the mid-twentieth century. It tracked the development of modern computer science, inspired and propelled in part by the work of British computer scientist Alan Turing. In the 1930s, Turing demonstrated mathematically that rules-based code could solve algorithmic problems, and it was he who developed the eponymous test for interrogating the presence of machine intelligence.

From those beginnings, the field of AI notched up a series of events and inflection points that moved the technology forward. At the 1956 Dartmouth Summer Research Project on Artificial Intelligence, researchers presented what has been dubbed the first AI program, Logic Theorist, and computer scientist John McCarthy coined the term “artificial intelligence.” In the decades after, computer science and computational capabilities both evolved and improved. While there was heady excitement over what AI could potentially accomplish, however, the hardware, software, and algorithms were insufficiently powerful.

Over time, the technology advancements needed for AI, such as computer storage, steadily emerged. In the 1980s, deep learning techniques were devised, opening the door for machine learning (rather than purely rules-based code). While initially conceived in the 1950s, it took several decades for a type of AI called expert systems to mature. These used symbolic logic, data-driven processing, and outputs that could be understood beyond the realm of complex mathematics. The excitement was such that by the end of the 1980s, more than half of Fortune 500 companies were creating or using expert systems.[2] Yet, for a variety of reasons, including the technical and cognitive limits of expert systems, this avenue of AI fizzled out.

In the 1990s, neural networks received more technical innovation and more effective algorithms. Massively parallel processing also received research attention, seen most publicly in IBM's Deep Blue computer, which in 1997 beat the chess world champion in a six-game competition. Thus, it took nearly half a century to progress from the origin of the concept of AI to a technology that exceeded human performance in a highly complex activity.

At the turn of the century, the pace of development in computational infrastructure and capabilities quickened. The capabilities in data storage, parallel processing, and the data generation and connectivity permitted by the advent of the Internet all moved toward the computational power needed to make real the loftiest AI ambitions. Continued innovation around artificial neural networks made possible the potential for things like computer vision recognition, wherein a cognitive tool could accurately classify an object in an image. Yet, this type of AI and others like it were flummoxed by a fundamental issue – for machines to learn what an image contained, those images had to be labeled by a human.

For example, if there is a photo of a lion on the African savannah approaching a herd of gazelles, the machine learning tool has no sense of what is what. It does not know which is the lion and which is the gazelle, or even the concept of an animal in the wild. As such, lofty projects set out to hand-label every object in massive databases of images. This became prohibitively laborious.

Then, in 2011, deep learning emerged in full. Stanford computer scientist Andrew Ng and Google engineer Jeff Dean constructed a neural network, pairing it with a dataset of 10 million images and a cluster of 1,000 machines. They let algorithms process the raw data, and in three days, the cluster had independently created categories for human faces and bodies, as well as cat faces. This was proof that computers could generate feature detectors without labels. It was the advent of unsupervised learning.[3]

Over the last decade, these and other types of AI have proliferated and are being deployed at scale by organizations across every industry and sector. This has been aided by enormous generation of data through connected devices, flexibility in cloud computing, and the development of critical hardware (e.g., the graphics processing unit). Today, organizations are operating in a period of vigorous innovation and exploration. They seek not just to automate components of the enterprise but to totally reimagine how business is conducted and identify use cases that were never before possible. To be sure, AI is no longer a "nice to have." It is a competitive necessity.

Basic Terminology in AI

AI is not one thing; it is many things. It is an umbrella term for a variety of models, use cases, and supporting technologies. Importantly, the development of one machine learning technique does not necessarily make another obsolete. Rather, depending on use cases, there are a variety of AI techniques that may be most appropriate.

AI raises a highly technical lexicon that can be opaque to people outside of the data science field. The concepts in AI describe complex mathematics that can leave nontechnical people unsure of how AI actually works. There is no shortage of writing that probes and contests definitions in this evolving field. Yet, we do not need math to grasp the basics of AI. Definitions of relevant and often-referenced terms include:

> *Machine learning (ML)* – At its most basic, ML consists of methods for automating algorithmic learning without human participation. The algorithm is supplied with data for training, and it independently "learns" to develop an approach to treating the data (based on whatever function the architect is optimizing). Machine learning methods might use both structured and unstructured data, though data processing for model training may inject some structure.
>
> *Neural network* – An NN loosely models how a brain functions, in as much as it uses connected nodes to process and compute data. It is not a distinct physical object but instead the way computations are set up in a virtual space within a computer. An NN contains an input layer, an output layer, and a number of hidden layers between them. Each layer is composed of nodes

and connections between nodes that together form a network of layers. Data is inserted into the input layer, computations are autonomously performed between hidden layers, and the algorithm produces an output.

Deep learning (DL) – A subset of ML, DL is largely (though not exclusively) trained with unstructured, unlabeled data. A DL algorithm uses a neural network to extract features from the data, refine accuracy, and independently adjust when encountering new data. The "deep" in DL refers to the number of layers in an NN. A challenge in DL is that as layers are added to the NN, the level of training error increases, and the task for data scientists is to adjust NN parameters until the algorithm is optimized to deliver an accurate output.

Supervised learning – In ML, one approach is to feed an algorithm labeled datasets. Humans curate and label the data before model training, and the model is optimized for accuracy with known inputs and outputs. In supervised learning, there are a variety of model types for classification (i.e., sorting data into appropriate categories) and for regression (probing relationships between variables).

Unsupervised learning – In this case, the training data is largely or entirely unlabeled and unstructured. The datasets are fed to an ML algorithm, and the model identifies patterns within the data, which it uses to reach an output that accurately reflects the real world. An example is the unsupervised learning approach Ng and Dean used in their 2011 image recognition experiment.

Reinforcement learning – Similar to how humans learn to act based on reward or reprimand, reinforcement learning is the ML approach where an algorithm optimizes its function by calculating an output and gauging the "reward," what could be simplistically called "trial and error."

While this list barely scratches the surface of AI vocabulary, it is sufficient for us to think critically about how AI training is conducted, how it can be applied, and where trust and ethics become important.

Types of AI Models and Use Cases

While there is a large degree of technical nuance and model variety, many of the AI tools used today can be categorized according to their basic operation and function. As a window into how AI is being used, review this sampling of AI functions and use cases:

> *Computer vision* – AI cannot "see" anything, but a computer vision model can process the bits of data that together constitute a digital image and, from that, determine mathematically what is likely to be in the image. Today, this is possible not just with static pictures but also with real-time video. We see computer vision used in autonomous vehicles, facial recognition, equipment monitoring, and much more.
>
> *Natural language processing (NLP)* – An NLP model can analyze, decipher, search, and generate language in the format humans use "naturally." The model does not "understand" language, but it can process and treat text such that the outputs are coherent and accurately reflect the data. These tools can classify, search, and create text. An example is an AI chatbot that can process a question from a customer and reply in a helpful way.
>
> *Speech recognition* – Text-to-speech programs are nothing new, but AI adds a layer of knowledge. As words, intonation, and speech patterns are deciphered, speech recognition tools can analyze the sentiments of the person speaking. For example, is the person speaking expressing anger or joy, frustration or satisfaction? The way in which speech is delivered impacts the context and meaning of the words. Using sentiment analysis with an NLP model can yield a powerful tool that can compute not just what a person says but also what they mean.
>
> *Planning, scheduling, and predicting* – In a complex organization, variables across business units and the speed at which conditions change can exceed human capacity to make fully informed decisions. Planning and scheduling were previously conducted by hand in spreadsheets. AI models today can offer granular insight across every business factor,

supporting informed decision making and even predicting the likelihood of an issue occurring and recommending solutions to avoid or mitigate it.

Recommendation systems – With the growth of online shopping and media, the general public is aware of recommendation systems that serve up products, content, or offers that are relevant to the user. These models can become extraordinarily sophisticated when paired with information about the user, such as their shopping and travel habits, their age, income and education, and their online activity in social communities. Deep insight into consumer personas allows an organization to offer an individual or group the right content, offers, or advertisements in the format and timing most likely to be compelling.

Robotics – While not a distinct type of AI, cognitive tools are essential for semi- or fully autonomous robotics. Using AI to operate a physical object requires a collection of models and data that allow a robot to function in the real world. This may include computer vision but also monitoring machine performance, changes in the environment, and the degree of predictive certainty in given actions. We see these collections of AI in places such as manufacturing, autonomous vehicles, and consumer products (e.g., robot vacuums).

Ultimately, these types of AI are just a sampling of the true potential in cognitive tools. There is so much left to be conceived and invented, and excitement over and eagerness to use AI is fueling innovation, investment, experimentation, and progress. An appropriate question for any organization exploring AI is not just asking what the tool can do, but also, what the organization might do with it.

New Challenges for the Modern AI Era

When AI existed only in research labs and its potential was largely experimental, questions about trust and ethics were mostly academic. It is only when we deploy these powerful tools at scale that we are forced to contend with the unanswered questions about the ethics of AI and whether we can trust this bold new era of machine intelligence. As usually happens, our technology advancement has preceded the evolution of the sociotechnical system needed to govern it toward our collective best interests.

After decades of work and innovation, AI has matured to a point where it now touches almost every aspect of our lives. It is not a one-off research project that escaped the lab but instead *the* transformational technology that will shape our future. As such, armed with a general appreciation for what AI is and how it works, we can begin the serious work of exploring how to make this technology something we can trust.

Notes

[1] Marie Curie, Nobel Lecture, December 11, 1911, Nobel Prize Outreach AB 2021, https://www.nobelprize.org/prizes/chemistry/1911/marie-curie/lecture/.

[2] David Schatsky, Craig Muraskin, and Ragu Gurumurthy, *Demystifying Artificial Intelligence* (Deloitte, November 4, 2014).

[3] Quoc V. Le et al., “Building High-Level Features Using Large Scale Unsupervised Learning,” *Proceedings of the 29th International Conference on Machine Learning* (2012).