Advances in Cognitive Systems 9 (2020) 3–12 Submitted 12/2020; published 12/2020

Explanation in Cognitive Systems

Pat Langley PATRICK.W.LANGLEY@GMAIL.COM
Center for Design Research, Stanford University, Stanford, CA 94305 USA
Institute for the Study of Learning and Expertise, Palo Alto, California 94306 USA

Abstract

In this essay, I discuss the importance of explanation to the computational study of cognitive systems. I distinguish between two common senses of the term – accounts of situations or events and the processes that produce them – as well as two distinct activities – constructing such accounts and communicating them to others. After this, I examine some representational issues that arise in both contexts, including the central role of knowledge and the varieties of explanatory structures. Next I consider in more detail the mechanisms that underlie the generation of explanations and the component abilities that support their communication. Finally, I conclude by noting some research challenges that cognitive systems researchers should pursue to further the field's understanding of these important mental faculties.

1. Introduction

One distinctive characteristic of human cognition is the ability to explain situations, events, and behaviors. People share many mental assets with other animals, including the capacity to categorize objects, carry out complex procedures, coordinate actions with others, and acquire new skills. However, dogs and cats, despite their excellent sensorimotor facilities, cannot diagnose a mechanical failure, understand the motives of a double agent, or clarify the reasons for their own activities. The production of rich explanations appears to be uniquely human, which means that it should be a central concern of the cognitive systems community. This topic has become even more important due to the popularity of statistical methods for learning effective but opaque expertise. Such systems often fare well on standardized tests for object recognition, caption generation, and robotic control, but they can explain neither the events they interpret nor the reasoning behind their conclusions. Techniques like the display of 'heat maps' offer some insight into decision making, but they bear little resemblance to people's accounts, which is the natural aim for cognitive systems research.

Two aspects of human explanations make them especially appropriate for study by our community. First, they invariably involve some form of *cognitive structure* that relates items of interest. For instance, a diagnosis links observed symptoms to hypothesized problems, often through multiple steps. Second, these structures typically comprise elements of *knowledge* that have been instantiated for the task at hand. Thus, the steps in a diagnosis might be instances of generic rules that relate symptoms to causes. Explanatory structures vary along a number of dimensions. They may be entirely qualitative, as in a geometry proof, or they may include quantitative annotation, as in the solution to a physics word problem. Accounts also differ in their complexity (e.g., the number of knowledge elements) and their depth (e.g., the length of reasoning chains). Nevertheless, they share many features that one can discuss in general terms.

We should distinguish between two uses of 'explanation' that commonly appear in English. The word sometimes refers to a mental, written, or spoken *structure* that serves to elucidate some phenomena or behaviors. Thus, we refer to a scientific explanation of pulsar cycles, an informal explanation of how a toilet flushes, or an explanation for one's home-buying decision. In other cases, the term denotes the *process* or *activity* of generating such an explanatory structure. We say that an astrophysicist engages in explanation of pulsar behavior, a plumber focuses on explanation of a leak, or a home buyer carries out explanation of his residential choice. This paper will use both senses of the term, but their meaning should be clear from the context.

We can further differentiate between two additional connotations of 'explanation'. The first meaning refers to the *construction* of accounts for observed situations or events. A geologist posits a set of processes for the origin of a landform, a mechanic hypothesizes the reasons a car does not start, and a reader infers the hidden goals of a novel's character. The result is a cognitive structure in the explainer's own mind. The second meaning instead deals with the *communication* of such mental structures once they exist. The geologist presents a talk about his account of a landform's evolution, the automobile mechanic includes diagnostic notes in an estimate, and the reader tells a friend his guesses about the character's motivations. However, this latter sense applies not only to sharing accounts of external events, but also to communicating why one made a given decision or generated a particular plan. Thus, it includes *self explanation* as an important special case.

Later in the essay, I elaborate on these two varieties of explanatory processes, as they emphasize different cognitive mechanisms that deserve separate treatment. Nevertheless, they operate over the same types of mental structures, so I first review some representational issues that arise in both settings. In the closing section, I discuss some challenges that remain for the study of explanation, broadly defined, that the cognitive systems community is uniquely qualified to address.

2. Representional Aspects of Explanations

We have seen that explanations are cognitive structures an intelligent system can construct or communicate, so both their form and content merit discussion. Such accounts relate a set of observed facts, along with optional queries or goals, to each other, so these elements serve as key building blocks. Explanations invariably draw on background knowledge, typically at the domain level (e.g., refrigerator operation, driving laws) but sometimes at the meta level (e.g., dialogue conventions). However, they incorporate not generalized knowledge elements themselves, but rather *instances* of such knowledge elements that connect facts or queries to each other.

In rule-based frameworks, explanations are organized as one or more proof trees with shared subproofs, where each rule instance links observed or inferred beliefs. For instance, an account for why an automobile does not start might connect observed behaviors through instantiated rules that describe a generic car's operation. In script and frame paradigms, the knowledge elements are large enough that some accounts involve a single instantiated structure, although they often combine more than one. An explanation can also take the form of an analogy, where knowledge corresponds to stored cases (linked facts), one of which maps onto elements of the new situation. Any formalism (e.g., rules, scripts, frames, or cases) that encodes knowledge structures can serve in this capacity.

In addition, explanations can differ in the ontological character of the knowledge elements on which they draw. These may denote logical relations, like those in geometry proofs, but they may also incorporate numeric calculations, as arise in solutions to textbook physics problems. Moreover, the knowledge elements can include likelihood information, as in the rules of a probabilistic context-free grammar. In such frameworks, explanations can have the same organization as in logical ones (e.g., proof trees), but they attach probabilities to constituents. Knowledge structures may also have a causal interpretation, which can in turn be either deterministic (e.g., a broken wire leads a starter to fail) or stochastic (e.g., a loose wire sometimes causes failure). Accounts that focus on an agent's behavior may be teleological in that they refer to the goals that guide its decisions and actions. Other explanations instead involve predictable patterns that lack further justification; many social norms and conventions (e.g., expected behavior in churches or restaurants) take this form.

Finally, observed facts can play two distinct roles in explanatory structures, as Langley and Meadows (2019) have noted. In *derivational* explanations, observations serve as root nodes in a set of connected proof trees, while rule instances or other instantiated knowledge structures show how they follow from other facts and assumptions. Many scientific explanations adopt this scheme, as do causal diagnoses and telelogical plans. In *associative* explanations, observed beliefs appear only as terminal nodes, which let one deduce new beliefs that follow from these facts. Such accounts use instantiated knowledge structures to connect obervations to each other, but not to derive them. Parse trees for sentences are classic instances of this paradigm, but script-based interpretations of stories also illustrate the idea. One can typically convert one form of explanation into the alternative, but they suggest different approaches to processing, to which I now turn.

3. Processes for Interpretive Explanation

As discussed earlier, a common sense of 'explanation' denotes the activity of understanding things one has observed or been told. This pursuit may focus on generalized accounts, as when a scientist seeks a deeper explication of empirical laws, but more typically it addresses particular events or situations. We can specify this task of *interpretive explanation* in terms of inputs and outputs:

- *Given:* A set of generic knowledge elements;
- *Given:* A set of observed situations, events, or activities;
- *Find:* Explanations that relate the observations through knowledge.

Again, explanations are always stated in terms of something already known, so existing expertise plays a key role in their construction, with instantiated knowledge elements providing the connective tissue that holds together the given observations or facts.

Explanations of external events can take any of the forms described earlier. For instance, suppose we know that a grandfather is the father of someone's father and that a father is someone's male parent. Also assume that we believe Abe is a parent of Bob, that Bob is a parent of Cal, and that both Abe and Bob are male. Finally, suppose that we are told Abe is a grandfather of Cal and we want to explain this relationship. In response, we can construct a proof that shows Abe is a father of Bob, that Bob is a father of Cal, and therefore that Abe is a grandfather of Cal. This logical structure connects the specified facts in terms of the available knowledge. In this sense, all deductive theorem provers support a form of interpretive explanation, and this ability carries over to reasoning techniques that engage in probabilistic inference.

Another important class of external explanation involves causal reasoning. For instance, suppose we know that flu can lead to fever and sneezing and that allergies can produce sneezing and itchy eyes. If we observe a patient with fever, sneezes, and itchy eyes, then we might hypothesize that he has both the flu and allergies. Or suppose we know that decreases in oil production produce higher gas prices, greater gas prices reduce traffic, more traffic causes more pollution, and higher pollution increases lung disease. If we observe that growth in oil production has been associated with more lung problems, then we can explain this through the chain from oil to gas price to traffic to pollution to disease. This example involves qualitative causal reasoning, as regularly practiced in economics, but quantitative versions that link variables with equations are also possible.

A third category of external explanations relies on plausible reasoning that is not deductively valid. Suppose that Ron, a student who always scores highest in a class, is murdered just before the final exam. Another student, Tim, wants a scholarship that requires very high grades, but Ron often ruins the curve. One explanation of the death is that Tim killed Ron to eliminate the competition and thus improve his score. This is an example of *abduction* (Peirce, 1878), which relies on introducing assumptions to explain observations. We seldom encounter murder scenes, but abduction abounds in everyday life. Mechanics diagnose plausible causes of automobile problems and doctors infer likely diseases from patients' symptoms. We posit others' beliefs and intentions based on their behavior, including utterances during dialogues. We make informed guesses about unobserved events when watching movies and reading stories. Even interpreting the meanings of individual sentences has a similar character. Such problems are so common that one might argue abduction is the default in human reasoning and deduction is the special case.

Computational mechanisms for explanation, especially ones that support abduction, have many applications. These include medical and mechanical diagnosis, natural language comprehension, and scientific reasoning and discovery. A popular arena is *plan understanding* (e.g., Geib, 2016; Meadows et al., 2014), which involves inferring another agent's goals and plans based on its actions or the states it traverses. Research on this problem often assumes knowledge takes the form of a hierarchical task network, with explanations encoded as hierarchical plans and associated tasks. A related task is *story understanding* (e.g., Schank & Abelson, 1977; Winston, 2012), which often requires interpreting interactions among agents. Here explanations incorporate not only inferences about participants' mental states, but also their models of others' mental states. Interpreting simple fables requires such reasoning (e.g., Meadows et al., 2014), as do the more complex tasks of watching games, appreciating movies, executing military operations, and understanding dialogues. The ability to explain others' behaviors also plays a central role in legal reasoning and moral judgement.

One approach to explanation, typically associated with derivational accounts, relies on query-driven, backward chaining (e.g., Gordon, 2018; Molineaux et al., 2012; Ng & Mooney, 1990). This scheme starts from an observation O that one wants to explain and retrieves rules with consequents that unify with O. One then chains through the rules' antecedents to create subqueries and applies the process recursively. This continues until reaching consequents that unify with other observations or, in abductive variants, with assumed beliefs. The result is a proof tree that derives O from other observations or default assumptions. One repeats this process for each observation, reusing subproofs where possible to generate a proof graph with observed beliefs as root nodes. A less common approach to creating explanations, usually linked to associative accounts, uses data-driven forward chaining (e.g., de Kleer, 1986; Langley & Meadows, 2019). This starts from observed facts and accesses rules with antecedents that unify with them, with only some antecedents matching in abductive variants. The technique then chains through instantiated consequents to infer beliefs and applies the procedure recursively to these derived elements. The process continues until it has constructed a set of linked proof trees in which all or most observations appear as terminal nodes. A third option treats explanation as a satisfiability problem that involves search through a space of possible worlds to find interpretations that are consistent with known constraints (e.g., Baral, 2003).

In each framework, the search process may produce multiple accounts of the same observations. Abductive methods that introduce plausible assumptions can lead to even more candidates. Naturally, we require some way to rank alternatives by their quality, and researchers have explored a number of such metrics. These include favoring explanations with fewer inference steps, fewer assumptions (Hobbs et al., 1993), higher probabilities (Charniak & Shimony, 1990), and greater coherence (Ng & Mooney, 1990). Each approach offers a means to evaluate competing accounts. For example, the principle of *parsimony* suggests that simpler explanations are more desirable, but Ng and Mooney (1990) offer a counterargument. Suppose our knowledge includes three rules: an upbeat person is happy, anyone who does well on a task is happy, and someone who takes an exam and studied does well on it. Suppose further that we hear John is happy and that John took an exam. Assuming that John is upbeat provides a simple explanation of his mood, but many would prefer an account that assumes John studied, which is more coherent because it connects all observed facts.

Generating all consistent explanations is often tractable in deductive settings, but it becomes unrealistic for abductive cases that allow assumptions. In such situations, a cognitive system must rely on heuristic search through the space of alternatives, using criteria like those above to evaluate partial accounts. Some researchers (e.g., Eckroth & Josephson, 2014) have defined abduction as *inference to the best explanation*, but this view comes from a misguided emphasis on finding optimal solutions. Clearly, people can formulate incomplete or incorrect explanations, yet it seems unfair to claim their behavior does not count as abductive inference. Heuristic search is not guaranteed to find the best solutions, but it offers a practical alternative that lets cognitive systems tackle complex tasks (Langley, 2017). A complicating factor is the need for incremental processing, as new observations can arrive after one has a tentative account in mind and can lead to inconsistencies, which in turn require revision of the candidate explanations.

4. Processes for Communicative Explanation

A different but equally important sense of the word ‘explanation’ refers to the activity of conveying an existing account to another party. As before, we can also specify this problem of *communicative explanation* in terms of inputs and outputs:

- *Given:* One or more explanations of situations or events;
- *Given:* Records of decisions made during their generation;
- *Given:* Questions about the explanations or decisions (optional);
- *Find:* Answers that clarify the explanations and decisions.

Again, the purpose here is not to generate explanatory structures but rather to share their contents, and the reasoning behind it, with others. This information may be conveyed in natural language, an annotated diagram, a formal notation, or any other way that encodes the relevant material.

The most obvious type of explanatory communication reports an account of external situations or events, such as those described in the previous section. However, the process can also provide information about plans, designs, or other mental structures that an agent has devised to achieve its own objectives. These are not interpretive explanations in the sense discussed earlier, but they take the same form as accounts of external events, and the same mechanisms can communicate their content. For instance, a plan that an agent generates to achieve its own goals may have the same content as one that it infers another agent has pursued. The plan-generating agent knows goals in advance and must determine actions, whereas the plan-understanding agent observes actions and must infer goals, but the resulting structures are equivalent and both are important. Langley (2019) has defined *explainable agency* as the ability to report on one's own plans and activities[1] and *justified agency* as the related ability to explain them in terms of social norms.

We can further distinguish types of communication that emphasize different facets of explanation, whether of external events or internal processing. These varieties include:

- *Structural accounts*, which clarify how a given collection of steps is *rational* in Newell's (1982) sense that an agent believes they could help achieve its goals. Thus, an explanation of external events includes inference steps that connect a query or conjecture to given facts or assumptions. Similarly, an agent's plan incorporates a sequence of actions that, if carried out, should produce an end state that satisfies some goal description while not violating any known constraints. The explanatory structure shows how the steps link the goals or query to the initial situation through knowledge: it focuses on the *means* of achieving objectives.
- *Preference accounts*, which explicate how some collections of actions or inferences are more desirable than others. Thus, an agent might note that it views one route plan as better than an alternative because it is shorter or traverses fewer traffic lights. An agent might prefer one account of external events because it has the fewest inference steps or makes the fewest assumptions. When multiple criteria come into play, preference explanations clarify their relative importance and how decisions resolve tradeoffs. Such accounts need not be quantitative; they may rely on symbolic rules that rank candidates without assigning scores. Rather than addressing the means of reaching objectives, preference explanations focus on solution *quality*.
- *Process accounts*, which describe the cognitive steps that led an agent to generate its plans or other mental structures. These revolve around the classic idea that complex reasoning requires heuristic search through a problem space. Here an agent might recount the history of its search, including the choices considered at each branch point, why it selected a given option, and when backtracking occurred, much like the accounts in verbal protocols but after the task has been completed. Such mental replays share features with preference accounts, but they focus on the detailed steps of search rather than on global evaluation of completed alternatives.

These are some obvious forms of communicative explanation, but others are possible. One variety clarifies how the agent revised a plan during execution because unexpected events occurred or how it updated an account of observations in response to new data. Another type conveys why the agent viewed a plan or account as unacceptable or nonviable. However, none require search themselves, as the reasoning chains, scores, and traces already exist as mental structures.

[1] Typical examples address why an agent took actions in the world, but it also applies to mental processing; thus, a cognitive system may first interpret a set of observed facts and later elucidate its inference steps to another agent.

The challenges here are very different from those that arise during interpretive explanation. Again, these may involve accounts of external events or they may involve plans, proofs, or designs related to the agent's own goals. In both cases, the cognitive structures link a set of concrete elements through instantiated versions of generic knowledge. The most straightforward approach to communicative explanation, at least for structural and preference accounts, simply reviews their component steps and notes why one alternative is better than others. However, more commonly the agent must respond to directed questions about specific aspects of one or more explanatory structures. Moreover, the agent may have created such accounts for many distinct situations in the past and it may be queried about any of them.

Responding to such questions appears to require three component processes. First, the agent must store the explanatory structure in memory, so it retains the content it needs to communicate in the future. In the simplest scheme, it would simply record the account itself, but a more complete approach would also cache information about the choices it considered during search, which ones it selected for further attention, and why it favored one option over others. For instance, a diagnostic agent might retain not only which fault hypotheses it entertained, but why it decided one was more plausible. Similarly, a planning agent might store not just the final route selected, but which turns it considered and where it backtracked during search. In addition to storing these details, the agent must index them in a manner that makes them accessible later. This means linking elements of the explanation to informative cues that distinguish them from others. Indexing is even more important for agents that are long lived and thus must store many explanations in an episodic memory.

Second, the agent must interpret the incoming question and use it to retrieve the relevant portion of a stored explanation from memory. This involves translating the query into a retrieval cue, mapping it to an appropriate index, and returning the indexed structure. Different types of indices will be useful for different aspects of explanatory structures. For example, when we ask a plan-understanding agent why someone took an observed action, it might retrieve parts of a stored account that refer to this action and how it enabled an inferred goal. In the same way, when we query a plan-execution agent about why it abandoned an intended route, it might access where it hit unexpected construction that caused the decision. The retrieval process should return only those aspects of an explanation that are relevant to the question at hand. When an agent has stored many separate accounts in memory, it should also identify which one to access for an answer.

This reliance on retrieval raises the question of what counts as a legitimate communicative explanation of one's decision making. People are reasonably good at generating verbal protocols during problem solving, but they are notoriously unreliable at reproducing their reasoning later and instead provide rationalizations. Such reconstructions are similar to accounts of external events, in that they explain partial memories in terms of plausible inferences over background knowledge. This form of communicative explanation is relevant to modeling humans, but it is less defensible when studying artificial systems, which need not suffer from memory limitations. Thus, I will limit my comments here to communication that operates over accurate traces of mental processes.

Finally, once the agent has retrieved the relevant portion of an explanation from episodic memory, it must identify the aspects appropriate for the query, translate them into an understandable form (e.g., natural language or a formal notation), and share the answer with the questioner. This communication should include no more detail than necessary to clarify the content, so the agent must also decide on the appropriate level of abstraction. Naturally, different types of query will invoke different types of answers. For instance, responses to questions about explanatory structure will present reasoning chains, whereas ones about choices the agent considered will list alternatives and selection criteria. Similarly, communications about explanations of external events may emphasize different elements (e.g., observations) than ones about an agent-generated plan (e.g., goals).

There has been some research on the component processes for storing, retrieving, and communicating an agent's reasoning. Studies of analogical planning (e.g., Jones & Langley, 2005; Veloso et al., 1995) have addressed the problems of storage and retrieval, but not in support of communicative explanation. Early expert systems could recount their reasoning chains when asked to defend their conclusions (e.g., Swartout, Paris, & Moore, 1991), but these focused on diagnosis. Johnson (1994) and Van Lent et al. (2004) extended the ideas to agents that recorded their decisions during execution of military missions and later answered questions about their reasons, including counterfactual ones. More recently, Briggs and Scheutz (2015) reported a robotic agent that answers five types of questions about why it could not carry out a task. However, we need far more research on this important topic, as discussed in the next section.

5. Open Challenges in Explanation

As already documented, there has already been substantial progress on the representations and mechanisms that underlie both interpretative and communicative explanation. Nevertheless, some key issues have not received the attention they deserve, and the cognitive systems community should delve deeper into this important collection of mental abilities. In this section, I consider some research challenges that have potential to drive further efforts in this area.

The process of interpretive explanation draws upon knowledge elements as building blocks to account for observed facts, but the knowledge base may be incomplete. When this occurs, humans can often bridge the gap by inferring rules or schemas to complete an account, much as classic abduction posits plausible beliefs. One approach to replicating this ability uses domain-independent rules that include no domain predicates but that can match or introduce them during the search for explanations. Researchers have applied this idea in physics problem solving (VanLehn & Jones, 1993) and high-level robotic control (Cropper & Muggleton, 2015) to extend incomplete knowledges bases with new domain-specific rules from a few training cases. There has also been some work on learning constraints (e.g., Law et al., 2018), but it has focused on techniques that require sizable labeled training sets. In contrast, we need research on methods that acquire such knowledge in a rapid, unsupervised manner that is integrated with the explanation process itself.

Once a cognitive system has constructed an explanation, it should store some generalized version of the structure in memory for future use. This can serve the same purpose as a 'lemma' in mathematics by reducing the number of steps needed to prove a higher-level conjecture. This idea was central to early work on 'explanation-based learning' (e.g., Mitchell et al., 1986) and macro-operation formation (e.g., Iba, 1987), but these paradigms relied mainly on deductive reasoning. We should extend this approach to aid the broader range of explanation methods discussed here. For example, abductive inference could benefit from high-level structures that encode long-distance relations which would otherwise require search. Extracting such templates from an explanation's constituents is a natural source for the *chunks* implicated in human cognition (Miller, 1956).

We also need more research on communicative explanation. One promising extension would support agent-recipient interactions that involve extended dialogues rather than sets of isolated questions. Here the agent would interpret later queries in the light of earlier utterances and even ask for clarification itself when ambiguity arises. In addition, we should augment explanation systems to take into account how they differ from the recipient. A simple scenario involves adopting a listener's perspective when describing spatial relations, but more complex situations require an agent to adapt its presentation based on the other's knowledge. An agent should convey information that it believes the questioner lacks but omit content that it thinks he already knows. This will sometimes require repeatedly updating a model of the questioner's knowledge and beliefs based on his queries.

Finally, we should develop cognitive systems that combine these distinct abilities. Consider an agent that accepts communicative explanations from a human but, when such reports are incomplete, invokes interpretive explanation to fill in the details. The system could also generate questions to discriminate among competing accounts, present them to the human, and incorporate the answers into its developing model. Once it constructs an account in sufficient detail, the agent could then use communicative explanation to pass on its understanding to a third party. A key difference here is that information would be collected not from passive observation but rather from active interrogation. Such an integrated agent would be fully in the spirit of the cognitive systems movement. Research along these lines would provide deeper insights – via interpretative explanation – and let us share those insights – via communicative explanation – about the nature of explanatory processes.

Acknowledgements

This essay was supported by Grants N00014-17-1-2434 and N00014-20-1-2643 from the Office of Naval Research and by Grant FA9550-20-1-0130 from the Air Force Office of Scientific Research, which are not responsible for its contents. I thank Will Bridewell, Ben Meadows, David Nicholas, and Mohan Sridharan for insightful discussions on the nature of explanation.

References

Baral, C. (2003). *Knowledge representation, reasoning and declarative problem solving*. Cambridge, UK: Cambridge University Press.

Briggs, G., & Scheutz, M. (2015). "Sorry, I can't do that:" Developing mechanisms to appropriately reject directives in human-robot interactions. *Proceedings of the AAAI Fall Symposium on AI and HRI*. Arlington, VA: AAAI Press.

Charniak, E., & Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 446–451). Cambridge, MA: AAAI Press.

Cropper, A., & Muggleton, S. H. (2015). Learning efficient logical robot strategies involving composable objects. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence* (pp. 3423–3429). Buenos Aires, Argentina: AAAI Press.

de Kleer, J. (1986). An assumption-based TMS. *Artificial Intelligence*, *28*, 127–162.

Eckroth, J., & Josephson, J. R. (2014). Anomaly-driven belief revision and noise detection by abductive metareasoning. *Advances in Cognitive Systems*, *3*, 123–142.

Geib, C. (2016). Lexicalized reasoning about actions. *Advances in Cognitive Systems*, *4*, 187–206.

Gordon, A. (2018). Interpretation of the Heider-Simmel film using incremental Etcetera Abduction. *Advances in Cognitive Systems*, *7*, 23–38.

Hobbs, J. R., Stickel, M. E., Appelt, D. E., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*, 69–142.

Johnson, W. L. (1994). Agents that learn to explain themselves. *Proceedings of the Twelfth National Conference on Artificial Intelligence* (pp. 1257–1263). Seattle, WA: AAAI Press.

Jones, R. M., & Langley, P. (2005). A constrained architecture for learning and problem solving. *Computational Intelligence*, *21*, 480–502.

Iba, G. A. (1989). A heuristic approach to the discovery of macro-operators. *Machine Learning*, *3*, 285–317.

Langley, P. (2017). Heuristics and cognitive systems. *Advances in Cognitive Systems*, *5*, 3–12.

Langley, P. (2019). Explainable, normative, and justified agency. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence* (pp. 9775–9779). Honolulu, HI: AAAI Press.

Langley, P., & Meadows, B. (2019). Heuristic construction of explanations through associative abduction. *Advances in Cognitive Systems*, *8*, 93–112.

Law, M., Russo, A., & Broda, K. (2018). Inductive learning of answer set programs from noisy examples. *Advances in Cognitive Systems*, *7*, 57–76.

Meadows, B., Langley, P., & Emery, M. (2014). An abductive approach to understanding social interactions. *Advances in Cognitive Systems*, *3*, 87–106.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, *1*, 47–80.

Molineaux, M., Kuter, U., & Klenk, M. (2012). DiscoverHistory: Understanding the past in planning and execution. *Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems* (pp. 989–996). Valencia, Spain.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*, 87–127.

Ng, H. T. & Mooney, R. J. (1990). On the role of coherence in abductive explanation. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 337–342). Cambridge, MA.

Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly*, *13*, 470–482.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Swartout, W., Paris, C., & Moore, J. (1991). Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert*, *6*, 58–64.

VanLehn, K., & Jones, R. M. (1993). Integration of analogical search control and explanation-based learning of correctness. In S. Minton (Ed.), *Machine learning methods for planning*. Los Altos, CA: Morgan Kaufmann.

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings of the Nineteenth National Conference on Artificial Intelligence* (pp. 900–907). San Jose, CA: AAAI Press.

Veloso, M., Carbonell, J., Pérez, A., Borrajo, D., Fink, E., & Blythe, J. (1995). Integrating planning and learning: The PRODIGY architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, *7*, 81–120.

Winston, P. H. (2012). The right way. *Advances in Cognitive Systems*, *1*, 23–36.