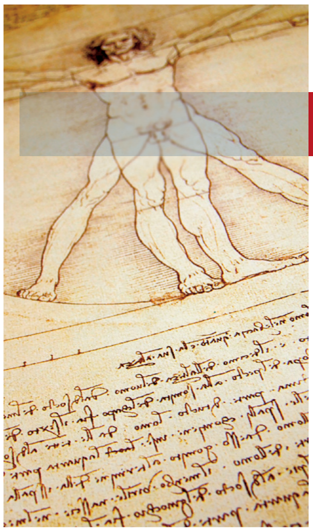

Editor: **Robert R. Hoffman,** Institute for Human and Machine Cognition, rhoffman@ihmc.us

Explaining Explanation, Part 1: Theoretical Foundations

Robert R. Hoffman, *Institute for Human and Machine Cognition*
Gary Klein, *Macrocognition, LLC*

This is the first in a series of essays that will explore the following questions: What is an explanation? How do people explain things? How might intelligent systems explain their workings? How might intelligent systems help humans be better understanders as well as better explainers? The first question speaks to theory. The second, to evidence, and the third and fourth to possibilities. The reason for these essays is the manifest programmatic interest in developing intelligent systems that help people make good decisions in messy, complex, and uncertain circumstances.

A thorough analysis of the subject of explanation would have to cover literatures spanning the entire history of Western philosophy from Aristotle onward. Research in cognitive psychology and scholarship in philosophy of science attest to the close relations of explanation and causal reasoning.[1,2] Explanations relate the event being explained to principles and invoke causal relations and mechanisms.[3] Research has likewise attested to the close relation of explanation and abductive inference.[4] Indeed, explanation had been defined as the finding of a best explanation, which is typically how Charles Sanders Pierce's concept of abduction[5,6] is understood. It is these elements of theory that we touch upon as key points in this first essay.

Causation Versus Causal Reasoning

Nearly every aspect of causation that has been proposed has come under scrutiny, including David Hume's notion that causes and effects are discrete events, and that causes must always precede effects.[7] One clear conclusion is that nearly any thing or event can count as a cause and nearly any thing or event can be thought of as the effect or the result of some other event.[8] Taxonomies of causes all assert that both causes and effects can be forces, events, beliefs, decisions, actions, and so on.[9,10] For example, an object can be a cause of something ("The mere sight of milk makes his skin itch"), or a property of an object can be a cause of something ("The red color of the sports car made the highway patrol officer skeptical"). The problem of creating causal explanations, and decomposing causal arguments into their propositional content, entails the problems of

- world knowledge,
- natural language understanding,
- developing an ontology for events, and
- developing an ontology for temporal relations.

One must be able to specify whether a causal force is continuous or discrete (the idea that a force might have a beginning and ending), distinguish forces from states, and so forth. Formalisms for causation typically just use the word "cause" as a predicate or operator in formulae asserting whether some proposition explains some other proposition, as in:

If α causes β, then $\alpha \rightarrow \beta$.

Such formulae in "causal calculus" do not contribute much that is of interest regarding the semantic and epistemic problems at hand. Causation becomes entailment or a form of conditional dependency. To date, the only solutions have been to restrict the range for the ontology (the topics to which a given decompositional analysis applies), for example, just reasoning about the causes of diseases. If the topic of interest is complex indeterminate causation in which human activity is central, then the problem again broadens to the problems of world knowledge, intention, and natural language understanding.

Our subject matter is causal reasoning, not causation—the way people formulate causal explanations and decide if an explanation is satisfactory, not the

1541-1672/17/$33.00 © 2017 IEEE
Published by the IEEE Computer Society

nature of causation itself. Furthermore, the vast bulk of the literature on the psychology of causal reasoning has involved analysis of reasoning about physical causation. (Will the water exiting a coiled hose continue to follow an arc?) We go well beyond that context. Causal reasoning in search of explanations is central to many of the high-level or macrocognitive functions that are crucial in modern sociotechnical work systems, and thus is central to planning and analyzing courses of action that actually have little to do with physical causation.

- Causal reasoning is central to *sensemaking*, the application of causal reasoning to understand events and to modify their causal models based on what is learned.
- Causal reasoning plays a central role in our *mental models* about how events transpire and what will happen if we intervene. Our mental models hinge upon knowledge and beliefs we summon to make sense of events. We might even define our mental models as the causal network that we understand to be operating to make things happen.
- Causal reasoning is central to *decision making*; the causal models people hold determine the way they recognize and categorize situations and the kinds of mental simulation they will perform to evaluate courses of action.
- Causal reasoning is central to *replanning*,[11] diagnosing why a plan might be going poorly and considering what needs to be altered.
- Causal reasoning is central to *coordination*, anticipating how individuals' actions will affect the team's activities.
- Causal reasoning is central to *anticipatory thinking*, using our mental models to prepare ourselves for possible events, particularly low-probability high-impact events.[12]

Much of the time, reasoning is the search for explanations of events that involve multiple parallel and interacting causes, and that are indeterminate because the events are of low intrinsic predictability due primarily to the vicissitudes of human activity and motivation.

Theoretical Foundations

We establish the theoretical foundations by referencing three descriptive models: (1) abduction, (2) retrospection and counterfactual reasoning, and (3) prospection, or projection to the future, which involves transfactual reasoning.

Abduction

Abductive reasoning has been given many diverse definitions in psychology, education, philosophy, and allied disciplines.[13] Looking across all these venues of scholarly and scientific activity, the concept becomes hard to nail down. Like critical thinking, cognitive flexibility, fluid intelligence, and creativity, abduction has become a nebulous concept that is sufficiently open to allow anyone to offer his or her own conceptual or working definition.

We prefer to fall back upon the original, seminal definitions. Actually the idea we now call abduction can be traced to Aristotle. In *Prior Analytics*, he discussed "reduction," or the explanation of two givens by a single conclusion. However, despite this historical precedent, most discussions attribute the concept of abduction to Peirce.

Peirce described logical inference in a way that differed from the logics of Aristotle and Hume. He rejected the idea that the roster of "acceptable types of inference" includes just the two classical types, deduction and induction, and his neoclassical type that he called abduction. In fact, he felt that one type covers the two classical forms:

RULE (All men are mortal)
Case 1 (Socrates is a man)
The Case falls under the rule (Socrates is mortal).

Induction is looking at n cases and having an expectation of what you would find for future cases. Thus, it can be said that induction is the same thing as "generalization," and it is always relative to some bounded set or limiting parameters. The difficulty here is that since any given case will have an unbounded number of hypothetical features about which one might form rules, one can have an unbounded number of rules. Although this might be a problem for logic, Peirce did not seem to see it as having any practical significance. In much of his discussion of abduction, he seems to consider it as a hybrid, that is, abductive reasoning also involves the use of both induction and deduction.[6] As Peirce put it, "abduction can partake of the nature of induction." Peirce also discussed how inductive inference can have the character of a perceptual judgment: "[O]ur first premises, the perceptual judgments, are to be regarded as an extreme case of abductive inference . . . the suggestion comes to us like a flash."[5] In other words, the very act of perceiving something, categorizing it, is a type of hypothesis formation, albeit an extreme case.

In writings spanning 1867–1902, Peirce defined abduction as a kind of informed guessing.[4,14] He also referred to abduction as "hypothetic inference"—the inferring of a hypothetical explanation from an observed surprising circumstance. A proposed explanation might be correct because if it were to be correct, the observed circumstance would necessarily occur. Abduction is distinct from both deduction and induction, which are both completely defined (in classical formal logics) by one or more assertions, one of which must be a generalization (that is, an assertion about a class). Abduction depends on propositions—from

Table 1. Peirce's decomposition of abductive inference.

Process	Requirements
1. Observation of an event or phenomenon.	Abduction has a "trigger": The observation of something that is interesting or surprising. The perception of the event or phenomenon (that is, categorization) hinges on the reasoner's knowledge and concepts. Discussions in the literature often refer to relatively 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.
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 (a sensemaking process). The derivation of an explanatory rule is a creative act.
3. Judging the plausibility of the candidate explanation(s).	Abduction is the search for a satisfying explanation. That judgment can be, but is not necessarily based on rationalist 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 typically (though not necessarily) results in a determination that a particular explanation is preferred. (This is Harman's[4] "inference to the best explanation," which supposes the rejection of all but one hypothesis.)
5. Extending the explanation.	Abduction involves going beyond the formation of a rule to the empirical testing of that rule. The determination of a preferred explanation is always tentative, that is, subject to disconfirmation by further evidence. Nevertheless, there is an accompanying expectation that further instances will conform to the preferred explanation.

the reasoner's knowledge—that are external to the calculus of the given assertions, that is, the observations and the explanatory hypothesis. Hence, abduction is not the same as inductive enumeration.[4]

Abduction can be understood as "rendering what might be thought of as a unique experience into an instance of a more general phenomenon."[15] John Josephson and Susan Josephson[16] (1995) described Peirce's meaning in a slightly more formal way[13,17,18]:

> D is a collection of data (facts, observations, and givens).
> H explains D (H would, if true, explain D).
> No other hypothesis can explain D as well as H does.
>
> ______________
>
> Therefore, H is probably true.

In other words, if the match between a set of data and a frame is more plausible than the match to any other frame, we accept the first frame as the likely explanation. An abductive inference (a hypothesis) is maintained until contradicted by experience or experience suggests a better (simpler, more general, and so on) hypothesis. Table 1 summarizes abduction as defined and discussed by Peirce.

In sum, abduction is the process of finding what is deemed to be the best explanation for some surprising complex event or phenomenon and then testing that explanation empirically. We forego a discussion of whether abduction is an ability, a skill, or a collection of component skills. Our main point is that discussions of abduction generally fail to retain one or another key aspect of Peirce's notion. especially surprise, affect, and empirical exploration (rows 1, 3, and 5 in Table 1). The role of affect in logic (and in discussions of Peircean logic) is often ignored. But looking across all discussions of abduction, we can generally conclude that abduction is rich with concepts and meanings, is highly dependent on knowledge, and moreover, is inference concerning complex events taken in context.

Abduction is both retrospective (explaining the past) and prospective (anticipating the future). It is possible to model these two forms.

Retrospection

The vast bulk of the literature on the psychology of explanatory reasoning (and the philosophy of causation) has involved analysis of reasoning about the causes of things that happened in the past, that is, retrospection. In reasoning about the past people often engage in counterfactual reasoning.

Causation and cause-effect relations have been be defined as forms of the counterfactual.[19] Although it is contestable whether a logic of counterfactuals is necessary for a logic of causation, it is certainly true that people reason about causation by using counterfactual reasoning (at least some of the time).[20] An analysis of causal reasoning must consider counterfactuals, if only because people often explain things by way of a counterfactual. With very few exceptions, all of the analyses of causal reasoning we have cited,[21] and many others not cited, are retrospective. That is, they deal with things that might or might not have happened in the past.

Planning and course of action analysis are necessarily and primarily *prospective*. In both the philosophical and psychological literatures, the process of explaining possible futures has been partitioned off as a separate topic, called "prediction." We do not think that that they are separable, as we will argue.

Prospection

Prospective causal reasoning involves things that might or might not be

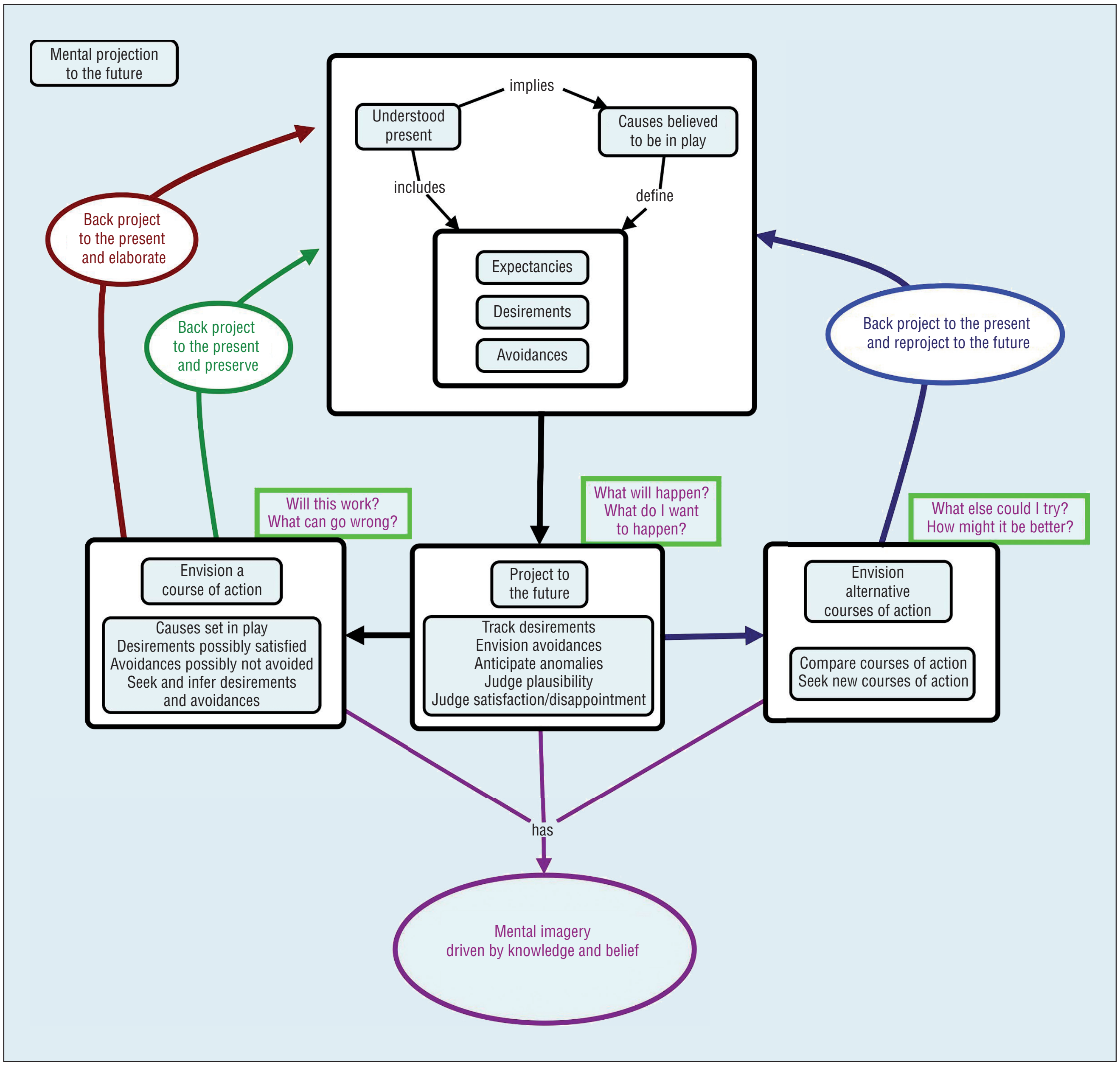

Figure 1. A macrocognitive model of "mental projection to the future." Notice that this model describes what happens when an individual envisions a course of action, considers the effects (desirable and undesirable) and then "back-projects" to the present to either modify the plan or consider alternative courses of action.[23]

happening now and that might or might not lead to something in the future. In classical terms, "What might happen in the future if *x* does not happen?" would be considered a counterfactual. But strictly speaking, this is not counterfactual since the facts have not happened yet, and hence are not really "facts." It is better to refer to this form of prospective reasoning as *transfactual*, as it transcends facts entirely and refers only to possible worlds. In addition, prospective reasoning includes abstractions in the form of durative events (a cause can continue on into an indeterminate future even after an effect has manifested). Because of the transfactual nature of prospective reasoning, there is a form of emergent: Something that was anticipated and did not happen. The nonoccurrence is a surprise. It would not have been predicted based on knowledge of the causes. "I did not expect X to occur and it did? So what happened?"

Prospective causal reasoning can involve asking additional questions.[22] One such question, of course, is "What will happen?" In this case, the distinction between explanation and prediction might seem fuzzy, but it is not. The purpose of the prospective reasoning might not just be to predict some event, but to anticipate things that would change the context, the ontology, or the variables that are currently considered when trying to predict and understand events. Thus, prospective causal reasoning also asks such questions as "What can I try? Will it work? What do I have to pay attention to?"[12] Prospection involves anticipating and preparing to respond to low-probability high-risk events, and not just predicting the most likely events.

Figure 1 presents a macrocognitive model of prospective causal reasoning, which shows how retrospection and prospection are linked. Beginning with the upper loop, the present is understood retrospectively in terms of the causes that are believed to have been in play. This understanding entails expectations for the future. We project past events, decisions, and perceived forces and abstractions forward in space and time to explain the present. Then we project from the present to the future, extrapolating forces, abstractions, and consequences from ongoing events and decisions. Referring to the outer closed loops, a course of action is imagined, as are its effects (desirable and undesirable) and these are then "back-projected" to the present to either modify the plan or consider alternative courses of action.[23]

Discussions of causal reasoning focus on the retrospective case, with rationality set in terms of the standard of logic or the axioms of probability. In real-world settings, the evidence for causation is typically too ambiguous to permit valid (that is, deductive) reasoning, so this is not a generally useful standard.

The Peircean model of abduction and the phenomenon of mental projection to the future must be foundational to a theory of causal reasoning and any naturalistic model of explanation. If abductive inference is regarded as a skill that is itself composed (somehow) of component abilities, what is required is a theory or model of how the components work together to result in abductive inferences. This should be treated for what it is—a question for empirical inquiry into how people actually reason. Empirical inquiry is the topic of the next essay in this series. We look to the actual occurrences of the phenomena of interest. As our research shows, there is great variety and diversity to causal reasoning, significantly broadening the scope and opportunity for study, modeling, and subsequent implementation. ■

References

1. S.M. Cohen, *The Philosophy of Aristotle*, 2008; http://faculty.washington.edu/smcohen/433/index.html.
2. J.D Trout, "Scientific Explanation and the Sense of Understanding," *Philosophy of Science*, vol. 69, no. 2, 2002, pp. 212–233.
3. T. Lombrozo, "Explanation and Abductive Inference," *Oxford Handbook of Thinking and Reasoning*, K.J. Holyoak and R.G. Morrison, eds., Oxford Univ. Press, 2012, pp. 260–276.
4. G.F. Harman, "Inference to the Best Explanation," *Philosophical Rev.*, vol. 74, no. 1, 1965, pp. 88–95.
5. C.S. Peirce, "Review of William James's Principles of Psychology," *Nation*, vol. 53, 1891, p. 32.
6. C.S. Peirce, *Harvard Lectures on Pragmatism*, chapter 5, 1903, pp. 171–174.
7. S. Mumford and R.L. Anjum, *Getting Causes from Powers*, 2011, Oxford Univ. Press.
8. R.R. Hoffman, G. Klein, and J.E. Miller, "Naturalistic Investigations and Models of Reasoning about Complex Indeterminate Causation," *Information and Knowledge Systems Management*, vol. 10, no. 4, 2011, pp. 397–425.
9. H.J. Einhorn and R.M. Hogarth, "Judging Probable Cause," *Psychological Bull.*, vol. 99, no. 1, 1986, pp. 3–19.
10. T.A. Groetzer, "Learning to Understand the Forms of Causality in Scientifically Accepted Explanations," *Studies in Science Education*, vol. 39, no. 1, 2003, p. 1074.
11. G. Klein, "Flexecution, Part 2: Understanding and Supporting Flexible Execution," *IEEE Intelligent Systems*, vol. 22, no. 6, 2007, pp. 108–112.
12. G. Klein, D. Snowden, and L.P. Chew, "Anticipatory Thinking," *Informed by Knowledge: Expert Performance in Complex Situations*, K.L. Mosier and U.M. Fischer, eds., Psychology Press, 2011, pp. 235–245.
13. A. Aliseda, *Abductive Reasoning: Logical Investigations into Discovery and Explanation*, Springer, 2006.
14. K.T. Fann, *Peirce's Theory of Abduction*, Nijhoff, 1970.
15. G. Shank, "The Extraordinary Ordinary Power of Abductive Reasoning," *Theory & Psychology*, vol. 8, no. 6, 1998, pp. 841–860.
16. J.R. Josephson and S.G. Josephson, eds., *Abductive Inference: Computation, Philosophy, Technology*, Cambridge Univ. Press, 1995.
17. B.D. Haig, "An Abductive Theory of Scientific Method," *Psychological Methods*, vol. 10, No. 4, 2005, pp. 371-388.
18. G. Minnameier, "The Logicality of Abduction, Deduction, and Induction," *Ideas in Action: Proc. Applying Peirce Conf.*, N. Bergman et al., eds., Nordic Pragmatism Network, 2010, pp. 239–251.
19. S.L. Morgan and C. Winship, *Counterfactuals and Causal Inference*, Cambridge Univ. Press, 2007.
20. R. Cowley, ed., *What If?: The World's Foremost Military Historians Imagine What Might Have Been*, Putnam Publishing Group, 2000.
21. S. Sloman, *Causal Models: How People Think about the World and Its Alternatives*, Oxford Univ. Press, 2005.
22. G. Klein and R.R. Hoffman, "Causal Reasoning: Initial Report of a

Naturalistic Study of Causal Reasoning," presentation at the 9th Int'l Conf. Naturalistic Decision Making, 2009.

23. G. Klein and B. Crandall, "The Role of Mental Simulation in Problem Solving and Decision Making," *Local Applications of the Ecological Approach to Human-Machine Systems*, P. Hancock et al., eds., 1995, pp. 324–358.

Robert R. Hoffman is a senior research scientist at the Institute for Human and Machine Cognition. His research interests include macrocognition and complex cognitive systems. Hoffman has PhD in experimental psychology from the University of Cincinnati. He is a Fellow of the Association for Psychological Science and the Human Factors and Ergonomics Society. He is a senior member of IEEE. Contact him at rhoffman@ihmc.us.

Gary Klein is chief scientist at Macrocognition, LLC. His research interests include naturalistic decision making. He is a Fellow of the American Psychological Association and of the Human Factors and Ergonomics Society. Contact him at gary@macrocognition.com.