Human beings are somewhat unique in their caregiving for their own sick and diseased. Around this seemingly preternatural urge has developed a tremendously sophisticated medical system wherein the sick are diagnosed and treated. It has long been acknowledged that the tools wielded by medical practitioners have the potential for great benefit and also for great harm, depending on how they are used. The scalpel, for instance, can be used to remove cancerous tumors to extend life. The same scalpel can cause irreparable harm and death. While every tool used in medicine has the potential to be directed either toward healing or harming, it is through a professed system of ethics that patients can trust that practitioners of medicine aim their tools toward the benefit, and away from the harm, of their patients.
MAXIMIZING BENEFIT AND MINIMIZING HARM: 
THE ROLE OF ETHICS

It seems that the same thing can be said of artificial intelligence (AI) systems: they are capable of great benefit and also great harm. In health care generally, AI systems may be beneficial by increasing access to care, making care more affordable, improving the overall quality of care, and fostering increased trust in the health-care system. On the flip side, the systems may also further entrench a two-tiered health-care system in which the insured get high-quality, personalized care and the uninsured and underinsured, who are often patients of color, get lower quality and often depersonalized care.3

AI systems also have the potential to dehumanize medical care and devalue the role of the therapeutic physician-patient relationship. AI systems that are able to independently make a diagnosis and prescribe a treatment are likely to increase the efficiency and decrease the cost of providing health care. Such systems could plausibly replace many of the roles played by physicians. It is likely that some of the functions performed by physicians today will be objectively improved when physicians collaborate with AI systems, and some functions may actually be performed better by AI systems alone. However, at its core, medicine is relational, and most patients place a high value on the relationship. Some patients may even value the relationship so much that they would accept a less perfect human to a more perfect AI system. Although AI systems may objectively perform some functions better than physicians, the relationship between the patient and their physician cannot be replaced by an AI system, regardless of how humanoid the system is. Furthermore, if
patients perceive that increased efficiency and decreased costs have led to the elimination of the physician-patient relationship, they may come to further distrust health-care institutions.

As AI systems disseminate into almost every aspect of society, particularly health care, one must ask how to ensure that their potential for good is being maximally actualized and the harm that they could cause is being assiduously avoided. Multiple frameworks and checklists have been developed and promulgated that seek to provide a structure for these systems that, if observed, would shift the balance toward benefit and away from harm. One of the issues becoming apparent in the process of adopting a framework for AI ethics is that these systems, in and of themselves, have no ethical obligation other than to perform as they claim. They are essentially amoral systems without a predetermined set of obligations or duties.

At this point, comparisons between AI and medicine diverge, since medicine has a long, well-established, and in many cases legally codified system of ethical obligations and duties. Before embarking on the application of these duties, it is helpful to first articulate what these ethical obligations are and how they are understood by most physicians today. Perhaps the best known and most widely utilized medical ethics approach is that described by Tom L. Beauchamp and James F. Childress in *Principles of Biomedical Ethics.* The system these authors describe has come to be known as principlism and encompasses four principles: respect for patient autonomy, beneficence, nonmaleficence, and justice. Applied to medical decision-making, these four principles guide physicians both in their own medical decision-making and also when engaging with the patient in shared decision-making.
It is within this ethico-medical context that AI systems are being introduced. Although AI systems are being implemented in almost every facet of health care, including operational and administrative arenas, this text specifically addresses the use of AI systems that are used to make clinical decisions and that involve both physicians and patients. These systems, often called artificial intelligence clinical decision support (AI CDS), are trained on historic patient information to make a desired clinical prediction. For instance, in designing a system that will predict which patients are at increased lifetime risk of developing kidney disease. The system will (ideally) have access to hundreds of thousands of patient records and either make a prediction based on outcome labels determined by human users (supervised learning) or discover patterns associated with its own computer-determined outcome labels (unsupervised). These systems learn from the medical histories and outcomes of past patients to predict clinical outcomes, disease classifications, and therapeutic responses for current and future patients. Importantly, and a topic of discussion later in the text, current patients can also contribute their health information to the system after it has made their prediction. In doing so, the system continues to learn new associations, which can shift over time (e.g., the COVID-19 pandemic affected time-honored associations between fever and bacterial sepsis).

This work focuses on AI CDS specifically for three important reasons. First, a general AI ethics framework that seeks to address every AI application would inevitably be either so general as to lack necessary granularity for certain applications or
so prescribed and detailed as to be overly restrictive to other applications. Second, medicine generally, and physicians specifically, already have well-defined and widely understood ethical codes. There is no need to reinvent or rediscover medical ethics; rather, medical ethics must be applied to AI systems that are used for medical decision-making. Third, these systems are being introduced into a human relationship between a physician and a patient and are best understood in that context. The relationship is built on the patient’s trust in their physician. The patient’s trust in the physician is justified and perpetuated because of the physician’s profession of ethical obligations to the patient.
This book also focuses primarily on those AI CDS systems that require collaboration between a physician and the system to make a clinical decision applicable to an individual patient. While AI CDS is being developed to provide clinical predictions directly to patients, these systems will likely be governed by a different set of normative considerations than will those that necessarily involve collaboration with a physician. It seems doubtful that the inevitable evolution of AI CDS will be toward direct to patient clinical predictions that make physicians obsolete. Whether it be because of patient demand, the inevitable technical limitations of these systems, a desire to maintain human oversight, or some combination of considerations, the most likely outcome is that physicians will collaborate with these systems routinely without these systems altogether replacing the physician in the relationship with the patient.

Finally, most of the consideration of AI CDS in this work focuses on systems that require physicians to initiate collaboration with the AI system and volitionally request clinical decision support. The AI decision is offered to physicians “on demand” when physicians request it. This is in contrast to AI CDS systems that are “pushed out” to physicians without the physician requesting collaboration. The ethical relevance of this distinction has several considerations. First, on demand clinical decisions presume that the physician is initiating the collaboration with a specified clinical aim, such as making a diagnosis, prescribing a treatment, or predicting an outcome. Second, the physician is prepared to act on the prediction, typically by accepting or rejecting it. Last, the physician is prepared to inform the patient of the prediction made by the system in the process of shared decision-making, even if the physician does not inform the patient that the decision was made in collaboration with an
AI system. In contrast, when physicians receive a clinical prediction without requesting it, they may be unprepared to assess and, in turn, act on the information and inform the patient. This could lead to unintended use of the prediction; failure to act on the prediction, even if it would be clinically beneficial to the patient; and failure to inform the patient of the prediction.

**THE STAKEHOLDERS IN THE PHYSICIAN-PATIENT-AI RELATIONSHIP**

Just as each clinical ethics dilemma involves the perspective and values of many different stakeholders—the patient, the physician and medical team, the patient's family, the hospital and health system in which the patient is receiving care, the community to which the patient belongs, the payer of the patient's hospital bill, and others—so too are their multiple stakeholders involved in the development, programming, deployment, use, and auditing of AI CDS. Although each of these AI CDS stakeholders has a different perspective and also a different value system, they all have a common relationship to the AI CDS: entering into the physician-patient relationship.

How, then, can it be assured that the AI CDS is directed toward the patient’s benefit and away from harm? It is through the expressed ethical commitment of each of these stakeholders, whom we define as the patient, the physician, the programmer, and the health system administrator and payer, that patients can be assured that AI CDS maximizes benefit and minimizes harm.

This book seeks to consider the pertinent ethical concerns, derived from the physician-patient relationship, that are relevant to each stakeholder. Beauchamp and Childress's four principles are a common starting point for understanding the ethical
concerns relevant to each stakeholder. This is not because principiplism is particularly well suited to AI CDS, but rather because it is the most commonly used and best-known framework in medicine today. In fact, as this text highlights, the four principles have several weaknesses generally and specifically when applied to AI CDS. For instance, the principle of autonomy is often thought of as encompassing informed consent, veracity, and confidentiality. While it is true that these other principles may demonstrate respect for patient autonomy, when it comes to the ethical issues facing stakeholders involved with AI CDS, informed consent, veracity, and confidentiality are important considerations in addition to, and distinct from, respect for autonomy.

When considering the stakeholders who have a relationship with AI CDS, it is important to acknowledge that the physician
has a direct relationship with both the patient and AI CDS, whereas the other stakeholders have direct relationships with AI CDS but indirect relationships with patients. The importance is twofold. First, it is into the physician-patient relationship that both AI CDS and the other stakeholders are introduced, a relationship governed by ethical considerations. By entering into a normative relationship, these other stakeholders are reasonably expected to not only understand the ethical norms of the relationship but also abide by them. Second, it is the physician whom the patient knows and sees, and with whom the patient talks and has the primary relationship. This is true even though other stakeholders may play an outsized role in the medical decisions made within the physician-patient relationship. The physician is vulnerable not only to the patient’s ire, but also to peer review, licensing boards, tort claims, and professional societies if the patient is harmed.

Take, for instance, the decision by health systems to sell patient health information to AI developers. A patient discovers that their health information has been sold and used to profile them across AI platforms—not by name or social security number, or medical record number, but by their own unique combination of health information that is specific to them and distinct from everyone else. The patient may be furious with their physician, blaming them for dereliction of their ethical obligation to maintain confidentiality, resulting in loss of trust. The physician, however, may be unaware that the health system sold patient data. They may have no idea how to respond to the patient. In fact, even the physician’s health data may have been sold. Patients may feel similarly angry and betrayed by their physician if they learn that AI CDS has been used to make high-stakes clinical predictions about them without their being
informed or giving consent. Although the physician has the most “skin in the game,” when nonphysician stakeholders enter into the relationship between the physician and the patient, the other stakeholders should abide by the ethical norms that govern the patient-physician relationship.

**NORMATIVE CONSIDERATIONS IN HUMAN RELATIONSHIPS AND AI COLLABORATIONS**

While myriad ethical frameworks are proposed for AI, this book takes a relational approach since medicine is ultimately relational. Applying a relational way of thinking about AI ethics could serve as a model for normative considerations of other AI applications. Humanity is ultimately relational, and patient-physician relationships are merely one kind of relationship, which occurs between one human experiencing illness and another human who has the knowledge and skill to treat the illness or promote greater health. While the normative considerations in medicine are well defined, in other human relationships the obligations are less well defined, but just as important. Introducing AI systems into society will require that humans understand not only collaborations between AI systems and humans, but also the relationships between humans to which the AI-human collaboration is most similar.

Many of the characteristics of the physician-patient relationship described in this book are presented as they would ideally be practiced. However, as health care has become increasingly commercialized it has, in many instances, become less personal and more transactional. One risk when introducing AI CDS into health systems is that systems that maximize cost savings, increase efficiency, and improve reimbursement will be
prioritized for development and deployment over systems that are purely clinically beneficial. It may also be that decisions about which AI system to develop or deploy are not made with input from the physicians collaborating with the system, nor from the patients and families affected by the system, but rather by executives who are responsible for the hospital or health system’s profitability. That is not to say that financial security is not a legitimate consideration and even ethically relevant. Rather, the concern is that the potential of AI CDS to improve clinical decision-making and overall patient care will be thwarted by the potential to contain costs and improve profit margins.

If concerns about AI CDS being used primarily to drive profit can be adequately addressed, then perhaps these systems could lead to the reinvigoration of the art of medicine. While it may seem paradoxical that an advanced technology would enhance the humanistic aspects of medicine, there are two important ways in which it has the potential to do so. First, as others have highlighted, AI systems can reduce the rote non-clinical activities that consume physicians’ time and energy. These activities are more burdensome and less professionally rewarding than caring for patients, which is a physician’s highest calling. However, there is another important role that AI systems have the potential to play that will increase the human dimensions of the physician-patient relationship. The corollary of this is that these systems have the potential to significantly reduce, but not altogether eliminate, clinical uncertainty. This will allow physicians to engage with patients and families in shared decision-making and to focus less on complex calculations and predictions based on error-prone hypothetical-deductive reasoning. Physicians will be able to focus more on problem solving, creativity, altruism, and the art of medicine,