Chapter 8
Explainability in Medical AI

Ron C. Li, Naveen Muthu, Tina Hernandez-Boussard, Dev Dash, and Nigam H. Shah

After reading this chapter, you should know the answers to these questions:

- What are the current trends in AI explainability research?
- What types of explainability paradigms can be conferred onto different machine learning (ML) models?
- What are the different methods by which ML models can be explained?
- How can principles of cognitive informatics be applied to explainability in medical AI?
- What is an 'emergent property' of a sociotechnical system?
- What regulatory frameworks have been put forth with regards to accountability of ML models?

Introduction

The current paradigm of artificial intelligence (AI) in medicine primarily relies on machine learning (ML) models as a means to provide insights—typically in the form of a diagnosis or prognosis—that can affect the health of individuals and populations. A model learned from past data is often a trigger that invokes a series of

R. C. Li (✉) · T. Hernandez-Boussard · D. Dash · N. H. Shah
Stanford University School of Medicine, Stanford, CA, USA
e-mail: ronl@stanford.edu

N. Muthu
University of Pennsylvania School of Medicine, Philadelphia, PA, USA

Children's Hospital of Philadelphia, Philadelphia, PA, USA

T. A. Cohen et al. (eds.), *Intelligent Systems in Medicine and Health*, Cognitive Informatics in Biomedicine and Healthcare,
https://doi.org/10.1007/978-3-031-09108-7_8

actions comprising a care workflow. We define a *model* as a function learned from data that maps a vector of predictors to a real-valued response. Predictors are also referred to as inputs, features, or variables; response is referred to as an outcome, output, label, or task. The "logic" of *how* ML models generate their estimates, and how those estimates translate into recommendations in the context of explicit or implicit policies, is often difficult for human beings to understand.

The high complexity and dimensionality of the relationships that ML models derive from data are often not interpretable by human reasoning, which is why many ML models are often referred to as "black box" models. However, we have known for decades that explainability is an important attribute of any human reasoning process (see Chap. 5) and clinicians have historically named it as a top requirement for a clinical decision support system [1]. Because model generated recommendations in medicine can affect high stakes decisions, the discussion around the "explainability" of both the model's output and the policy that translates that output into actions is particularly relevant for the safe and effective use of this technology.

Current Trends in AI Explainability Research

Explainability of ML models has been deeply explored across academia, industry, and government as a potentially critical component of applying AI into health care in a way that is usable, transparent, and trustworthy [2]. Key themes from current work in explainability center on how it is defined, the methods by which it can be achieved for different ML models, how it is evaluated, and whether it is truly useful when applying AI in healthcare settings [3]. Based on the current consensus definition of explainability, a ML model is considered explainable if the explanation satisfies two criteria: (1) it is "interpretable," meaning that the logic the model incorporates to make predictions is understandable by humans, and (2) it has fidelity, meaning that the explanation faithfully reflects the underlying logic of the task model (the model making predictions) [4].

There are now a range of methods described in the literature to generate explanations that attempt to satisfy these criteria, albeit with varying degrees of success [5]. These methods can be broadly divided into two categories: (1) using aspects of the model's intrinsic architecture (e.g. beta coefficients from a linear regression) to derive explanations, which can only be done for certain model architectures, and (2) post hoc methods, where separate interpretable models are developed to accompany the original, potentially "black box", model in order to approximate explanations between model features and the outcome. The majority of such post hoc methods fall into the category of attribution-based explanations, which use a variety of quantitative methods to attempt to measure the relative importance of the task model features in determining the outcome. These methods are typically applicable to more complex and non-linear model architectures such as neural networks that may deliver higher predictive performance at the expense of the lack of intrinsic model interpretability. Nevertheless, while these computational strategies can be used to approximate the relative importance of model variables, they do not reflect the true inner workings of the task model logic, so they may not satisfy the fidelity clause of explainability. Further, statistical explanations, even for the more "easily explainable" linear task models, still require an additional layer of human interpretation that can vary and may not faithfully reflect the underlying model mechanism.

In light of these limitations, whether explainability is truly useful when applying AI to health care continues to be debated. Explainable models are thought to facilitate users' ability to understand and improve the model, discover new insights learned from the model, and even to be more empowered to manage social interactions with other humans when using the model [6]. Qualitative stakeholder studies have also indicated that clinicians seem to want to understand explainable variables when exposed to predictive models in order to assess whether they align with their clinical judgement [7]. The healthcare AI field has indeed been moving forward, with increasing interest demonstrated in government research and development, venture capital and industry, as well as in professional societies as these entities encourage the development of methods, financing, and regulations that encourage explainability in medical AI [8, 9]. However, there remains some skepticism that explainability can truly enhance the usefulness of AI in health care, as well as concern that it may even lead to harm. For example, explanations, especially if they do not sufficiently satisfy the clauses of interpretability and fidelity, may give users a false sense of security, especially since they typically require some level of statistical comprehension and nuance to understand them, even for linear models [10].

Applying Additional Context to Understand Explainability in Medical AI

How we think about the meaning and purpose of explainability and its incorporation deserves deeper examination because the answers to these questions may depend on the context in which the model is deployed. This chapter applies principles of **cognitive informatics** to delve into these questions. Consider the following hypothetical scenarios:

1. An AI software product is used to analyze chest CTs as part of an automated system for lung cancer screening. Patients with chest CTs that are flagged by the AI software as high risk are automatically referred for biopsy.
2. A physician and nurse for a hospitalized patient each receives an AI generated alert that a patient for whom they both are caring is at risk of developing respira-

tory failure in the near future and recommends mechanical ventilation. They proceed to meet and discuss next steps for the patient's clinical management.

3. A consumer smartwatch outfitted with AI capabilities, detects cardiac arrhythmias and notifies a user that an irregular heart rate has been detected recommending that the user consult a physician for further evaluation. After performing a full clinical assessment, the physician orders a continuous cardiac monitoring study for a formal diagnostic evaluation.

Although each scenario includes an AI solution, the nature of the task performed by the AI enabled tool and how it is incorporated into patient care differ. The first scenario describes an AI approach that drives the diagnosis of lung cancer and automatically triggers an intervention without any mediation by humans. The second AI scenario also drives the diagnostic and management process for a high risk medical condition, but the process is mediated by humans. In example three, the AI system supplies diagnostic insights, but is intended only to be supplementary information for a formal evaluation; however, at the population level the use of such a system does impact the total amount of work done by those that have to perform the formal evaluation.

The level of risk associated with the task and extent of human involvement in the delivery of care has broad implications for how to approach the purpose of AI explainability as well as the kinds of explanations provided. For example, in scenario one, where the system drives high stakes clinical care without any mediation by human clinicians, it may be important for patients, as well as the clinicians, to understand the tool's "reasoning" behind its conclusions, similar to how a patient would want a physician to explain the reasoning behind a cancer diagnosis. The health system employing this AI solution and regulatory bodies may also require in-depth understanding of how the ML model generates its predictions and the level of model performance for quality assurance. In scenario two, the AI system interacts with human clinicians who need to synthesize the prediction with the rest of their clinical evaluation in order to make a decision about the patient's management. While the clinicians need to trust the tool for its advice to be adopted, the mechanics of how the ML model generated the prediction may be less important to the clinicians than a conceptual understanding of why the program predicted this patient to be at risk that they can mentally incorporate into the rest of their clinical assessment. In scenario three, trust in the AI advisor is similarly important, but insight into the "how" and "why" of the AI prediction may be less relevant to the non-clinician layperson user since the AI prediction is only meant to be supplemental to a formal evaluation by a physician and does not directly drive care management.

These scenarios demonstrate that the thinking around the need for AI explainability must move beyond a binary "yes/no" paradigm to *it depends* and *for what purpose?* Explainability can be for several purposes: understanding how the relationships between variables generate the output of the ML model, a conceptual appreciation for why certain predictions are formed from the available data, or simply as a surrogate for trust in the model's performance. As illustrated by these scenarios, the purpose of explainability depends on the nature of the task, the recipient(s) of the predictions, and the broader environment in which the AI system is deployed.

Deciding the appropriate purposes of AI explainability requires an understanding of how AI interacts with human users and the implications these interactions have on downstream clinical outcomes. In order to capture the depth and breadth of how explainability affects medical AI, we must consider three levels of impact that explainability can have on how AI systems are shaped: information processing by the individual human user, the interactions between people and AI agents, and the **emergent properties** of the broader AI-enabled **sociotechnical system**.

For example, when assessing how to incorporate AI explainability into scenario two, we would first consider what the physician and nurse individually need to understand about the prediction in order to make sense of it in the context of their understanding of the patient (i.e. do they need to understand *how* the program generated the prediction of respiratory failure in order to make sense of the rest of the patient's clinical findings or do they primarily need to know that the predictions are rarely wrong). Second, we would query how explainability would affect the ability of the physician and nurse to interact with the AI agent to make shared decisions (i.e. given the social nature of human cognition, can the AI system function as an effective teammate?) Finally, these interactions with the AI system and the physician and nurse may have downstream ripple effects that may ultimately affect the clinical outcomes in unpredictable ways, such as impact on communication patterns, culture, and patient safety.

The goals of this chapter are (1) to describe the different purposes of AI explainability, (2) to present a framework for assessing the different needs for AI explainability by examining how an AI system interacts with human cognitive processes situated in sociotechnical systems, and (3) to discuss how this framework can be applied to real world examples of AI in medicine and implications for regulatory approaches.

Three Purposes of AI Explainability

Explainability is a tricky notion given the lack of consensus in the form of explainability desired and when. For AI systems in medicine, we consider three purposes of explainability: (1) to allow the study of a ML model and perform quality assurance and/or improvements, (2) to help the user(s) of the AI system to gain contextual understanding of the model's prediction in order to incorporate into their subsequent decisions and actions, and (3) to facilitate trust in AI systems (Chap. 18) [11]. To the ML engineer, explainability often refers to the ability to articulate which variables and their combinations, in what manner, led to the output produced by the model [12]. This approach to explainability requires an understanding of the computational relationships among the variables and architecture needed to generate the model outputs, which is often highly complex in ML. For example, an ML model that predicts respiratory failure may generate predictions from hundreds of thousands of features derived from clinical data that may not be clinically meaningful (e.g. the log of the mean blood pressure over 24 h cubed), and the computational relationships among these features are often high dimensional and difficult to represent in any clinically understandable way. Sometimes, features may be included in an ML model due to pure statistical associations but not indicate any potential causal relationship that would be helpful to a clinician seeking insight into what about a patient's clinical status may be increasing their risk of respiratory failure (e.g. hair color may be a feature in a model that predicts respiratory failure given a possible statistical association in the training set, but this would not offer a clinically meaningful explanation for why the patient may be a risk for respiratory failure). The *purpose* of this type of explainability is typically to allow engineers to perform quality assurance and to replicate or improve on the ML model, whereas a user of the AI system, such as a clinician or patient, may not find this type of explanation helpful (Fig. 8.1).

To the clinician or patient user of medical AI, explainability is more likely to be important for enriching their understanding of the prediction in the context of the clinical situation and providing information that would allow them to trust the performance of the AI system. For example, a team of physicians and nurses who are alerted by an AI system that their patient is at risk of going into respiratory failure and that the situation may warrant mechanical ventilation will typically want to understand which clinical variables contributed most to the model's prediction. Here, explainability allows the clinician users to make sense of the prediction in the context of the rest of their evaluation as well as potentially to use that information to tailor their subsequent decisions and actions to respond to the risk. The precise

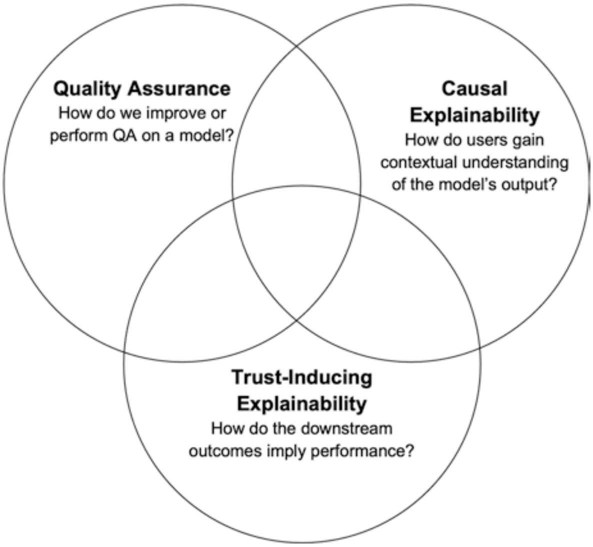
Quality Assurance
How do we improve or perform QA on a model?
Causal Explainability
How do users gain contextual understanding of the model's output?
Trust-Inducing Explainability
How do the downstream outcomes imply performance?

Fig. 8.1 Intersecting purposes for AI model explainability. One model may benefit from one or a combination of purposes depending on end-user and stakeholder needs

mathematical representation of the features is likely far less important to the clinical team than an understanding that the AI system detected the patient's deteriorating mental status and increasing respiratory rate over the past 12 h as factors contributing to the risk of needing mechanical ventilation. Such summative insight, along with information about the model's accuracy and how it was trained and validated may be enough to trust the AI system even without an in-depth trace-back of the inner workings of the model. To a layperson user of the AI system, such as the owner of the AI-enabled smartwatch in scenario three that detected an arrhythmia, understanding the context of the prediction may even be less important than having information to trust the system, especially if the information generated is only supplemental to an evaluation by a physician.

Sometimes, the same information about explainability can be applied to all three purposes of model engineering, enriching user understanding, and facilitating trust. For example, the commonly used **Pooled Cohort Equations** for the prediction of 10 year risk of atherosclerotic cardiovascular disease (**ASCVD**) is a linear regression that relies on variables including age, sex, race, blood pressure, cholesterol, history of diabetes, smoking history, and use of antihypertensive, statin, or anti-platelet medications [13]. These variables all happen to also be components of a patient's medical history that a physician would review to assess ASCVD risk, so knowledge of these variables and weights would fulfill the purposes of understanding the inner mechanics of the model as well as deriving clinically meaningful explanations for the patient's clinical condition and facilitating trust in the model prediction.

Given two models of equal performance, one a black-box and one an explainable model, most users, when asked, prefer the explainable model [14]. However, in many practical scenarios, models that may not be as easily interpreted can lead to better end user outcomes and may even be desirable in certain situations [15]. For example, when users are asked to accept or reject the price of a New York City apartment based on an explainable model, which tells them the features used such as the number of bedrooms and bathrooms, the distance to subways or schools) or a black box model which does not, the participants receiving explanation were more likely to accept wrong predictions than those who were shown the black box output.

In parallel, it is worth considering whether rigorous validation and high accuracy and consistency of the ML model alone could be sufficient for building user trust [16]. For example, one does not need to have an explainable model for a rain forecast as long as it is correct enough, often enough, to rely on to carry an umbrella. Trust in the model's output can be established by rigorous testing and prospective assessment of how often the model's predictions are correct and calibrated, and for assessing the impact of the interventions on the outcome. At the same time, prospective assessment can be costly.

A request for "explainability" in medical AI can be separated into a request for explaining model mechanics (perhaps better phrased as transparency of the modeling), a need by the user to understand the clinical context of the AI predictions, or a need to establish user trust. We will explore how principles of cognitive informatics can be applied to untangle the kind of explainability needed in a particular application.

Expanding the Conception of AI Explainability Based on Cognitive Informatics

As discussed in the section on "Three Purposes of AI Explainability", the current science has already established that explainability is not a single easily-defined construct that can be conceived as present or absent, and is substantially dependent on what "explainability" is meant to help the person receiving the "explanation" accomplish. In addition to the concepts of causality, surrogacy for trust and functional understanding of AI construction for an engineer, knowledge from cognitive informatics and related disciplines also suggest numerous nuances to "explainability" that need to be further explored. Specifically, theories of human information processing, conceptions of humans interacting with AI agents, and complex sociotechnical systems theory all suggest that there is yet much to be learned about how explainability is applicable to healthcare settings.

Human Information Processing

Consider the physician in the second example presented at the beginning of the chapter: while the physician having a difficult conversation with a patient about a poor prognosis, the physician is interrupted by an alert that suggests that a different patient in a different building is at risk of developing respiratory failure and may need escalation of care and mechanical ventilation. How will the physician respond? Will the physician leave the conversation? Seek more information? Dismiss the alert?

Models of human information processing have been part of the cognitive psychology and human factors literature for decades (see also Chaps. 1 and 5). Early models were developed in the 1950s and 1960s [17]. At their most basic, these human information processor models note that there's a layer of information processing between human perception of inputs/stimuli (involving encoding perceived stimuli in the context of mental models, comparing various options and choosing a response) and outputs/execution of action. Importantly, cognitive processing is both "top-down" and "bottom-up"—what we perceive and process is filtered by what our attention is directed towards. A commonly used model of human information processing in healthcare settings is the model in **situation awareness theory** (Fig. 8.2). Among other applications, situation awareness theory has been used in the improvement of the recognition of clinical deterioration in hospitalized patients as well as the diagnostic reasoning process [18]. The human information processing model underlying situation awareness theory resembles other goal-directed linear models of human information processing such as **Norman's theory of action** and **Rasmussen's decision ladder** [19–21]. This model suggests that when humans perceive information, they then comprehend the information (see discussion of mental models below), project the expected future states based on this information as well

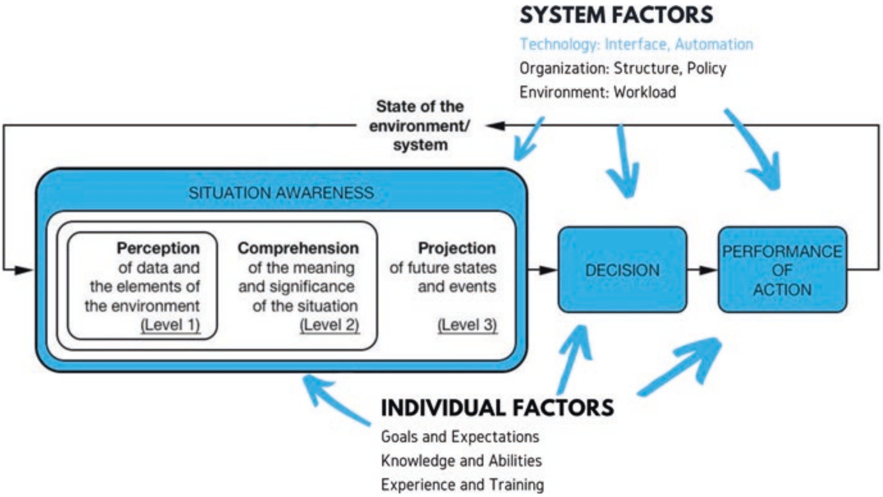
SYSTEM FACTORS
Technology: Interface, Automation
Organization: Structure, Policy
Environment: Workload
State of the environment/ system
SITUATION AWARENESS
Perception of data and the elements of the environment (Level 1)
Comprehension of the meaning and significance of the situation (Level 2)
Projection of future states and events (Level 3)
DECISION
PERFORMANCE OF ACTION
INDIVIDUAL FACTORS
Goals and Expectations
Knowledge and Abilities
Experience and Training

Fig. 8.2 Endsley's model of situational awareness. This model describes situational awareness in dynamic decision making and notes how technology such as AI can affect each step in human information processing [22]

as various choices that the person might make, then make a choice based on the desired future state, finally acting on the decision.

In our scenario, the physician may perceive the alert and see if there's any other information that can quickly allow for comprehension and verification of the current situation. If no supplemental information allows for verification of the alert, the only information that the physician will use to project the future state of the patient is what remains in the physician's memory. Any new information obtained by sensors or documented in the electronic health record that may be relevant and explain why the alert triggered will not be used by the physician. Because such information is never perceived, the physician may simply project that the patient's risk of respiratory failure has not changed and that the alert is simply incorrect, resuming their focus on the challenging conversation.

Comprehension of the alert and the situation does not happen in a vacuum. Human experts rely on mental models stored in long-term memory to translate perceived information into comprehension that can support reasoned projection of future states and subsequent decision-making. A mental model is a person's explanation of "how things work in the world" and allows one to predict what will happen if one takes a particular action (see also Chaps. 1 and 5). Experts are able to do this very efficiently by framing new information in the context of existing mental models built from experience (such as knowledge about disease processes, how previous patients with similar appearance have progressed in their illness, etc.).

So consider again how AI predictions alone, unaccompanied by additional explanatory information that matches the mental models of the user, may fail to produce action. For example, suppose that the alert in our example was received by a relatively inexperienced physician, and this physician knew that the alert was appearing for a patient admitted with a neurological concern. In this inexperienced physician, the mental model of neurologic abnormalities may not yet have linked "impending respiratory failure" to the possibility that such failure is a consequence of a neurologic problem interfering with the central respiratory control system. Because the physician experiences an apparent mismatch between the mental model of the patient (this patient will experience a deterioration with primarily neurologic changes such as altered mental status) and the alert (warning of impending respiratory failure), the alert may again be ignored in the absence of another explanation that helps the physician to establish quickly that the alert is not a false alarm.

Of course, human comprehension is not always a linear process. Other models of human comprehension, such as **Klein's data-frame theory** of sense-making and theories of information foraging in the human-computer interaction literature, explain that information processing is often an iterative process [23]. People tend to gather just enough information to "satisfice" and allow them to apply a mental model that helps them to understand the current situation [24]. If all relevant information had been gathered, a different mental model might have been applied, as a key piece of information may have reframed the situation. This also suggests an aspect of sufficiency for optimal stopping to the concept of explanation (see also Chap. 5). In the example discussed, if the most salient information presented with the risk prediction alert biases the physician towards framing the risk as a primary respiratory failure that is viewed as unlikely in the patient, the physician may be satisfied and dismiss the alert. On the other hand, if the explanatory information supplementing the AI risk prediction helps physicians frame the patient as potentially experiencing a neurologic disruption of the respiratory control system, they may be far more likely to seek more information and act.

Through all of these concepts of human information processing, explainability can be considered in the context of the "gulf of evaluation": the degree to which a person can use information to make sense of a situation and determine how well their goals have been met. To the extent that a user can perceive information or knowledge in an AI system and quickly make sense of the real world based on the explanation provided, they are more likely to make the optimal decision. If the explanatory information that allows the person to make sense of the situation is missing or requires substantial effort to glean, such as needing to click through multiple screen transitions, the person is much more likely to fail to appropriately use AI.

Human-AI Agents

In the previous section, we primarily conceived of explanation in AI systems as relevant at one point in time: when a person receives information from the AI system and may make a decision to act. However, AI systems can be complex enough, especially if there is a component of automation, that it can be conceived of as an independent agent. The "agency" of the AI system comes from implicit goals in any automated steps the technology might take (e.g. the automated referral for biopsy in the first example from the beginning of the chapter) as well as the "conversation" that occurs over a series of interactions between people and the technology that they are using. In this context, explanations may be considered in terms of their ability to make the AI agent's intentions and actions understood in order to produce predictable interactions that allow work towards a common goal [25].

The conception of humans and AI systems as agents that interact has been present from some of the foundational work in clinical informatics [26]. In the MYCIN and EMYCIN systems, the role of the AI agent was as a "consultant" for the clinical user (see Chap. 2). In Clancey's GUIDON project, the system was a tutor for students. In this work, an early discovery was the need for users to understand "why" the systems were acting as they did. The systems attempted to make explicit as explanation the internal goal and strategy of the AI system. Often these goals and strategies are implicit, which can be challenging for a human who seeks to interact with the system. In the absence of explicit explanation, people must infer goals and strategies for the agent with which they interact, which is even more challenging in the case of adaptive AI systems that change their behavior over time.

Conceiving of the interaction between humans and an AI agent as a "dialogue" was also established in the projects that were derived from MYCIN (see also Chap. 9). A conversation implies a conception of explanation tied to intelligibility. A phenomenon studied in aviation is that of "automation surprises", where the AI system acts in a way that is not expected and is thus not comprehensible to the person without an explanation. In this case, humans may assume that automation has failed, leading the person to take inappropriate actions. Another issue may arise when the set of actions that are offered do not match what the user is expecting, limiting the "conversational" nature of the interaction. For example, if the physician interacting with an AI-enabled sepsis alert is expecting to gather more information through lab testing to assess patient risk for sepsis but the system forces a decision about antibiotics before the physician can obtain that information, then the physician may wonder "why" the AI system is "recommending" antibiotics, even though the AI system is merely providing an incomplete set of potential choices.

As with human information processing, it is beyond the scope of this chapter to review all of the ways that explanations may function in a back-and-forth series of interactions between human and AI agents. Appreciation of humans as agents interacting with potential AI agents, however, suggests there is much still to be learned about the role of explanation in such interactions. Explanation is critically important so that a person can predictably interact with an AI system to achieve one's goals.

Sociotechnical Systems

When considering the impact of AI systems on health care, we need to examine the broader care delivery system in which the AI technologies are being incorporated. Healthcare is complex - meaning that the delivery of health care, whether it is diagnosing and identifying disease, providing therapies, or implementing interventions to prevent disease, occurs in complex sociotechnical systems. Complexity exists because (1) there are numerous relationships/interactions among the many entities that are involved, and (2) health care is a human-driven process (since care is provided by and for humans, and human behavior is adaptive to changes in the environment). While the previous section introduced the idea of interaction between the human and the AI system, outcomes in health care are mediated by numerous nonlinear interactions between people, processes, and technologies. With nonlinear systems, changes in the input do not always lead to proportional changes in output. Outcomes from complex systems cannot be predicted by examining the properties of just one component of the system. The system must be examined as a whole, and we need to assess how changes to any particular component impact the rest of the system in order to understand how it could affect the outcome. These outcomes are known as emergent properties as they emerge only when the system exists as a whole but not within or between any individual components.

Health care has made important progress in recognizing that processes and outcomes that we see are not products of individual actions but instead emerge from a complex set of interactions between people, the tools they use, and the processes/organization of their work environment. Patient safety is a great example of what is now often conceptualized as an emergent property of the care delivery system. The Systems Engineering Initiative for Patient Safety (**SEIPS**) model is now well established in health care and has been applied to numerous healthcare projects to design tools and technology in healthcare delivery (Fig. 8.2) [27]. Patient safety, defined as the prevention of unintended harm to the patient, cannot be attributed to any one part of the work system alone but emerges from how each part of the care delivery system interacts with the others (Fig. 8.3).

When introducing AI systems to improve health care, we need to think about how the system changes the existing sociotechnical work environment to mediate the outcome. In other words, how does the AI system interact with the other people, processes, and technologies (including other AI systems)? A common assumption in health care is that digital tools improve the reliability of care because humans are error-prone. In reality, any introduction of technology adds new components and thus new “failure” points for safety. In order to function effectively, the person in the sociotechnical environment needs the AI system not just to explain its goals and intentions, but also to convey how it is interacting with the other elements of the sociotechnical environment.

Trust is a concept that is discussed often in the context of explainable AI—understandably, as it may be essential to sustained use of any given AI system. While trust is traditionally viewed as a property of the human-AI interaction, it may also be useful to conceive of trust as another emergent property like patient safety. Over time, when people observe interactions and the outcomes of interactions with AI systems in their work, people will develop a set of expectations on how to best interact with AI systems. This will come not just from their own experience but also through observing other people, socialization of the technology in the popular press and the culture of their work organization. If these expectations are violated without

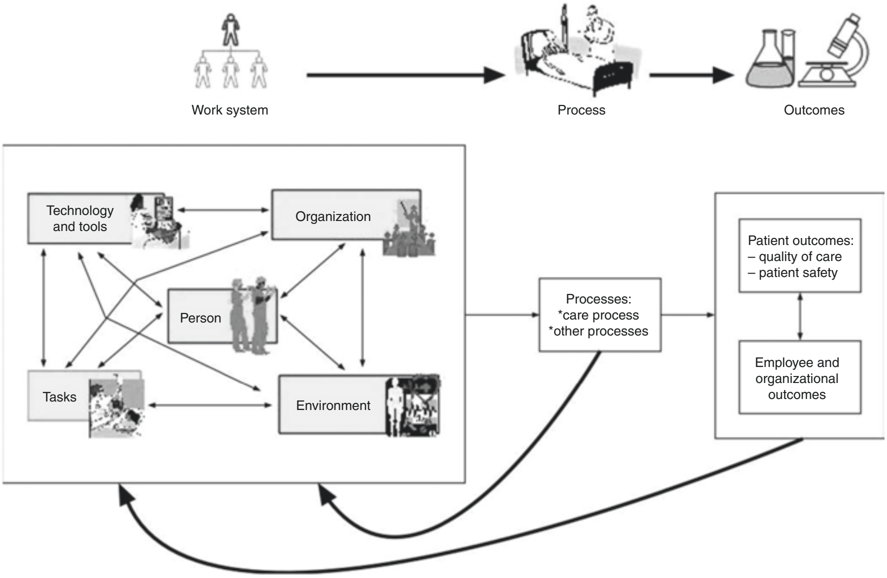

Fig. 8.3 The systems engineering initiative for patient safety (SEIPS) model. An example of a framework for understanding the structures, processes and outcomes in health care and their interactions, which result in the 'emergent' outcomes such as patient safety [27]

explanation, trust in a given AI system can be lost. Furthermore, knowing that attention to information processing has a "top-down" component (i.e. humans pay selective attention to information they are already expecting), subtle unexpected changes in the system may never be perceived by the user. For example, consider the example of the AI system that automatically screens chest CTs for lung cancer and refers patients for biopsies. If the threshold for referral is updated with only a subtle notation and without explanation to the patient or physician, the change in system's behavior may be noticed only when there are dramatic outcomes, such as a large increase in the number of referrals or a missed referral in a patient with lung cancer. At that point, physicians and patients may lose trust in the system.

This is why trust is best assessed not as "present" or "absent" but in terms of whether trust is appropriately calibrated. Under-trust is traditionally the focus of healthcare AI literature, given the limited adoption of AI systems to date. Much effort is spent on increasing trust in AI systems because of the adoption problem. However, over-trust is just as important. Over-trust occurs when the clinician comes to depend on the AI system, either because of the clinician's lack of expertise or because other pressures from the sociotechnical system such as the workload potentially drive inappropriate reliance on technology. A goal of explanation for AI systems is to optimize the calibration of trust. The goal is neither under-trust when the AI system is enabling the correct action but also not over-trust when the AI system is operating outside of its optimal scope or the user should not trust the AI system without further patient assessment.

This understanding of sociotechnical systems suggests that AI explanations may not translate across clinical contexts in which the AI system might be used. Rather, local customization of the explanation and user interface may be necessary to situate the AI system in the sociotechnical environment. Design principles that incorporate contextual constraints such as ecological interface design may be useful on this front. To the extent that the interface can mirror the external world, and the AI system's actions are placed within the context of that external world, people can be more successful in their essential roles for patient safety: anticipating errors and acting as a source of resilience.

Ultimately, the lenses of human information processing, human-AI agents, and sociotechnical systems suggest that there may be no single "universally" suitable design for optimal AI explanations and approaches to AI explanation will continue to evolve.

Implications of Explainability on Bias and the Regulatory Environment

It is important to understand the effect of explainability on accountability and the regulation of AI. As mentioned, complex AI systems often include elaborate data transformation using hundreds of thousands of features derived from clinical data that may or may not be meaningful. As these complex systems increasingly drive clinical decisions, it is important to acknowledge the legitimate concerns about the intentional and unintentional consequences of these AI systems. Explainability presents an opportunity to understand better the changing landscape of accountability and regulation.

Explainability and Inherent Biases

An emerging body of evidence suggests that AI systems can make unfair and discriminatory decisions, replicate or develop biases, and behave in inscrutable and unexpected ways in highly sensitive environments that put human interests and safety at risk. Therefore, it is important to consider how explainability may mitigate such biases and affect our own inherent biases using the three purposes of explainability mentioned in section "Three Purposes of AI Explainability".

For the ML engineer, explainability offers an opportunity to identify and mitigate potential biases in AI systems. However, this is only possible through the transparent reporting of the AI details, such as recommended by the Minimal Information for Medical AI Reporting, or the **MINIMAR standards** [28]. Such standards demand information related to the data source and cohort selection, demographics of the training population, model features and design characteristics, as well as evaluation criteria and validation steps. Data biases (see Chap. 3) are common across most data sources, therefore the transparent reporting of the MINIMAR concepts ensures that the interpretability of the model and can help end users (e.g., providers, healthcare systems, etc.) to identify the populations for which the AI can be applied—a step at mitigating inherent data biases. Furthermore, having a clear definition of the model output is imperative. Datasets with inconsistent, imperfect, or even incorrect labels used for training and testing data allow one's own biases to enter into the model and affect explainability. Understanding the intent of the model (e.g. predict mortality or predict transfer to ICU), composition of the training data, development of the ground truth, model architecture, and data transformation enables the ML engineer to assess biases in the data and algorithmic fairness empirically. While the model architecture may not be critical for explainability, understanding sample representativeness in AI models may help clinicians decide if the prediction is applicable to their patient population. For example, many models predicting adverse events in Type II diabetic (T2DM) patients have not reported Hispanics in their training data. Given Hispanics higher prevalence and complication rates for T2DM compared to Whites, it is essential clinicians have this type of explainability to determine if they incorporate the AI prediction into their clinical decision making. This type of explainability is the foundation of developing trust for both the clinician as well as the patient population, as it is only through transparency and explainability that one can mitigate biases in AI that contribute to unfair and discriminatory decisions that put human interests and safety at risk.

Effect of Explainability on Accountability for Decision Making

Accountability, in this context, means the ability to determine whether a decision was made in accordance with procedural and substantive standards and to hold someone responsible if those standards are not met. Therefore, it is important that developers understand and integrate current standards within the model's design and during development. Explainable models must be developed through team efforts involving knowledge experts, decision makers, and end-users. The incorporation of procedural and substantive standards must be clearly presented to end-users across platforms.

Human-interpretable information about the factors used in a decision and their relative weight is necessary. This is likely the most common understanding of what constitutes an explanation for a decision. A list of the factors that went into a decision, ideally ordered by the significance to the output, can provide accountability by confirming that proper procedures were followed.

While there is significant support for explanations as a tool for holding AI accountable, there are also concerns about the costs of generating such explanations. True explainability could inhibit innovation by forcing transparency around key model features which may be seen as industry trade secrets. However, the lack of incentives, restrictions around data sharing and data privacy, and the acceptance of stealth science in industry has created an environment that allows AI to be implemented without understanding how the model was developed, from what data was the model learned, and using what data was the model deemed satisfactory for use. Accountability can only fairly proceed if transparency is provided regarding the data, model, and standards woven into the AI model.

The Current Regulatory Framework and Explainability

Across the global AI regulatory environment, explainability is the center of accountability (see Chap. 18). In 2017, the International Medical Device Regulators Forum (**IMDRF**) came together to develop a path for standardized AI regulations, including a risk-based framework [29]. The hypothetical scenarios described in the introduction provide examples of the different levels of risk to consider and the importance of explainability in each scenario. The level of regulation and necessary documentation are determined by the risk-based framework, as described above. In addition to the risk-framework, the European Union has put forth the General Data Protection Regulation (**GDPR**) mechanism which ensures users (or patients) have a right to information about the existence, logic, and potential consequences of AI-driven decision-making systems. The GDPR establishes rules and regulations for privacy and permissions and gives control to individuals. Patients must not only consent to the collection of the data but also to each use of their data. For AI developers, this requires that they explain *in plain language* how data will be used as part of the consent process. Many interpret this as the “right to explanation”. Systems are now aiming to produce more explainable models; design an explanation interface; and understand the human requirements for useful explanations [30]. However, there are also concerns about the costs of generating explainable AI in regard to engineering challenges, the effect on innovation and trade secrets; and the cost of system accuracy or other performance objectives.

Application of Explainability to Real World Examples of Medical AI

The following real world examples of medical AI can be used to understand the differences among the purposes of explainability, targets and downstream actions among the three methods of explainability, as well as how the cognitive informatics concepts we have described apply to particular use cases of medical AI.

Example: Continuous Blood Glucose Monitoring for Patients with Type 1 Diabetes

Current Type 1 Diabetes management approaches are largely limited to non-closed loop systems that depend on the patient checking blood glucose levels and administering themselves insulin either through a pump or a syringe. The iLet bionic pancreas from Beta Bionics [31] aims to simplify the latter, interfacing with an embedded glucose monitor and continuously dosing insulin similar to a native pancreas. It needs to be able to adapt to blood glucose variation patterns and function autonomously. The end users of this device are endocrinologists and their patients who will likely not benefit from a mechanistic or 'engineer's' explainability but rather on 'trust inducing explainability'—relying on outcome data that shows a closed loop feedback system for glucose control has minimal hypoglycemic events and maintains glucose levels within an acceptable range that is conducive for improved long term outcomes. From a cognitive informatics perspective, human information processing needs with this system are likely very different from how patients are counseled about diabetes management now, learning to "count carbohydrates" and estimate how much insulin to self-administer. However, explanation for this highly autonomous agent may need to convey information like how the insulin administered is based on the blood sugar goal or how well the overall blood sugar control has been, allowing a patient to not only monitor the system but also troubleshoot and recover from malfunctions without experiencing life-threatening hypoglycemic or hyperglycemic events. From a regulatory perspective, given this automated closed-loop system, and lack of a physician intermediary while care is being delivered, the regulatory concerns are high although this device would be categorized as a medical device rather than **SaMD** (software as a medical device) per the previously mentioned framework put forth by IMDRF.

Example: Digital Image Analysis Tools Assisting in Histopathological Diagnoses

Proscia's digital pathology tools are designed to drive clinical management by analyzing pathology samples and prioritizing certain cases for review by pathologists, especially cases that are flagged by the system to demonstrate high risk features. For example—biopsy samples of precancerous lesions that have high risk features are prioritized for expedited review to allow for earlier management of a potential cancer diagnosis. Such a system, which is tasked in prioritizing certain cases for review for the end user (in this case a Pathologist), will need to demonstrate the reasons for prioritization. In this case, the Pathologist is tasked with making the final diagnosis. Clinical data may help with contextual explainability but these tools may benefit most from having surrogates for the underlying mechanistic ‘engineering’ processes. In the case of AI/computer vision systems, segmentation and bounding box techniques can assist in establishing this surrogate, but may need significant more time invested into labeling training data. Image classification techniques that do not have segmentation, object detection, or bounding boxes built in will need to depend on context to engender trust in the end user. These tools have to rely on proper functioning within a human-AI team; the level at which this tool can operate autonomously should be carefully conveyed to the healthcare team to optimize calibration in such a tool. This will avoid under-trust leading to under-utilization and over-trust leading to an inappropriate amount of dependency. From a regulatory perspective, digital pathology tools fall under Category II of the IMDRF framework as they drive clinical management of serious conditions and would likely benefit from independent review.

Example: Wearable Devices Informing Clinical Management

Finally, we offer an emerging use-case scenario where wearable data is informing clinical management. Wearables are starting to incorporate not only heart rate information to show variability, and correlate the rate to motion sensors to determine types of activity and levels of sleep, but also to oxygen sensors and basic one lead rhythm monitors such as the **ECG** App on the Apple Watch. This particular app is designed to detect atrial fibrillation, low and high heart rates, and to provide a summary of heart rate variability. Algorithms that assist in aggregation of clinical data will likely need to depend on causal explainability—information that becomes important with context. In the case of an ECG app assisting in detection of atrial fibrillation, the output is to be taken in conjunction with patient data—possibly complaints of palpitations or indications of a history of cardiac disease that would predispose to a diagnosis of atrial fibrillation. Human information processing models are important for information that is best analyzed in context (in this case, patient symptoms and history). Along with contextual explanatory information presented with the model output, the end users’ experiences within their sociotechnical environment will also drive each user’s trust in the prediction and subsequent decision-making. Regulatory concerns with wearables that inform clinical management are largely dependent on the manufacturer and on the element for which it chooses to obtain clearance. The ECG app has **FDA** clearance as a Class II device but the pulse oximeter function is described for ‘general wellness’ and thus does not have FDA clearance as a medical device.

Conclusion

The question of whether explainability is useful for medical AI must be expanded to include considerations around (1) the type and purpose of the explanation and (2) the type of human-machine interaction in which explainability may play a role in mediating the desired outcome. Finally, the degree of explainability may impact how bias and accountability are incorporated into the medical AI product and how it may be regulated.

Questions for Discussion

- How can the different methods and purposes of explainability be applied to different AI use cases?
- How do principles of human information processing and information flow across teams affect how AI explainability should be approached?
- What frameworks are important to consider for an AI agent, with a similar underlying model, that is deployed across different environments?
- What potential pitfalls might there be with the current regulatory framework with regards to AI explainability?
- Who are the common stakeholders and what motivations do they have with regards to accountability in AI systems?
- What are some of the potential underlying causes of unintended intrinsic biases within AI systems?

Further Reading

General Data Protection Regulation (GDPR), https://gdpr-info.eu/.

- Official EU documentation of General Data Protection Regulation, including recitals and key issues. This regulation for consumer privacy is a reference for other countries and regions as they craft their own versions. E.g. California Consumer Privacy Act has many similarities with GDPR.

Gilpin, L.H., Bau. D., Yuan, B.Z., et al. Explaining Explanations: An Overview of Interpretability of Machine Learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 80–89. Available at https://arxiv.org/pdf/1806.00069.

- An exploration into best practices of explainability, the insufficiency of current approaches and future directions for explainable artificial intelligence. Being aware of the work being done in the non-clinical realm will help inform efforts with regards to explainable medical AI.

Markus, A.F. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics. 2021;113:103655. https://doi.org/10.1016/j.jbi.2020.103655.

- In this paper is an exploration of quantitative metrics regarding explainable AI. Although the field is far from a consensus, having quantitative metrics will allow for model comparison with regards to explainability similar to how model performance is compared today.

Carayon, P., Hundt, A.S., Karsh, B.T., et al. Work system design for PATIENT safety: The SEIPS model. Qual Saf Health Care. 2006 Dec;15 Suppl 1(Suppl 1):i50–8. http://dx.doi.org/10.1136/qshc.2005.015842.

- This paper provides an overview for the Systems Engineering Initiative for Patient Safety framework, which is applied to describe complex work systems in healthcare and provides a tool to examine the context for the downstream impact of explainable AI in healthcare workflows.

Brady, P.W., Wheeler, DS, Muething, S.E., Kotagal, U.R. Situation awareness: A new model for predicting and preventing patient deterioration. Hosp Pediatr. 2014;4(3):143–6. https://doi.org/10.1542/hpeds.2013-0119.

- This paper describes an example of how AI is applied to a use case in pediatrics, and how explainability facilitates a team dynamic that was important in mediating the outcome.

References

1. Teach L, Shortliffe H. An analysis of physician attitudes regarding computer-based clinical consultation systems. Comput Biomed Res. 1981;14(6):542–58. https://doi.org/10.1016/0010-4809(81)90012-4.
2. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med. 2020;3(1):47. https://doi.org/10.1038/s41746-020-0254-2.
3. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the 2018 CHI conference on human factors in computing systems. Montreal: ACM; 2018. p. 1–18. https://doi.org/10.1145/3173574.3174156.
4. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th international conference on data science and advanced analytics (DSAA). Turin: IEEE; 2018. p. 80–9.
5. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655. https://doi.org/10.1016/j.jbi.2020.103655.
6. Montavon G, Vedaldi A, Hansen LK, Muller K-R. Explainable AI: interpreting, explaining and visualizing deep learning. New York: Springer; 2019. p. 439.
7. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Doshi-Velez F, Fackler J, Jung K, Kale D, Ranganath R, Wallace B, et al., editors. Proceedings of the 4th machine learning for healthcare conference. Ann Arbor: PMLR; 2019. p. 359–80.
8. Turek M. Explainable artificial intelligence. Defense Advanced Research Projects Agency. 2021. https://www.darpa.mil/program/explainable-artificial-intelligence.
9. Top Explainable AI Companies. Venture Radar. Available from: https://www.ventureradar.com/keyword/Explainable%20AI.

10. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284–6. https://doi.org/10.1126/science.abg1834.
11. Miller K. Should AI Models be explainable? That depends. HAI. 2021. Available from: https://hai.stanford.edu/news/should-ai-models-be-explainable-depends.
12. Slack D, Friedler S, Scheidegger C, Roy CD. Assessing the local interpretability of machine learning models. Workshop on human-centric machine learning at the 33rd conference on neural information processing systems. 2019.
13. Muntner P, Colantonio LD, Cushman M, Goff DC, Howard G, Howard VJ, et al. Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations. JAMA. 2014;311(14):1406. https://doi.org/10.1001/jama.2014.2630.
14. Lipton ZC. The mythos of model interpretability. Queue. 2018;16(3):31–57.
15. Holm EA. In defense of the black box. Science. 2019;364(6435):26–7.
16. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Wortman Vaughan JW, Wallach H. Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI conference on human factors in computing systems. Yokohama: ACM; 2021. p. 1–52.
17. Hilgard ER, Bower GH. Theories of learning by Ernest R. Hilgard and Gordon H. Bower. Appleton-Century-Crofts; 1966.
18. Brady PW, Wheeler DS, Muething SE, Kotagal UR. Situation awareness: a new model for predicting and preventing patient deterioration. Hosp Pediatr. 2014;4(3):143–6. https://doi.org/10.1542/hpeds.2013-0119.
19. Singh H, Giardina TD, Petersen LA, Smith MW, Paul LW, Dismukes K, et al. Exploring situational awareness in diagnostic errors in primary care. BMJ Qual Saf. 2012;21(1):30–8. https://doi.org/10.1136/bmjqs-2011-000310.
20. Norman DA. The psychology of everyday things. New York: Basic Books; 1988.
21. Rasmussen J. Information processing and human machine interaction: an approach to cognitive engineering. New York: North-Holland; 1986.
22. Endsley MR. Toward a theory of situation awareness in dynamic systems. Human Factors. 37(1):32–64.
23. Klein G, Phillips JK, Rall EL, Peluso D. A data-frame theory of sensemaking. Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making; 2007, pp. 113–55.
24. Simon HA. Rational choice and the structure of the environment. Psychol Rev. 1956;63(2):129–38.
25. Lewis M. Designing for human-agent interaction. AIMag. 1998;19(2):67.
26. Clancey WJ. From guidon to neomycin and Heracles in twenty short lessons. AIMag. 1986;7(3):40.
27. Carayon P, Schoofs Hundt A, Karsh B-T, Gurses AP, Alvarado CJ, Smith M, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care. 2006;15(1):50. https://doi.org/10.1136/qshc.2005.015842.
28. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27(12):2011–5. https://doi.org/10.1093/jamia/ocaa088.
29. Sun J, et al. IMDRF essential principles of safety and performance of medical devices and IVD medical devices introduction and consideration. Zhongguo Yi Liao Qi Xie Za Zhi. 2021;45(1):62–6.
30. Gunning D. Broad agency announcement explainable artificial intelligence (XAI). Technical report. 2016.
31. Jafri RZ, Balliro CA, Sherwood J, Hillard M, Ekhlaspour L, Hsu L, Russell SJ. 77-OR: first human study testing the iLet, a purpose-built bionic pancreas platform. Diabetes, 2019;68(1):77. https://doi.org/10.1016/S2213-8587(15)00489-1.