Psychology and AI at a Crossroads: How Might Complex Systems Explain Themselves?

ROBERT R. HOFFMAN
Institute for Human and Machine Cognition

TIMOTHY MILLER
University of Melbourne

WILLIAM J. CLANCEY
Institute for Human and Machine Cognition

A challenge in building useful artificial intelligence (AI) systems is that people need to understand how they work in order to achieve appropriate trust and reliance. This has become a topic of considerable interest, manifested as a surge of research on explainable AI (XAI). Much of the research assumes a model in which the AI automatically generates an explanation and presents it to the user, whose understanding of the explanation leads to better performance. Psychological research on explanatory reasoning shows that this is a limited model. The design of XAI systems must be fully informed by a model of cognition and a model of pedagogy, based on empirical evidence of what happens when people try to explain complex systems to other people and what happens as people try to reason out how a complex system works. In this article we discuss how and why C. S. Peirce's notion of abduction is a best model for XAI. Peirce's notion of abduction as an exploratory activity can be regarded as supported by virtue of its concordance with models of expert reasoning that have been developed by modern applied cognitive psychologists.

KEYWORDS: abduction, Peirce, artificial intelligence, explanation, expert reasoning, training

Artificial intelligence (AI) systems are being applied ever more widely (e.g., helping corrections officers who have to make parole decisions helping financiers who decide about loan applications). A regulation posed by the European Union (Goodman & Flaxman, 2016) asserts that users have a "right to an explanation" concerning algorithm-based systems. For decision makers who rely on analytics and data science, explainability is also a pressing issue. They need to be confident that their decisions are reasonable when they rely on the outputs of machine learning systems (sometimes called deep nets or black boxes). People whose lives are affected by AI systems, and people who rely on AI systems to do their work, need to understand how the AI works, the mistakes it can make, and the safety measures surrounding it. Numerous computer scientists have advocated a requirement that AI systems be explainable, under-

American Journal of Psychology
Winter 2022, Vol. 135, No. 4 pp. 365–378

standable, transparent, and interpretable (Lipton, 2016; Miller, Howe, & Sonenberg, 2017). This has motivated a number of research projects in what is now called explainable AI (XAI). Work in this area gained significant momentum from funded programs, including a focus on AI explainability on the part of the U.S. Department of Defense, beginning about 2016 (see Gunning & Aha, 2019).

The design of XAI systems must be fully informed by a psychological model based on empirical evidence of what happens when people try to explain complex systems to other people and what happens as people try to reason out how a complex system works (Miller, 2019). Therefore, the creation of XAI systems necessarily invokes a collaboration between psychologists and AI system developers.

Recent discussions of explainable AI have not much considered the role of abduction in explanatory reasoning (see Mueller, Hoffman, Clancey, Emrey, & Klein, 2019). In this article we discuss why and how C. S. Peirce's notion of abduction applies to XAI systems. We then elaborate Peirce's model based on findings about expert reasoning, which have been empirically derived by applied cognitive psychologists.

Explainable AI Systems

For the developers of AI systems, an explanation has to express a formal justification for why a system was architected the way it was. Developers need to know that the AI's algorithm produces correct solutions. When explaining their XAI systems to other computer scientists, system developers seek "transparency" and "interpretability" (Biran & Cotton, 2017; Doshi-Velez & Kim, 201; Lipton, 2016). It is important to note that these are defined formally, not used in the ordinary senses of the words. A system is interpretable if the algorithms can be modeled by some other, simpler and well-understood formal system. Hence, the explanations generated by XAI systems have taken such forms as numerical matrices of feature weights, decision trees, rule hierarchies, and Bayesian probability networks (see Gunning, Vorm, Wang, & Turek, 2021; Mueller et al. 2019).

This is fine, but a formal analysis that works for a computer scientist will usually make no sense to users. In some XAI research it was initially assumed that a good explanation for a computer scientist would serve as a good explanation for users. First attempts to demonstrate the effectiveness of formalist explanations (for improving the user's performance) had mixed results but showed that "explanations are more helpful when an AI is incorrect and are particularly valuable for edge cases" (see Gunning et al., 2021, p. 8). This meant that formal interpretability is not enough (see also Chowdhury & Lake, 2018).

Prospective adopters and users need to know that the AI will work well when it is deployed. Users need to understand how the AI works, expressed in everyday language or easily understood graphics (see Krause, Perer, & Bertini, 2016). Users want AI systems to provide explanations that accord with their current context and that increase trust in and understanding of the AI system. This includes information about the features that a classifier uses, about how the classifier works, and about how certain the AI is its determinations (Lim & Dey, 2009).

In other words, what counts as a good explanation depends on the intended beneficiary of the explanation and their context and goals.

The initial scheme for research that was intended to evaluate the performance of XAI systems is presented in Figure 1 (see Mueller et al., 2019).

This model was a useful starting point for developing experimental designs and procedures for empirical evaluation, especially since it showed some of the things that would have to be measured (the dark nodes in Figure 1). However, the model does not take into account the findings from psychological research on how people explain complex systems to other people (see Hoffman, Klein, & Miller, 2011; Klein, Hoffman, & Mueller, 2019; Klein, Hoffman, et al., 2021). Nor does it take into account the challenges that were discovered in the attempts, dating to the 1980s, to create intelligent tutoring systems (ITSs; see Clancey & Hoffman, 2022). The XAI model was essentially spoonfeeding: The AI generates an explanation and delivers it to the user, the user understands it, then the performance of the human–AI system improves. At least, this was the researchers' working assumption.

Whether considered from a formal perspective or a lay perspective, the avowed goal is for the XAI to provide correct, understandable, and sufficient explanations. (For reviews, see Biran & Cotton, 2017; Doshi-Velez & Kim, 2017; Gerlings, Shollo, & Con-

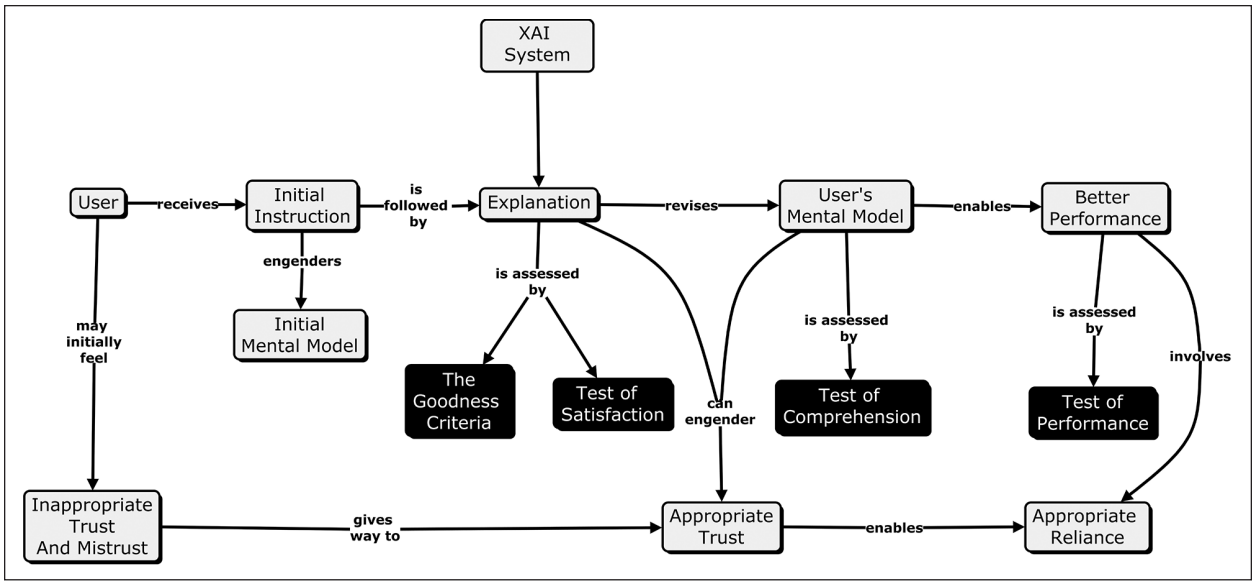

FIGURE 1. The initial XAI process model

stantiou, 2021; Mueller et al., 2019; Walton, 2011.) It is generally agreed that "enhancing the explanatory power of intelligent systems can result in systems that are easier to use, and result in improvements in decision-making and problem-solving performance" (Nakatsu, 2004, p. 575).

This article presents a synopsis of what C. S. Peirce said about abductive reasoning, from logical and psychological perspectives. A reconciliation of the logical and psychological perspectives on abduction provides a definition that applies to explainable AI. We then expand that Peircean model by noting its concordance with what has been discovered in research on the reasoning of experts.

Abduction in Philosophy

Abduction as the process of explanation has been widely discussed and cited as a central aspect of critical thinking, cognitive flexibility, fluid intelligence, and creativity (e.g., Douven, 2011a, 2022; Moore & Malinowski, 2009). Abduction has been defined as an accomplishment, an ability, a skill, and an amalgam of component skills such as evaluating evidence, forming mental models, and recognizing cues. Many scholars have discussed hypothesis formation and testing as the basis of scientific reasoning (e.g., Bruner, 1985; Collins, 1985; Glaser, 1984; Pfeiffer, Feinberg, & Gelber, 1987; Hempel, 1965; Nummedal, 1987; Selz, 1935).

The logical concept of abduction might actually be traced to Aristotle. In *Prior Analytics* (69a 20ff; Smith, 1989), Aristotle discussed what he called "reduction": the transformation of the subject–predicate relations in the major and minor premises of a syllogism (e.g., "Socrates is mortal," "Socrates is a man"), so that each of the four basic types (or "figures") of syllogisms could be expressed as a type in which the conclusion is an explanatory rule, or we might say an assertion about a class (e.g., "All men are mortal"). Despite this historical precedent, most discussions attribute the concept of abduction to the early American philosopher–scientist Charles Sanders Peirce.

In classical formal logics, deduction and induction are completely defined over sets of assertions, one of which must be a generalization (that is, an assertion about a class). The standard forms are presented in Table 1.

In contrast, abduction depends on propositions from the reasoner's knowledge—propositions that come from beyond the given rule and the given observations. Hence, abduction is not the same as inductive enumeration. Nevertheless, to philosophers it is a form of inference (Douven, 2011a; Harman, 1965; Lipton, 2004; see also Hempel, 1965). A proposed explanation is correct if the observed circumstance would necessarily occur. In this view, abduction would be distinct from both deduction and induc-

TABLE 1. Standard Syllogism-Style Representations of Deduction and Induction

Deduction	
Rule: All instances of x belong in Class X.	All humans are mortal.
Observation: Q is an instance of x.	Socrates is a human.
Inference: Q belongs in Class X.	*Therefore*
	Socrates is mortal.
Induction	
Observation: Instance Q of x belongs in Class X.	Socrates is human and a mortal.
Observation: Instance R of x belongs in Class X.	Pliny the Elder is human and a mortal.
Observation: Instance S of x belongs in Class X.	Plato is human and a mortal.
Inference: All instances of x belong in Class X.	*Therefore*
	All humans are mortal.

tion. Indeed, it would be a third basic or fundamental type of inference. One might say something like:

> If H (hypothesis) is a correct model of the system of interest, then the assertions D (facts) are logically consistent with H's structure/processes, i.e., D is modeled by H; H models D.

This would retain abduction in the province of classical logic. But Peirce described logical inference in a way that differed from the logics of Aristotle and Hume, and he was not consistent in his writings on this topic (see Beckwith, 2018). He did not seem to mean that abduction was a third classical type (in addition to deduction and induction). Rather, he argued that one single type—that a given case or set of cases follow a rule—covers the two classical forms.

Figure 2 demonstrates the difference between deduction, abduction, and induction. Deduction starts with some generalized knowledge (i.e., the rule: "All men are mortal") and an instance ("Socrates is a human") and makes an inference about that instance ("Socrates is mortal"). Induction, on the other hand, takes instances (usually many instances) and a property about the instances and forms a rule that would apply to future cases. For example, if we see many people (instances) and observe that they are all mortal, we infer that all people are mortal. Thus, it can be said that induction is the same as generalization. To Shank (1998, pp. 847–848), abduction is "rendering what might be thought of as a unique experience into an instance of a more general phenomenon."

There are particular problems for logic here. For example, since any given case will have an unbounded number of hypothetical features (potential premises) about which one might form rules, one can have an unbounded number of rules (see Douven, 2011a). Another problem is that the premise in deduction (the given rule) is itself the result of induction. A third is that abduction is question-begging because it entails a miracle (i.e., the spontaneous creation of a hypothesis; see Niiniluoto, 1999). While these are problems for logic, they do not seem to have been a concern for the Peirce the pragmaticist.

Peirce referred to abduction as "hypothetic inference"—the inferring of a plausible explanation for an observed circumstance. One starts with an interesting event or phenomenon and conjures (perhaps created de novo or perhaps previously derived using induction) to make an inference about that instance (one would assume, an inference about causation). Recent philosophy has maintained the close linkage to classical logic in treatments of abduction. For example, Josephson and Josephson (1995) described abduction in a formal way (see also Aliseda, 2007; Haig, 2009; Minnameier, 2010):

> D is a collection of facts, observations, or "givens,"
>
> H explains D, that is, H would, if true, explain D,
>
> No other hypothesis can explain D as well as H does,
>
> Therefore, H is probably true.

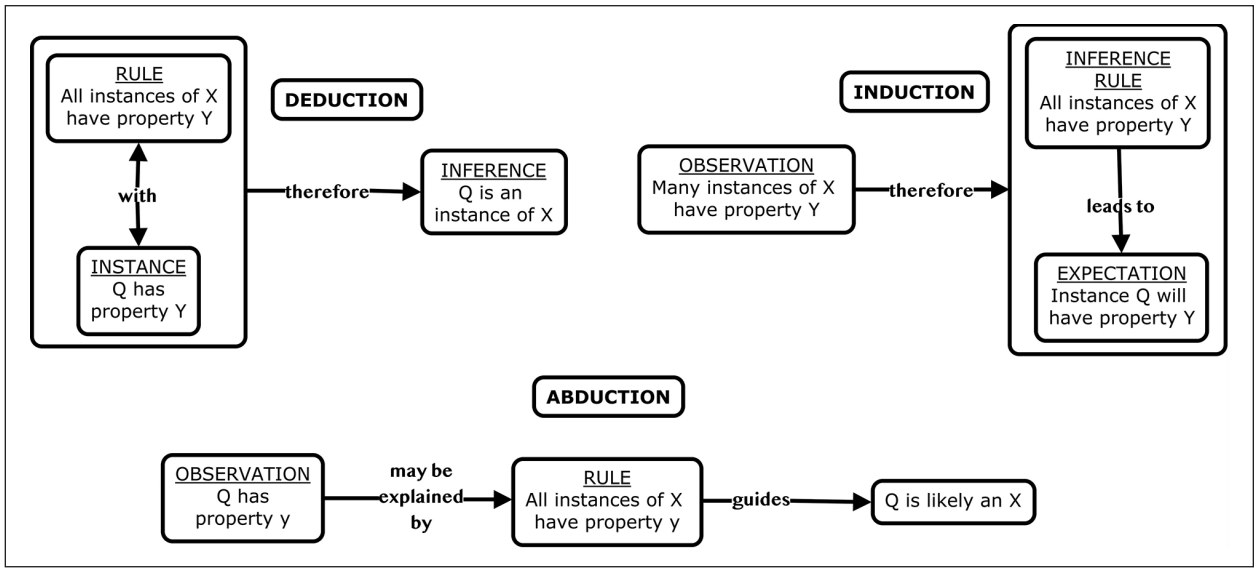

FIGURE 2. The difference between deduction, induction, and abduction

This final proposition supposes the rejection of all but one hypothesis. Abduction is commonly defined as "inference to a best explanation." This is reflected in the third proposition in the Josephson and Minnameier formalism. Peirce's work is precursor to this notion, but it is not strictly his definition of abduction, and the identification of the two poses dangers for philosophy of science (see Douven, 2011b; Thagard, 1978). (For reviews, see Niiniluoto, 1999; Macauliffe, 2015.)

As we now discuss, the concept of abduction also has a legacy in the field of AI.

Abduction in the Field of Artificial Intelligence

If abduction involves creativity, the discovery of new ideas, and insight, could it be formalized and implemented computationally? In the field of AI, abduction has been considered a form of induction, as in case-based reasoning (see Lenat, 2013, p. 54). However, a number of AI researchers have not identified abduction with induction, and have been investigating the process of abductive reasoning, for topics such as fault diagnosis and medical diagnosis. This led to mechanizations and formalizations of aspects of the abductive reasoning process that clearly distinguish it from induction and gave it an empirical backing (see Pople, 1973).

Abduction has been referred to as the process of inferring causal processes based on observations, for example, in the literature on medical diagnostic systems (see Finin & Morris, 1988). One of the best examples of the implementation of abductive reasoning is that of Fraser et al.'s (1989) project to develop an ITS for antibody identification in immunohematology. They conducted knowledge elicitation interviews with a domain expert. The interviews focused on the recall of critical incidents in which an initial diagnosis proved wrong. The researchers' primary intent was to pin diagnostic errors on one or another cognitive bias (e.g., overestimating the likelihood of a given hypothesis). (While the researchers were able to do this, they also learned that the expert was aware of possible biases and deliberately performed checks on her probabilistic reasoning by such means as examining base rates.) The expert's process of inferring a diagnosis based on the available evidence was referred to as abduction, but otherwise the concept of abduction (as an act of inference) played no role in the research. That said, the description of the expert's process for testing hypotheses fits with the Peircean model (i.e., explicit check on the lab technician's process for performing a test to develop converging or disconfirming evidence).

ITSs have the same goal as XAI: to support the sensemaking process. An ITS that comes close to the Peircean model is an ITS for physics problem solving developed by Makatchev, Jordan, and VanLehn (2004a, 2004b). Given that learner's mental

models of complex systems are likely to be incomplete and inconsistent (diSessa, 1993, 2018), the ITS of Makatchev et al. was intended to help students recognize and correct their misconceptions. In the implementation, student essays about physics problems were rendered as sets of propositions. "The student essay is viewed as a fragmentary, incomplete, and possibly incorrect proof. Our task is to complete that proof insofar as possible" (p. 193, Makatchev, Jordan, & VanLehn, 2004b). The proofs were coherent arguments that were consistent with students' mental models. In the implementation, it was the ITS (not the learner) who engaged in abductive reasoning, inferring the "best proof" in order to guide the tutoring process. The implementation approximates the Peircean scheme in certain respects. First, the implementation included plausibility analysis (by assigning weights for alternative explanations). (See row 3 in Table 2, below.) Second, although it might be asserted that abduction involves inferring explanations from observations, it might also be said that observations are made as a consequence of the explanations. (See row 5 in Table 2, below.)

Although the work of Makatchev et al. fits the Peircean model in these ways, it is not entirely clear what is "abductive" about the proofs. Granted that the research is aimed at a readership of computer scientists, it often seems unsatisfyingly distanced from the psychology of the matter. As was argued even in the 1990s, it is potentially misleading to label students' prior conceptions as misconceptions. "Students have often been viewed as holding flawed ideas that are strongly held, that interfere with learning, and that instruction must confront and replace. . . . This view overemphasizes the discontinuity between students and expert scientists" (Smith, diSessa, & Roschelle, 1993, p. 116). Indeed, one might check by asking whether the student actually believes the "misconceptions." The assumption of the Makatchev et al. proof method is that the misconceptions are student beliefs rather than bridges constructed by the tutor, that is, supplied by the model builder/programmer.

The term *abduction* is sometimes invoked in research on intelligent systems but actually does little heavy lifting, either in cognitive modeling or in implementations. For example, in the field of machine learning, Mooney (2000) defined abduction as the process of revising a knowledge base in order to fit the data. Abduction was defined formally as the inductive inference across multiple cases of an explanatory hypothesis (cause–effect relation), such that its conjunction with observations of a particular case is true, and its conjunction with background knowledge is also true. Mooney's linkage of abduction to induction accords with Peirce's assertion that abduction "partakes of the nature" of induction, when an inference is taken beyond the given observations (Peirce, 1878).

A distinction must be drawn between abduction (of a sort) that is the program's process of constructing model of student's reasoning and abduction by the student in the sense of producing the observed behavior. In 1970s and most of 1980s, the ITS community equated the program's model of the student's knowledge with the student's knowledge. Although that may be functionally useful for implementation and testing, it is on shaky grounds psychologically.

Overall, although we see elements of the concept of abduction in the ITS work—abduction as Peirce described it—there is still a long way to go. The fuzziness of the concept of abduction manifests itself in the ITS literature. Indeed, computer scientists have been at liberty to propose "new forms of abduction" (e.g., Stickel, 1991) that are not obviously tied to any Peircean notion. We see in the AI literatures some scattered elements of the full Peircean model. Indeed, one might say that abduction is a convergence of some of the fundamental problems of the field of AI and its application to human learning and reasoning.

Peirce the Psychologist

From 1867 to 1908, Peirce discussed abduction in different contexts and with different focus points. Peirce's discourses on abduction dealt not only with the logical view (representing abductive inferences formally) but also with a psychological view (Peirce, 1891b, 1903a). To Peirce, the match between a set of data and a preferred explanation is more plausible than the match to other explanations, and so we accept the preferred one as the likely explanation. More than this, abduction involves justification. A hypothetic inference is maintained until contradicted by experience or until experience suggests a better explanation (e.g., simpler, more general, more plausible). Although abduction can be described as if it were a single, punctuated act of reasoning, like the making of a logical inference from given premises (as

in Figure 2), Peirce the psychologist regarded abduction as an exploratory discovery process involving the observation of something that is surprising, qualitative reasoning (judgments of plausibility), and insight (Bellucci, 2015; Douven, 2011a). Peirce was quite explicit about this, using the classical form:

> The surprising fact, C, is observed;
>
> But if A were true, C would be a matter of course,
>
> Hence, there is reason to suspect that A is true. (Peirce, 1903b)

Thus, the model presented in Figure 2 is too simple to capture the process of abduction as Peirce described it. Table 2 summarizes what Peirce said about abduction in his various writings (see also Fann, 1970; Niiniluoto, 1999).

It is noteworthy that the requirements in Table 2 are consistent with the logic-inspired views of Harman (1965), Josephson and Josephson (1995), and Lipton (2004), although the requirements go well beyond their logical expressions. The requirements in Table 2 are also consistent the psychology-inspired view of Lombrozo (2012) based on studies of explanatory reasoning. This suggests a reconciliation of the logical view and the psychological view.

Reconciling the Logical and the Psychological

There is a tendency for writing on abduction either to define it philosophically in terms of other fuzzy concepts (e.g., critical thinking, explanation) or to define it logically as a form of inference based on a set of premises. However, there is an alternative, in which abduction is understood as exploratory modeling, a constructive reasoning activity that is highly dependent on knowledge. "The process of explaining recruits prior beliefs and a host of explanatory preferences, such as unification and simplicity, that jointly constrain subsequent processing" (Lombrozo, 2012, p. 260). The reliance on propositions that are external to the syllogism (no matter how many premises it begins with) is central in the Peircean definition of abduction.

Referring to row 1 in Table 2, the reasoner's motivation for creating a model is that the observed event or phenomenon is interesting or surprising. The observed event or phenomenon is at least nontrivial; that is, it is complex and involves interactions whose properties (structure and behaviors) are not sufficiently understood. This regards abduction as an accomplishment, an exploratory activity that begins with some information (e.g., data, assumptions, hypotheses, partial models) about some system of

TABLE 2. Peircean Psychological Model of Abduction

Process	Requirements
1. Observation of an event or phenomenon.	The observed event or phenomenon is interesting or surprising.
	The perception of the event or phenomenon (i.e., categorization) hinges on the reasoner's knowledge and concepts.
2. Generation of one or more possible explanations for some observed event or phenomenon.	The understanding of the event or phenomenon hinges on the reasoner's knowledge and concepts. Abduction partakes of the nature of both creativity and informed guessing ("guesses guided by reasons"; Peirce, 1878, p. 479).
3. Judging the plausibility of the candidate explanations.	The judgment can be but is not necessarily based on considerations of necessity and sufficiency.
	The judgment can be but is not necessarily based on the estimation of probabilities or likelihoods.
4. Resolving the explanation.	The plausibility judgment results in a determination that an explanation is viable.
5. Extending the explanation.	The determination is always tentative, that is, subject to disconfirmation by further inquiry, even though there is an assumption that further instances will conform to the explanation.

Note. Discussions in the literature often refer to simple examples of abduction, but it is clear especially in Peirce's writing that there is an assumption that the observed event or phenomenon is at least nontrivial, that is, it is a complex event or phenomenon (i.e., it is a system of interactions whose properties [structure and behaviors] are not sufficiently understood).

interest, assimilated into an initial yet incomplete understanding (diSessa, 2018), producing a more complete model of that system, one that is consistent with the given information.

The psychological perspective on Peirce's notion of an explanation for a surprising event conjures Wittgenstein's argument about the role of tacit knowledge in language understanding (Wittgenstein, 1953). A phenomenon or event could not be a surprise unless the observer already had in mind some expectation, understanding, or mental model of the phenomenon. It follows that a surprise is not just an act of recognition (of a disconnect between a tacit model and experienced phenomenon). Rather, there must be a perspective shift or some consideration of the notion that there might be more than one possible model. Reasoning about possible models involves a plausibility judgment.

Figure 3 presents a process model version of the Peircean process described in Table 2. In the spirit of Peirce's writings, this describes the process of "reasoning from surprise to inquiry," as he phrased it in a letter written in 1905 (Bellucci, 2015).

To Peirce, abduction involves active exploration, the empirical assessment of competing hypotheses (Capaldi & Proctor, 2008). Abduction is an activity that is extended in time, having its own structure and dynamics. It is not a punctuated act of reasoning like making a logical inference. But the classical forms (deduction and induction) are involved in Peircean abductive exploration. Referencing row 5 in Table 2, this is where abduction involves deduction and induction as integral to the process of empirical evaluation. In some of his discussions of abduction Peirce considers abduction as a hybrid, that is, abductive reasoning "partakes of the nature of induction"

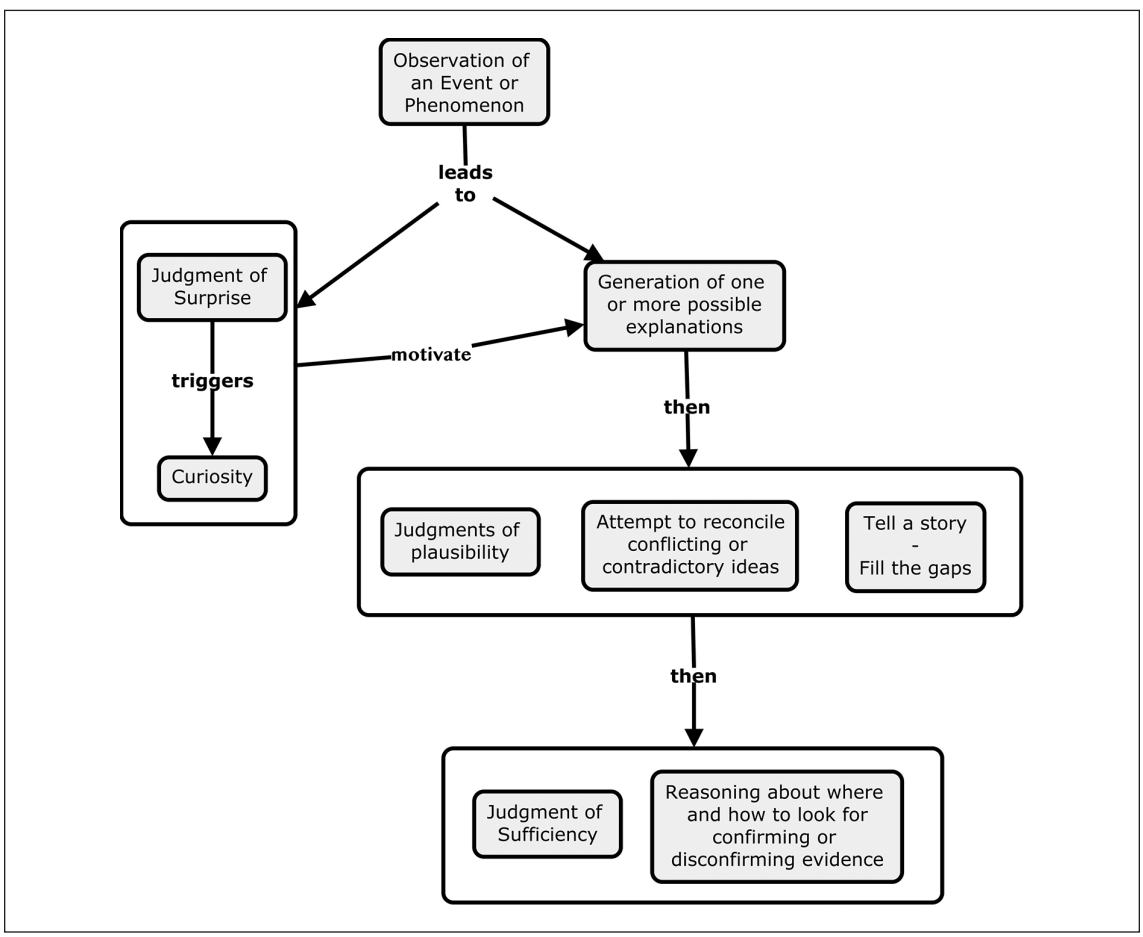

FIGURE 3. A process model representing Peircean abduction

(Peirce, 1878, 1903c). In other words, the classical forms of reasoning can be thought of as stages in abductive scientific research (Bellucci, 2015; Douven, 2011b).

The viability of the Peircean model presented in Figure 3 is reinforced by is concordance with recent research in two areas of applied cognitive psychology: research on the teaching of critical thinking skills and research on the reasoning of experts, spanning the poles of a proficiency continuum.

Evidence About the Peircean Model of Abduction: Research on the Training of Critical Thinking Skills

Peirce referred to the phenomenon in which an explanation pops out in the reasoner's awareness: "The suggestion comes to us like a flash" (Peirce, 1891a). Is that moment to be regarded as an act of inference? If so, it is certainly not of a classical form. To Pierce the abductive derivation of a rule is a creative act, an insight (Peirce, 1903b).

Some researchers have demonstrated success at teaching critical thinking skills defined so as to accord with the concept of abduction (e.g., Schank, 2011). van Dongen, Schraagen, Eikelboom, and te Brake (2003) conducted critical thinking training by using an ITS system that encouraged trainees to list multiple alternative hypotheses and then list both the confirming and disconfirming evidence for each hypothesis. The manifest purpose of the training tool was to mitigate confirmation bias, but the tutoring involved practice on deciding which of a set of alternative hypotheses was the better hypothesis.

Another study that comes close to implementing the Peircean concept of abduction is one by van den Bosch and de Beer (2007) on training for decision making. The researchers described two elements of critical thinking:

> *Building a story.* Explaining a situation, integrating assumptions and uncertainties. This corresponds to rows 1 and 2 in Table 2.
>
> *Testing and evaluating a story.* Identifying incomplete and contradictory information; evaluating the plausibility of the story. This corresponds to rows 3 and 4 in Table 2.

Tests and measures of critical thinking and explanatory reasoning have tapped selected aspects of abductive reasoning (e.g., recognition-primed decision making, plausibility judgment) but do not evaluate reasoning quality across all of the defining aspects of abductive reasoning, and apparently little work has been done on the skill or capacity for generating plausible hypotheses in the first place (Lombrozo, 2012).

Peirce's concept of abduction, and most of the secondary literature on it, has orbited the topic of scientific reasoning (e.g., Hanson, 1958; Peirce, 1878). But another new empirical source is the research on the psychology of expertise.

Abduction as Sensemaking by Domain Experts

Using methods of observation and cognitive task analysis, applied cognitive psychologists have studied the reasoning of experts in many diverse domains and professions (Ericsson, Hoffman, Kozbelt, & Williams, 2018; Hoffman, 2007; Klein, Orasanu, Calderwood, & Zsambok, 1993; Ward, Schraagen, Gore, & Roth, 2019). A leading model of expert reasoning accords well with Peirce's model of scientific abduction and also adds some specifications to it.

Research has shown that in most cases, the expert recognizes a situation as a familiar one and immediately engages in appropriate actions. This is called recognition-primed decision making (Klein, 1989). There is no deliberation over alternative hypotheses, and indeed, in many situations (e.g., firefighting, surgery) there is no time for deliberation.

The other scenario is one that is concordant with Peircean abduction. The data–frame model of sensemaking (Klein, Phillips, Rall, & Peluso, 2007) posits that there is some sort of trigger, when the expert observes something that is surprising, something that does not map well to a known pattern. The expert forms an initial mental model (i.e., inference to a plausible explanation) and then engages in an empirical exploration, in which the reasoner seeks new data, infers possible relationships, tracks anomalies, gauges data quality, and looks for evidence to support and refute the explanation. The data–frame process model maps clearly onto the description presented in Table 2 and Figure 3. This convergence of evidence and models bolsters one's confidence in proposing that a valid consensus model of abduction has been adduced.

Peirce's context, and that of the research on expertise, involves a focus on how people understand things they observe in the world, as it were (e.g., in

"pure" physics, the kinetics of gasses). But Peircean modeling can be used now to describe the process of creating and empirically evaluating explanatory hypotheses about how an AI system works.

A Conceptual Framework for Explainable AI

We present a conceptual framework specific to the methods of explainable AI based on the Peircean model of abduction. The key point is that people who are involved in an interaction with an XAI system are abducting to understand what, how, and why the AI does what it does, and not just make sense of the world that is being observed or controlled (e.g., forecasting the weather is guided by the outputs of computational models). The process of sensemaking, or self-explaining a complex system, is deliberative and effortful (Chi, Roy, & Hausmann, 2008; Chi, Siler, Jewong, Yamauchi, & Hausmann, 2001; Klein et al., 2019; Klein, Hoffman, et al., 2021; Renkl & Eitel, 2019). The process of explaining often takes the form of a dialog in which an explainer and a learner collaborate, explore, and co-adapt (Clancey, 1987; Walton, 2011).

First, we define the term *explainable AI*:

> Explainable AI is the development of AI systems capable of engaging in meaningful interactions with people to support their abductive reasoning.
>
> Explainable AI has the purpose of helping people develop good mental models of how the AI system works and when, why, and how it fails.
>
> Explainable AI requires that the user–AI relationship be one of interdependence: The user learns and benefits from the AI, but additionally, the XAI improves based on the actions and feedback of the user, such as improving its ability to adjust inputs or eliminate certain hypotheses.

It is important for those who are researching and building explainable systems to think beyond the initial spoonfeeding paradigm (see Figure 1), which is simply about automatically generating "one size fits all" statements. Statements such as "This is a bird because it has features a, b, and c of other birds" can to some extent be useful for determining when and how a model might fail or succeed and can be useful in building initial mental models. But for genuine understanding of AI systems, spoonfed explanations can leave a wide epistemic gap that users have to bridge. Indeed, spoonfeeding can actually impede sensemaking. Recent experiments indicate that if someone already agrees with an answer from an AI model, they sometimes ignore the explanations, and if they disagree, they do not always engage with the explanations (Gajos & Mamykina, 2022).

So if explainable AI is not about producing explanations, what should it be about? It should be about providing tools that support people to realize and specify their mental model of a system and explore the behavior of the AI, working collaboratively as well as individually. This includes understanding why the AI made particular decisions (so-called local explainability) and understanding more generally how the AI works (so-called global explainability). For example, in many current XAI systems for object recognition, local explanations take the form of "heat maps" that use rainbow colors to highlight the areas that were most heavily weighted by an algorithm. (Computer scientists refer to these as showing "saliency," or what the AI is "seeing" or "paying attention to.") In some XAI systems (e.g., for the control of autonomous systems), local explanations take the form of decision rules, which are often cryptic and complex. (For a review, see Mueller et al., 2019.)

The Peircean model of abduction (see Table 2 and Figure 3) entails a conceptual framework for explainability requirements for XAI tools, presented in Table 3. These requirements offer distinct challenges to computer scientists.

A glance at research in XAI will show that support for abduction is not the conceptual framework of most projects. Current XAI systems fall flat on rows 1 and 2 in Table 3. Some XAI systems preempt exploration by having some sort of window always open that shows some information about a particular decision or classification (e.g., similar instances, lists of features, a saliency map, a decision rule). If the AI does something that is surprising, this same basic display is all the user has to go on. Currently, much XAI research is focused on row 3, plausibility judgment. Such judgments express the justification that computer scientists present to other computer scientists, to explain "why we built it that way." This is for good reason: The complexity and opacity of AI models means that revealing reasons is a difficult technical challenge. However, this does not mean

TABLE 3. Supporting Abduction in AI

Process	Explainability requirements
1. Observation of an event or phenomenon.	Design clear interfaces to make it easy to identify what has happened.
	Design tools that make it easy to highlight events that could be considered anomalous or unusual for a particular domain.
2. Generation of one or more possible explanations for some observed event or phenomenon.	Design interfaces and affordances that either list potential (archived) hypotheses or help people compose a list of new potential hypotheses. Hypotheses in this case would be the "causes" or "reasons" that influence the AI's outputs.
3. Judging the plausibility of the candidate explanations.	Design interactions and affordances that make it easy for people to learn how the causes (inputs) affect system outputs. This would include support for contrastive analysis: How would a decision of the AI change if some variable had been different? Why would an AI decision *not* change if some variable were different?
4. Resolving the explanation.	Make it easy for people to explore how the AI operates when it is pushed to the boundaries of its competence envelope, to surmise the *when*, *how*, and *why*.
5. Extending the explanation.	Make it easy for people to revisit and revise any of their determinations if they receive new information or there are new surprises, or if they have new insights.
	Make it easy for people to explore follow-up questions that the explanatory or sensemaking process may have raised for them.
	Make it easy for people to access and share information about the AI.

that the other important steps in the model can be safely neglected. XAI requires good tools to support understanding rather than the automatic generation and presentation of "explanations," which does not go far enough in supporting abductive reasoning in many scenarios.

Prospects

We know that concepts of abduction can be implemented. Makatchev, Jordan, and VanLehn (2004a, 2004b) demonstrated that an ITS can generate plausible hypotheses about students' reasoning. This is suggestive of the possibility of supporting people's abductive reasoning about how an AI system works and how it fails. Mooney (2000) implemented abduction in the form of a process that would modify its knowledge base to make it consistent with evidence. This accords with the Peircean notion of plausibility judgment and the resolution of explanations (see Table 3, rows 3 and 4). The field of ITSs has demonstrated that it is possible for a computer system to engage in meaningful interaction with learners and facilitate their sensemaking. Clancey and Hoffman (2022) reviewed a number of ITSs, listing some specific capabilities that have been implemented and evaluated and that align with the requirements for AI systems that would support abductive reasoning (see Table 3, rows 3, 4, and 5). For example, an ITS can promote understanding by enabling the trainee to make rapid comparisons of cases (exploration). An ITS can help the trainee reflect on experience to integrate fragmentary general and situated knowledge.

Clearly, there is no single clear or easy path to the computational modeling of the full process of exploratory sensemaking as Peirce described it. It might be worthwhile to pursue this, to develop intelligent systems that support people to perform rigorous abductive reasoning, and to allow for the assessment of abductive reasoning as a learnable skill.

NOTES

The authors thank the anonymous reviewers for their helpful comments.

This material is approved for public release. Distribution is unlimited. This material is based on research sponsored by the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711 and by Australian Research Council Discovery Project grant DP190103414, "Explanation in Artificial Intelligence: A Human-Centered Approach." The U.S. government is authorized to reproduce and dis-

tribute reprints for governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. government.

Address correspondence about this article to Robert R. Hoffman, Institute for Human and Machine Cognition, Pensacola, FL 32592 (email: rhoffman@ihmc.us).

REFERENCES

Aliseda, A. (2007). Abductive reasoning: Challenges ahead. *Theoria, 60,* 261–270.

Aristotle. (1989). *Prior analytics* (R. Smith, Trans.). Hackett Publishing.

Beckwith, A. (2018). C. S. Peirce and abduction inference. *Johnson Community College Honors Journal, 10*(2). https://scholarspace.jccc.edu/honors_journal/vol10/iss1/2

Bellucci, F. (2015). Charles Sanders Peirce: Logic. In *Internet encyclopedia of philosophy*. https://iep.utm.edu/peir-log/

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*. http://home.earthlink.net/~dwaha/research/meetings/ijcai17xai/1.%20(Biran%20&%20Cotton%20XAI-17)%20Explanation%20and%20Justification%20in%20ML%20-%20A%20Survey.pdf

Bruner, J. (1985). On teaching thinking: An afterthought. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills, Vol. 2: Research and open questions* (pp. 597–608). Erlbaum.

Capaldi, E. J., & Proctor, R. W. (2008). Are theories to be evaluated in isolation or relative to alternatives? An abductive view. *American Journal of Psychology, 121,* 617–641.

Chi, M. T. H., Roy, M., & Hausmann, R. G. (2008). Observing tutorial dialogs collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32,* 301–341.

Chowdhury, R., & Lake, M. (2018). Is explainability enough? Why we need understandable AI. *Forbes.* https://www.forbes.com/sites/rummanchowdhury/2018/06/04/is-explainability-enough-why-we-need-understandable-ai/?sh=33ed372d62f4

Chi, M. T. H., Siler, S. A., Jewong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471–533.

Clancey, W. J. (1987). *Knowledge-based tutoring*. MIT Press.

Clancey, W. J, & Hoffman, R. R. (2022). Methods and standards for research on explainable artificial intelligence: Lessons from intelligent tutoring systems. *Applied AI Letters*. https://doi.org/10.1002/ail2.53

Collins, A. (1985). Teaching reasoning skills. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills, Vol. 2: Research and open questions* (pp. 579–608). Erlbaum.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10,* 105–225.

diSessa, A. A. (2018). A friendly introduction to "Knowledge in Pieces": Modeling types of knowledge and their roles in learning. In G. Kaiser et al. (Eds.), *Invited lectures from the 13th International Congress on Mathematical Education*. ICME-13 Monographs. https://doi.org/10.1007/978-3-319-72170-5_5

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*, 1702.08608v2.

Douven, I. (2011a). Abduction. In *Stanford encyclopedia of philosophy.* https://plato.stanford.edu/entries/abduction/

Douven, I. (2011b). Supplement to Abduction. In *Stanford encyclopedia of philosophy.* https://plato.stanford.edu/entries/abduction/

Douven, I. (2022). *The art of abduction.* MIT Press.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, M. (2018). *Cambridge handbook of expertise and expert performance* (2nd ed.). Cambridge University Press.

Facione, P. A., & Facione, N. C. (1992). *The California critical thinking skills test and manual.* California Academic Press. http://www.insightassessment.com

Fann, K. T. (1970). *Peirce's theory of abduction*. Nijhoff.

Finin, T., & Morris, G. (1988). *Abductive reasoning in multiple fault diagnosis*. Technical Report, Department of Computer and Information Science, University of Pennsylvania [https://repository.upenn.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1730&context=cis_reports

Fraser, J. M., Strohm, P., Smith, J. W. Jr., Svirbely, J. R., Rudmann, S., Miller, T. E., . . . Smith, P. J. (1989). Errors in abductive reasoning. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* (pp. 1136–1141). IEEE.

Gajos, K. Z., & Mamykina, L. (2022). Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (pp. 794–806). *arXiv,* 2202.05402.

Gerlings, J., Shollo, A., & Constantiou, I. (2021). Reviewing the need for explainable artificial intelligence (XAI). In *Proceedings of the 54th Hawaii International Conference on Systems Science* (pp. 1284–1293). https://hdl.handle.net/10125/70768

Glaser, R. (1984). Education and thinking. *American Psychologist, 39,* 93–104.

Goodman, B., & Flaxman, S. (2016). *European Union regulations on algorithmic decision making and a "right to explanation."* Presented at the ICML Workshop on Human Interpretability in Machine Learning, New York, NY.

Gunning, D., & Aha, D. W. (2019, Summer). DARPA's Explainable Artificial Intelligence (XAI) program. *AI Magazine,* 44–58.

Gunning, D. Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's Explainable AI program: A retrospective. *Applied AI Letters.* https://doi.org/10.1002/ail2.61

Haig, B. D. (2009). Inference to the best explanation: A neglected approach to theory appraisal in psychology. *American Journal of Psychology, 122,* 219–234.

Hanson, N. R. (1958). *Patterns of Discovery: An inquiry into the conceptual foundations of science*. Cambridge University Press.

Harman, G. F. (1965). Inference to the best explanation. *Philosophical Review, 74,* 88–95.

Hempel, K. (1965). *Aspects of scientific explanation*. Free Press.

Hoffman, R. R. (Ed.). (2007). *Expertise out of context: Proceedings of the Sixth International Conference on Naturalistic Decision Making.* Taylor and Francis.

Hoffman, R. R., Klein, G., & Miller, J. E. (2011). Naturalistic investigations and models of reasoning about complex indeterminate causation. *Information and Knowledge Systems Management, 10*, 397–425.

Josephson, J. R., & Josephson, S. G. (Eds.). (1995). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Klein, G. (1989). The recognition-primed decision model. In W. B. Rouse (Ed.), *Advances in man–machine systems research* (vol. 5, pp. 47–92). JAI Press.

Klein, G., Hoffman, R. R., & Mueller, S. T. (2019). *Naturalistic psychological model of explanatory reasoning: How people explain things to others and themselves*. Presentation at the International Conference on Naturalistic Decision Making, San Francisco, CA. https://www.researchgate.net/publication/335083297

Klein, G., Hoffman, R. R., Mueller, S. T., & Newsome, E. (2021). Modeling the process by which people try to explain complex things to other people. *Journal of Cognitive Engineering and Decision Making, 15,* 213–232.

Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C. E. (Eds.). (1993). *Decision making in action: Models and methods*. Ablex.

Klein, G. A., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In R. R. Hoffman (Ed.), *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making* (pp. 113–155). Erlbaum.

Krause, J., Perer, A., & Bertini, E. (2016). Using visual analytics to interpret predictive machine learning models. *arXiv,* 1606.05685.

Lenat, D. B. (2013). Ontology alignment of learner models and domain models. In R. A. Sottilare, A. Graesser, X. Hu, & H. Holden (Eds.), *Design recommendations for intelligent tutoring systems* (Vol. 1, pp. 49–56). U.S. Army Research Laboratory.

Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. In P*roceedings of the 11th International Conference on Ubiquitous Computing* (pp. 195–204). Association for Computing Machinery.

Lipton, P. (2004). *Inference to the best explanation*. Routledge.

Lipton, Z. C. (2016). *The mythos of model interpretability*. Presented at the 2016 ICML Workshop on Human Interpretability in Machine Learning. *arXiv,* 1606.03490v3

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford University Press.

Macauliffe, W. H. B. (2015). How did abduction get confused with inference to the best explanation?, *Transactions of the Charles S. Peirce Society, 51,* 300–319.

Makatchev, M., Jordan, P. W., & VanLehn, K. (2004a). Abductive proofs as models of students' reasoning about qualitative physics. In *Proceedings of the International Conference on Cognitive Modelling (ICCM 2004)* (pp. 166–171). International Conferences on Cognitive Modeling.

Makatchev, M., Jordan, P. W., & VanLehn, K. (2004b). Abductive theorem proving for analyzing student explanations to guide feedback in intelligent tutoring systems. *Journal of Automated Reasoning, 32,* 187–226.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267,* 1–38.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum. *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*. http://home.earthlink.net/~dwaha/research/meetings/ijcai17xai/6.%20(Miller+%20XAI-17%20extended)%20Explainable%20AI%20-%20Beware%20of%20Inmates%20Running%20the%20Asylum.pdf

Minnameier, G. (2010). the logicality of abduction, deduction, and induction. In N. Bergman, S. Paavola, A.-V. Pietarinen, & H. Rydenfelt (Eds.), *Ideas in action: Proceedings of the Applying Peirce Conference* (pp. 239–251). Nordic Pragmatism Network.

Mooney, R. J. (2000). Integrating abduction and induction in machine learning. In P. Flach & A. Kakas (Eds.), *Abduction and induction* (pp. 181–191). Kluwer Academic Publishers.

Moore, A., & Malinowski, P. (2009). Meditation, mindfulness and cognitive flexibility. *Consciousness and Cognition, 18,* 176–186.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human–AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. Report on award no. FA8650-17-2-7711, DARPA XAI Program. DTIC accession number AD1073994. arXiv: 1902.01876.

Nakatsu, R. T. (2004). Explanatory power of intelligent systems: A research framework. In *Proceedings of Decision Support in an Uncertain and Complex World: The IFIP TC8/WG8.3 International Conference*. https://pdfs.semanticscholar.org/8cff/ab3d8abb2f0c6014b1121c5ac77e1195d933.pdf

Niiniluoto, I. (1999). Defending abduction. *Philosophy of Science, 66,* S436–451.

Nummedal, S. G. (1987). Developing reasoning skills in college students. In D. E. Berger, K. Pezdek, & W. P. Banks (Eds.), *Applications of cognitive psychology: Problem solving, education, and computing* (pp. 87–97). Erlbaum.

Peirce, C. S. (1878). Illustrations of the logic of science. *Popular Science Monthly, 13,* 470–482.

Peirce, C. S. (1891a). *Collected papers of Charles S. Peirce (1931–1958)* (CP 5.542, 5.544-5, 5-157). C. Hartshorne, P. Weiss, & A. W. Burks (Eds.). Harvard University Press.

Peirce, C. S. (1891b). Review of William James's *Principles of Psychology*. *Nation, 53,* 32.

Peirce, C. S. (1903a). *Harvard lectures on pragmatism* (CP 5, 171–174). Robin Catalog. Harvard University.

Peirce, C. S. (1903b). [CP] *Collected Papers of Charles S. Peirce (1931–1958)*. (CP 5.189). C. Hartshorne, P. Weiss, & A. W. Burks (Eds.). Harvard University Press.

Peirce, C. S. (1903c). [EP] The essential Peirce: Selected philosophical writings (1992–1998). (CP 5.145). The Peirce Edition Project (Eds.). Indiana University Press.

Peirce, C. S. (1908). A neglected argument for the reality of God. *Hibbert Journal, 7,* 90–112. Reprinted (CP 6.452–485); (*Selected Writings,* 358–379); (*Essential Pierce* 2, 434–450).

Pfeiffer, K., Feinberg, G., & Gelber, S. (1987). Teaching productive problems solving attitudes. In D. E. Berger, K. Pezdek, & W. P. Banks (Eds.), *Applications of cognitive psychology: Problem solving, education, and computing* (pp. 99–107). Erlbaum.

Pople, H. E. (1973). On the mechanization of abductive logic. *International Joint Conference on Artificial Intelligence, 73,* 147–152.

Renkl, A., & Eitel, A. (2019). Self-explaining: Learning about principles and their application. In J. Dunlosky & K. Rawson (Eds.), T*he Cambridge handbook of cognition and education* (Cambridge Handbooks in Psychology, pp. 528–549). Cambridge University Press. doi:10.1017/9781108235631.022

Schank, R. C. (2011). *Teaching minds: How cognitive science can save our schools.* Teachers College Press.

Shank, G. (1998). The extraordinary powers of abductive reasoning. *Theory and Psychology, 8*(6), 841–860.

Selz, O. (1935). *Versuche zur Hebung des Intelligenzniveaus: Ein Beittrag zur Theoriet der Intelligenz und ihrer erziehlichen Beeinflussung* [Experiments for increased intelligence level: A contribution to the theory of intelligence and its educational impact]. *Zeitschrift für Psychologie, 134,* 236–301.

Smith, R. (Ed.). (1989). *Prior analytics*. Hackett Publishing.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences, 3,* 115–163.

Stickel, M. (1991). A prolog-like inference system for computing minimum-cost abductive explanations in natural language interpretation. *Annals of Mathematics and Artificial Intelligence, 4,* 89–105.

Thagard, P. (1978). Best explanation: Criteria for theory choice. *Journal of Philosophy 75*(2), 76–92.

van den Bosch, K., & de Beer, M. M. (2007). Playing a winning game: An implementation of critical thinking training. In R. R. Hoffman (Ed.), *Expertise out of context* (pp. 177–198). Erlbaum.

van Dongen, K., Schraagen, J. M., Eikelboom, A., & te Brake, G. (2003). Supporting decision making by a critical thinking tool. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 517–521). Human Factors and Ergonomics Society.

Walton, D. (2011). A dialogue system specification for explanation. *Synthese, 182*(3), 349–374.

Ward, P., Schraagen, J. M., Gore, J., & Roth, E. (Eds.). (2019). *The Oxford handbook of expertise*. Oxford University Press.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Blackwell.